Noname manuscript No.
(will be inserted by the editor)

Null Space Conditions and Thresholds for Rank Minimization

Benjamin Recht · Weiyu Xu · Babak Hassibi

Received: date / Accepted: date

Abstract Minimizing the rank of a matrix subject to constraints is a challenging problem that arises in many applications in Machine Learning, Control Theory, and discrete geometry. This class of optimization problems, known as rank minimization, is NP-HARD, and for most practical problems there are no efficient algorithms that yield exact solutions. A popular heuristic replaces the rank function with the nuclear norm—equal to the sum of the singular values—of the decision variable and has been shown to provide the optimal low rank solution in a variety of scenarios. In this paper, we assess the practical performance of this heuristic for finding the minimum rank matrix subject to linear equality constraints. We characterize properties of the null space of the linear operator defining the constraint set that are necessary and sufficient for the heuristic to succeed. We then analyze linear constraints sampled uniformly at random, and obtain dimension-free bounds under which our null space properties hold almost surely as the matrix dimensions tend to infinity. Finally, we provide empirical evidence that these probabilistic bounds provide accurate predictions of the heuristic's performance in non-asymptotic scenarios.

AMS (MOC) Subject Classification 90C25; 90C59; 15A52.

Keywords. rank, convex optimization, matrix norms, random matrices, compressed sensing, Gaussian processes.

B. Recht
Department of Computer Sciences, University of Wisconsin, 1210 W Dayton St, Madison, WI 53703
E-mail: brecht@cs.wisc.edu

W. Xu
Electrical Engineering, California Institute of Technology
E-mail: weiyu@systems.caltech.edu

B. Hassibi
Electrical Engineering, California Institute of Technology
E-mail: bhassibi@systems.caltech.edu

1 Introduction

The *rank minimization* problem consists of finding the minimum rank matrix in a convex constraint set. Though this problem is NP-Hard even when the constraints are linear, a recent paper by Recht et al. [29] showed that most instances of the linearly constrained rank minimization problem could be solved in polynomial time as long as there were sufficiently many linearly independent constraints. Specifically, they showed that minimizing the *nuclear norm* (also known as the Ky Fan 1-norm or the trace norm) of the decision variable subject to the same affine constraints produces the lowest rank solution if the affine space is selected at random. The nuclear norm of a matrix—equal to the sum of the singular values—can be optimized in polynomial time. This paper initiated a groundswell of research, and, subsequently, Candès and Recht showed that the nuclear norm heuristic could be used to recover low-rank matrices from a sparse collection of entries [8], Ames and Vavasis have used similar techniques to provide average case analysis of NP-HARD combinatorial optimization problems [1], and Vandenberghe and Zhang have proposed novel algorithms for identifying linear systems [23]. Moreover, fast algorithms for solving large-scale instances of this heuristic have been developed by many groups [7,21,24,26,29]. These developments provide new strategies for tackling the rank minimization problems that arise in Machine Learning [2,3,31], Control Theory [6,17,16], and dimensionality reduction [22,36,37].

Numerical experiments in [29] suggested that the nuclear norm heuristic significantly out-performed the theoretical bounds provided by their probabilistic analysis. They showed numerically that random instances of the nuclear norm heuristic exhibited a *phase transition* in the parameter space, where, for sufficiently small values of the rank the heuristic always succeeded. Surprisingly, in the complement of this region, the heuristic never succeeded. The transition between the two regions appeared sharp and the location of the phase transition appeared to be nearly independent of the problem size. A similar phase transition was also observed by Candès and Recht when the linear constraints merely revealed the values of a subset of the entries of the matrix [8].

In this paper we provide an approach to explicitly calculate the location of this phase transition and provide bounds for the success of the nuclear norm heuristic that accurately reflect empirical performance. We describe a *necessary* and sufficient condition for the solution of the nuclear norm heuristic to coincide with the minimum rank solution in an affine space. This condition, first reported in [30], characterizes a particular property of the null space of the linear map which defines the affine space and is generalized from similar properties in compressed sensing [12,38,33]. We then show that when the null space is sampled from the uniform distribution on subspaces, the null space characterization holds with overwhelming probability provided the number of equality constraints exceeds a threshold. We provide explicit formulas relating the dimension of the null space to the largest rank matrix that can be found using the nuclear norm heuristic. We also compare our results against the empirical findings of [29] and demonstrate that they provide a good approximation of the phase transition boundary especially when the number of constraints is large.

1.1 Main Results

Let X be an $n_1 \times n_2$ matrix decision variable. Without loss of generality, we will assume throughout that $n_1 \leq n_2$. Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a linear map, and let $b \in \mathbb{R}^m$. The main optimization problem under study is

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & \mathcal{A}(X) = b\,. \end{array} \tag{1}$$

This problem is known to be NP-HARD and is also hard to approximate [26]. As mentioned above, a popular heuristic for this problem replaces the rank function with the sum of the singular values of the decision variable. Let $\sigma_i(X)$ denote the i-th largest singular value of X (equal to the square-root of the i-th largest eigenvalue of XX^*). Recall that the rank of X is equal to the number of nonzero singular values. In the case when the singular values are all equal to one, the sum of the singular values is equal to the rank. When the singular values are less than or equal to one, the sum of the singular values is a convex function that is strictly less than the rank. This sum of the singular values is a unitarily invariant matrix norm, called the *nuclear norm*, and is denoted

$$\|X\|_* := \sum_{i=1}^{r} \sigma_i(X)\,.$$

This norm is alternatively known by several other names including the Schatten 1-norm, the Ky Fan norm, and the trace class norm.

As described in the introduction, our main concern is when the optimal solution of (1) coincides with the optimal solution of

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \mathcal{A}(X) = b\,. \end{array} \tag{2}$$

This norm minimization problem is convex, and can be efficiently solved via a variety of methods including semidefinite programming. See [29] for a survey and [7, 23, 24] for customized algorithms.

We characterize an instance of the affine rank minimization problem (1) by three dimensionless parameters that take values in $(0, 1]$: the *aspect ratio* γ, the *constraint ratio* μ, and the *rank ratio* β. The aspect ratio is the number of rows divided by the number of columns: $\gamma = n_1/n_2$. The constraint ratio is the number of constraints divided by the number of parameters needed to fully specify an $n_1 \times n_2$ matrix. That is, $m = \mu n_1 n_2$. The rank ratio is the minimum rank attainable in Problem (1) divided by the number of rows of the decision variable. That is, if the minimum rank solution of (1) has rank r, then $\beta = r/n_1$. The main focus of this paper is determining for which triples (β, γ, μ) the problem (2) has the same optimal solution as the rank minimization problem (1).

Our first result characterizes when a particular low-rank matrix can be recovered from a random linear system via nuclear norm minimization.

Theorem 1 (Weak Bound) *Set $n_1 \leq n_2$, $\gamma = n_1/n_2$, and let X_0 be an $n_1 \times n_2$ matrix with of rank $r = \beta n_1$. Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{\mu n_1 n_2}$ denote the random linear transformation*

$$\mathcal{A}(X) = \mathbf{A}\operatorname{vec}(X)\,,$$

where each entry of $\mathbf{A}$ *is sampled independently from a normal distribution with mean zero and variance one. Define the function* $\varphi : (0,1] \to \mathbb{R}$ *by*

$$\varphi(\gamma) := \frac{1}{2\pi} \int_{(1-\sqrt{\gamma})^2}^{(1+\sqrt{\gamma})^2} \sqrt{\frac{-z^2 + 2(1+\gamma)z - (1-\gamma)^2}{z}} dz \,. \tag{3}$$

Then whenever

$$\mu \geq 1 - \left(\varphi\left(\frac{\gamma - \beta\gamma}{1 - \beta\gamma} \right) \frac{(1-\beta\gamma)^{3/2}}{\gamma} - \frac{8}{3\pi} \gamma^{1/2} \beta^{3/2} \right)^2 \tag{4}$$

there exists a numerical constant $c_w(\mu, \beta, \gamma) > 0$ *such that with probability exceeding* $1 - e^{-c_w(\mu,\beta,\gamma)n_2^2 + o(n_2^2)}$,

$$X_0 = \arg\min\{\|Z\|_* \ : \ \mathcal{A}(Z) = \mathcal{A}(X_0)\} \,.$$

In particular, if β*,*γ*, and* μ *satisfy (4), then nuclear norm minimization will recover* X_0 *from a random set of* $\mu\gamma n_2^2$ *constraints drawn from the Gaussian ensemble almost surely as* $n_2 \to \infty$*.*

Formula (4) provides a lower-bound on the empirical phase transition observed in [29]. Since $\mathcal{A}$ is Gaussian, the null space of $\mathcal{A}$, that is the set of Y such that $\mathcal{A}(Y) = 0$, is identically distributed to the uniform distribution of $(1-\mu)n_1n_2$ dimensional subspaces. Since the constraint set is uniquely determined by the null space of $\mathcal{A}$, Thereom 1 holds for any distribution of linear maps whose null spaces are uniformly distributed. From this perspective, the theorem states that the nuclear norm heuristic succeeds for almost all instances of the affine rank minimization problem with parameters (β, γ, μ) satisfying (4). A particular case of interest is the case of square matrices ($\gamma = 1$). In this case, the Weak Bound (4) takes the elegant closed form:

$$\mu \geq 1 - \frac{64}{9\pi^2} \left((1-\beta)^{3/2} - \beta^{3/2} \right)^2 \,. \tag{5}$$

Figure 1(a) plots these thresholds for varying γ. The y-axis here denotes the ratio of the number of parameters of a low-rank matrix divided by the number of measurements. The *model size* is the number of parameters required to define a low rank matrix. An $n_1 \times n_2$ matrix of rank r is defined by $r(n_1 + n_2 - r)$ parameters (this quantity can be computed by calculating the number of parameters needed to specify the singular value decomposition). In terms of the parameters β and γ, the model size is equal to $\beta(1 + \gamma - \beta\gamma)n_2^2$. The bounds in Figure 1a demonstrate that when the number of measurements is a constant fraction of the total number of entries in the unknown matrix, the nuclear norm heuristic will succeed as long as the number of constraints is a constant factor larger than the number of *intrinsic* parameters of a low-rank matrix. In other words, the number of measurements required to recover a low-rank matrix scales proportionally to the model size in some measurement regimes. Larger oversampling is required to recover matrices with larger aspect ratio γ. In fact, as γ approaches 0, an oversampling of the model size by a factor of 2 suffices for exact recovery for most values of μ and β. These patterns are also observed experimentally in Section 3.

The second theorem characterizes when the nuclear norm heuristic succeeds at recovering *all* low rank matrices.

Theorem 2 (Strong Bound) *Let $\mathcal{A}$ be defined as in Theorem 1. Define the two functions*

$$f(\gamma,\beta,\epsilon) = \frac{\varphi\left(\frac{\gamma-\beta\gamma}{1-\beta\gamma}\right)(1-\beta\gamma)^{3/2} - \frac{8}{3\pi}\gamma^{3/2}\beta^{3/2} - 4\epsilon\varphi(\gamma)}{1+4\epsilon}$$

$$g(\gamma,\beta,\epsilon) = \sqrt{2\beta\gamma(1+\gamma-\beta\gamma)\log\left(\frac{3\pi}{\epsilon}\right)}\,.$$

with φ as in equation (3). Then there exists a numerical constant $c_s(\mu,\beta) > 0$ such that with probability exceeding $1 - e^{-c_s(\mu,\beta)n^2+o(n^2)}$, for all *$\gamma n \times n$ matrices X_0 of rank $r \leq \beta\gamma n$*

$$X_0 = \arg\min\{\|Z\|_* \;:\; \mathcal{A}(Z) = \mathcal{A}(X_0)\}$$

whenever

$$\mu \geq 1 - \sup_{\substack{\epsilon>0 \\ f(\beta,\epsilon)-g(\beta,\epsilon)>0}} \gamma^{-2}\left(f(\beta,\epsilon) - g(\beta,\epsilon)\right)^2 . \tag{6}$$

In particular, if β, γ, and μ satisfy (6), then nuclear norm minimization will recover all rank r matrices from a random set of $\gamma\mu n^2$ constraints drawn from the Gaussian ensemble almost surely as $n \to \infty$.

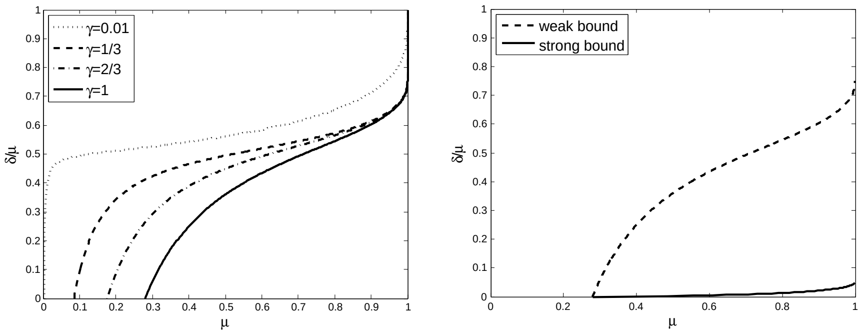
Figure 1(b) plots the bound from Theorems 1 and 2 with $\gamma = 1$. We call (4) the *Weak Bound* because it is a condition that depends on the optimal solution of (1). On the other hand, we call (6) the *Strong Bound* as it guarantees the nuclear norm heuristic succeeds, *no matter what the optimal solution*, as long as the true minimum rank is sufficiently small. The Weak Bound is the only bound that can be tested experimentally, and, in Section 3, we will show that it corresponds well to experimental data. Moreover, the Weak Bound provides guaranteed recovery over a far larger region of the (β,μ) parameter space. Nonetheless, the mere existence of a Strong Bound is surprising, and results in a far less conservative bound than what was available from previous results (c.f., [29]).

Both Theorems 1 and Theorem 2 were first announced in [30] in the case of $\gamma = 1$ without proof. The present work provides the previously unpublished proofs and generalizes to matrices with arbitrary aspect ratios.

1.2 Related Work

Optimization problems involving constraints on the rank of matrices are pervasive in engineering applications. For example, in Machine Learning, these problems arise in the context of inference with partial information [31] and multi-task learning [3]. In Control Theory, problems in controller design [17,27], minimal realization theory [16], and model reduction [6] can be formulated as rank minimization problems. Rank minimization also plays a key role in the study of embeddings of discrete metric spaces in Euclidean space [22] and of learning structure in data and manifold learning [36].

In certain instances with special structure, the rank minimization problem can be solved via the singular value decomposition or can be reduced to the solution of a linear system [27,28]. In general, however, minimizing the rank of a matrix



(a) (b)

Fig. 1 (a) Bounds on low-rank recovery varying the aspect ratio of the matrix. Here the x-axis denotes the measurement ratio, μ, or the number of linear equations divided by the total number of entries in the unknown matrix. The y-axis denotes the ratio of the number of parameters in the unknown matrix to the number of measurements. δ is the number of true parameters (called the *model size*) divided by the number of entries. As described in the text, $\delta = r(n_1 + n_2 - r)/(n_1 n_2)$. γ is the aspect ratio of the matrix equal to n_1/n_2. (b) The Weak Bound (4) versus the Strong Bound (6). The axes here are the same as in (a), and $\gamma = 1$ in both cases.

subject to convex constraints is NP-HARD. Even the problem of finding the lowest rank matrix in an affine space is NP-HARD. The best exact algorithms for this problem involve quantifier elimination and such solution methods require at least exponential time in the dimensions of the matrix variables.

Nuclear norm minimization is a recent heuristic for rank minimization introduced by Fazel in [15]. When the matrix variable is symmetric and positive semidefinite, this heuristic is equivalent to the “trace heuristic” from Control Theory (see, e.g., [6,27]). Both the trace heuristic and the nuclear norm generalization have been observed to produce very low-rank solutions in practice, but, until very recently, conditions where the heuristic succeeded were only available in cases that could also be solved by elementary linear algebra [28]. As mentioned above, the first non-trivial sufficient conditions that guaranteed the success of the nuclear norm heuristic were provided in [29].

The initial results in [29] build on seminal developments in “compressed sensing” that determined conditions for when minimizing the ℓ_1 norm of a vector over an affine space returns the sparsest vector in that space (see, e.g., [10,9,5]). There is a strong parallelism between the sparse approximation and rank minimization settings. The rank of a diagonal matrix is equal to the number of non-zeros on the diagonal. Similarly, the sum of the singular values of a diagonal matrix is equal to the ℓ_1 norm of the diagonal. Exploiting the parallels, the authors in [29] were able to extend much of the analysis developed for the ℓ_1 heuristic to provide guarantees for the nuclear norm heuristic.

Building on this work, Candès and Recht showed that most $n \times n$ matrices with rank at most r can be recovered from a sampling of on the order of $(n^{1.2}r)$ of the entries [8] using nuclear norm minimization. In another recently provided extension, Meka et al. [26] have provided an analysis of the multiplicative weights algorithm for providing very low-rank approximate solutions of systems of inequalities. Ames and Vavasis have demonstrated that the nuclear norm heuristic can solve many instances of the NP-Hard combinatorial optimization problems maximum clique and maximum biclique [1].

Focusing on the special case where one seeks the lowest rank matrix in an affine subspace, Recht et al. generalized the notion of "restricted isometry" from [10] to the space of low rank matrices. They provided deterministic conditions on the linear map defining the affine subspace which guarantees the minimum nuclear norm solution is the minimum rank solution. Moreover, they provided several ensembles of affine constraints where this sufficient condition holds with overwhelming probability. They proved that the heuristic succeeds with large probability whenever the number m of available measurements is greater than a constant times $2nr \log n$ for $n \times n$ matrices. Since a matrix of rank r cannot be specified with less than $r(2n - r)$ real numbers, this is, up to asymptotic scaling, a nearly optimal result. However, the bounds developed in this paper did not reflect the empirical performance of the nuclear norm heuristic. In particular, it gave vacuous results for practically sized problems where the rank was large. The results in the present work provide bounds that much more closely approximate the practical recovery region of the heuristic.

The present work builds on a different collection of developments in compressed sensing [12–14, 33, 38]. In these papers, the authors studied properties of the null space of the linear operator that gives rise to the affine constraints. The bounds resulting from these approaches were significantly sharper than those obtained in earlier work on sparse recovery such as [10]. In particular, the thresholds obtained in [11] closely approximate actual experimental results. The null space criteria described in Section 2.1 generalize the concepts of the same name in Compressed Sensing.

Unfortunately, the polyhedral analysis of the null spaces arising in compressed sensing does not extend to the low-rank matrices as the unit ball in the nuclear norm is not a polyhedral set. Figure 2 plots a simple three dimensional example, depicting the unit ball of the nuclear norm for matrices parameterized as

$$\left\{ X \ : \ X = \begin{bmatrix} x & y \\ y & z \end{bmatrix} , \ \|X\|_* \leq 1 \right\} . \tag{7}$$

In order to extend null space analysis to the rank minimization problem, we need to follow a different path. In [33], the authors provide a probabilistic argument specifying a large region where the minimum ℓ_1 solution is the sparsest solution. This works by directly estimating the probability of success via a simple Chernoff-style argument. Our work follows this latter approach, but requires the introduction of specialized machinery to deal with the asymptotic behavior of the singular values of random matrices. We provide a sufficient statistic that guarantees the heuristic succeeds, and then use comparison lemmas for Gaussian processes to bound the expected value of this heuristic (see, for example, [20]). We then show that this random variable is sharply concentrated around its expectation.

1.3 Notation and Preliminaries

For a rectangular matrix $X \in \mathbb{R}^{n_1 \times n_2}$, X^* denotes the transpose of X. vec(X) denotes the vector in $\mathbb{R}^{n_1 n_2}$ with the columns of X stacked on top of one another. A *projection operator* always refers to a square matrix P such that $P^2 = P$. We

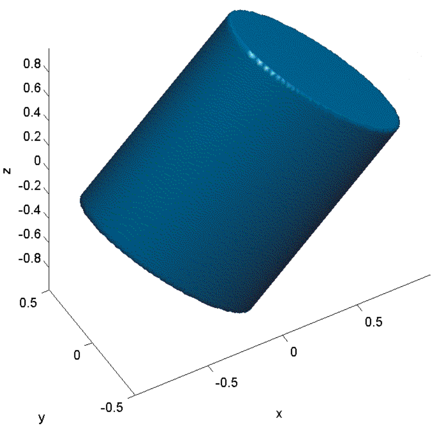


Fig. 2 The unit ball of the nuclear norm. The figure depicts the set of all matrices of the form of equation (7) with nuclear norm less than one.

say that a projection operator projects onto a d-dimensional subspace if its range has dimension d.

For vectors $v \in \mathbb{R}^d$, the only norm we will ever consider is the Euclidean norm

$$\|v\|_{\ell_2} = \left(\sum_{i=1}^{d} v_i^2\right)^{1/2} .$$

On the other hand, we will consider a variety of matrix norms. For matrices X and Y of the same dimensions, we define the inner product in $\mathbb{R}^{n_1 \times n_2}$ as $\langle X, Y\rangle := \operatorname{trace}(X^*Y) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{ij}Y_{ij}$. The norm associated with this inner product is called the Frobenius (or Hilbert-Schmidt) norm $||\cdot||_F$. The Frobenius norm is also equal to the Euclidean, or ℓ_2, norm of the vector of singular values, i.e.,

$$\|X\|_F := \left(\sum_{i=1}^{r} \sigma_i^2\right)^{\frac{1}{2}} = \sqrt{\langle X, X\rangle} = \left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{ij}^2\right)^{\frac{1}{2}}$$

The operator norm (or induced 2-norm) of a matrix is equal to its largest singular value (i.e., the ℓ_∞ norm of the singular values):

$$\|X\| := \sigma_1(X).$$

The nuclear norm of a matrix is equal to the sum of its singular values, i.e.,

$$\|X\|_* := \sum_{i=1}^{r} \sigma_i(X) .$$

These three norms are related by the following inequalities which hold for any matrix X of rank at most r:

$$||X|| \leq ||X||_F \leq ||X||_* \leq \sqrt{r}||X||_F \leq r||X||.$$

To any norm, we may associate a *dual norm* via the following variational definition

$$\|X\|_d = \sup_{\|Y\|_p=1} \langle Y, X \rangle \,.$$

One can readily check that the dual norm of the Frobenius norm is the Frobenius norm. Less trivially, one can show that the dual norm of the operator norm is the nuclear norm (See, for example, [29]). We will leverage the duality between the operator and nuclear norm several times in our analysis.

Finally, we define the random ensemble of $d_1 \times d_2$ matrices $\mathfrak{G}(d_1, d_2)$ to be the Gaussian ensemble, with each entry sampled i.i.d. from a Gaussian distribution with zero-mean and variance one.

2 Proofs of the Probabilistic Bounds

We now turn to the proofs of the probabilistic bounds (4) and (6). We first review necessary and sufficient null space conditions for the nuclear norm presented in [30]. Then, noting that the null space of $\mathcal{A}$ is spanned by Gaussian vectors, we use bounds from probability on Banach Spaces to show that the respective sufficient conditions are met when the Weak Bound (4) or the Strong Bound (6) hold. This will require the introduction of two useful auxiliary functions whose actions on Gaussian processes are explored in Section 2.4.

2.1 Sufficient Conditions for null space Characterizations

Whenever $\mu < 1$, the null space of $\mathcal{A}$ contains a non-zero matrix. Note that X is the unique optimal solution for (2) if and only if for every Y in the null space of $\mathcal{A}$

$$\|X + Y\|_* > \|X\|_* \,. \tag{8}$$

The following theorem, originally proven in [30], generalizes this null space criterion to a critical property that guarantees when the nuclear norm heuristic finds the minimum rank solution of $\mathcal{A}(X) = b$ as long as the minimum rank solution is sufficiently small.

Theorem 3 *Let X_0 be the optimal solution of (1) and suppose* $\mathrm{rank}(X_0) \leq r$.

1. *If for every nonzero Y in the null space of $\mathcal{A}$ and for every decomposition*

$$Y = Y_1 + Y_2,$$

where Y_1 has rank r and Y_2 has rank greater than r, it holds that

$$\|Y_1\|_* < \|Y_2\|_*,$$

then X_0 is the unique minimizer of (2).

2. *Conversely, if the condition of part 1 does not hold, then there exists a vector $b \in \mathbb{R}^m$ such that the minimum rank solution of $\mathcal{A}(X) = b$ has rank at most r and is not equal to the minimum nuclear norm solution.*

For completeness, a short proof of this theorem is included in the appendix. For the purposes of proving the Strong Bound, the following theorem gives us a sufficient but more easily analyzed condition that implies Condition 1 in Theorem 3.

Theorem 4 *Let $\mathcal{A}$ be a linear map of $n_1 \times n_2$ matrices into $\mathbb{R}^m$. Let P and Q be projection operators onto r-dimensional subspaces of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ respectively. Suppose that for every Y in the null space of $\mathcal{A}$*

$$\|(I-P)Y(I-Q)\|_* \geq \|PYQ\|_* \,. \tag{9}$$

Then for every matrix Z with row and column spaces equal to the range of Q and P respectively,

$$\|Z+Y\|_* \geq \|Z\|_*$$

for all Y in the null space of $\mathcal{A}$. Moreover, if the condition (9) holds for every pair of projection operators P and Q onto r-dimensional subspaces, then for every Y in the null space of $\mathcal{A}$ and for every decomposition $Y = Y_1 + Y_2$ where Y_1 has rank r and Y_2 has rank greater than r, it holds that $\|Y_1\|_* \leq \|Y_2\|_*$.

We will need the following lemma

Lemma 1 *For any block partitioned matrix*

$$X = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

we have $\|X\|_* \geq \|A\|_* + \|D\|_*$.

Proof This lemma follows from the dual description of the nuclear norm:

$$\|X\|_* = \sup\left\{ \left\langle \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}, \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right\rangle \quad \middle| \quad \left\| \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \right\| = 1 \right\} \,, \tag{10}$$

and similarly

$$\|A\|_* + \|D\|_* = \sup\left\{ \left\langle \begin{bmatrix} Z_{11} & 0 \\ 0 & Z_{22} \end{bmatrix}, \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right\rangle \quad \middle| \quad \left\| \begin{bmatrix} Z_{11} & 0 \\ 0 & Z_{22} \end{bmatrix} \right\| = 1 \right\} \,. \tag{11}$$

Since (10) is a supremum over a larger set that (11), the claim follows.

Theorem 4 now trivially follows.

Proof (of Theorem 4) Without loss of generality, we may choose coordinates such that P and Q both project onto the space spanned by first r standard basis vectors. Then we may partition Y as

$$Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

and write, using Lemma 1,

$$\begin{aligned} \|Y-Z\|_* - \|Z\|_* &= \left\| \begin{bmatrix} Y_{11} - Z & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \right\|_* - \|Z\|_* \\ &\geq \|Y_{11} - Z\|_* + \|Y_{22}\|_* - \|Z\|_* \\ &\geq \|Y_{22}\|_* - \|Y_{11}\|_* \end{aligned}$$

which is non-negative by assumption. Note that if the theorem holds for all projection operators P and Q whose range has dimension r, then $\|Z+Y\|_* \geq \|Z\|_*$ for all matrices Z of rank r and hence the second part of the theorem follows.

2.2 Proof of the Weak Bound

Now we can turn to the proof of Theorem 1. The key observation in proving this theorem is the following characterization of the null space of $\mathcal{A}$ provided by Stojnic et al [33]

Lemma 2 *Let $\mathcal{A}$ be sampled from $\mathfrak{G}(\mu n_1 n_2, n_1 n_2)$. Then the null space of $\mathcal{A}$ is identically distributed to the span of $n_1 n_2(1-\mu)$ matrices G_i where each G_i is sampled i.i.d. from $\mathfrak{G}(n_1, n_2)$. In other words, we may assume that $w \in \ker(\mathcal{A})$ can be written as $\sum_{i=1}^{n_1 n_2(1-\mu)} v_i G_i$ for some $v \in \mathbb{R}^{n_1 n_2(1-\mu)}$.*

This is nothing more than a statement that the null space of $\mathcal{A}$ is a random subspace. However, when we parameterize elements in this subspace as linear combinations of Gaussian vectors, we can leverage Comparison Theorems for Gaussian processes to yield our bounds.

Let $M = n_1 n_2(1-\mu)$ and let $G_1, \ldots, G_M$ be i.i.d. samples from $\mathfrak{G}(n_1, n_2)$. Let X_0 be a matrix of rank βn_1. Let P_{X_0} and Q_{X_0} denote the projections onto the column and row spaces of X_0 respectively. By Theorem 4 and Lemma 2, we need to show that for all $v \in \mathbb{R}^M$,

$$\left\|(I - P_{X_0})\left(\sum_{i=1}^{M} v_i G_i\right)(I - Q_{X_0})\right\|_* \geq \left\|P_{X_0}\left(\sum_{i=1}^{M} v_i G_i\right) Q_{X_0}\right\|_* . \qquad (12)$$

That is, $\sum_{i=1}^{M} v_i G_i$ is an arbitrary element of the null space of $\mathcal{A}$, and this equation restates the sufficient condition provided by Theorem 4. Now it is clear by homogeneity that we can restrict our attention to those $v \in \mathbb{R}^M$ with Euclidean norm 1. The following lemma characterizes when the expected value of this difference is nonnegative:

Lemma 3 *Let $n_1 = \gamma n_2$ for some $\gamma \in (0,1]$ and $r = \beta n_1$ for some $\beta \in (0,1]$. Suppose P and Q are projection operators onto r-dimensional subspaces of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ respectively. For $i = 1, \ldots, M$ let G_i be sampled from $\mathfrak{G}(n_1, n_2)$. Then*

$$\begin{aligned}\mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1} \left\|(I-P)\left(\sum_{i=1}^{M} v_i G_i\right)(I-Q)\right\|_* - \left\|P\left(\sum_{i=1}^{M} v_i G_i\right) Q\right\|_*\right] \\ \geq \left(\alpha_1(1-\beta\gamma)^{3/2} - \alpha_2 \gamma^{3/2}\beta^{3/2}\right) n_2^{3/2} - \sqrt{M\gamma n_2}\,,\end{aligned} \qquad (13)$$

where $\alpha_1 = \varphi\left(\frac{\gamma-\beta\gamma}{1-\beta\gamma}\right) + o(1)$ and $\alpha_2 = \frac{8}{3\pi} + o(1)$.

We will prove this lemma and a similar inequality required for the proof of the Strong Bound in Section 2.4 below. But we now show how using this lemma and a concentration of measure argument, we can prove Theorem 1.

First note, that if we plug in $M = (1-\mu)n_1 n_2$, divide the right hand side by $n_2^{3/2}$, and ignore the $o(1)$ terms, the right hand side of (13) is non-negative if (4) holds. To bound the probability that (12) is non-negative, we employ a powerful concentration inequality for the Gaussian distribution bounding deviations of smoothly varying functions from their expected value.

To quantify what we mean by smoothly varying, recall that a function f is *Lipschitz* with respect to the Euclidean norm if there exists a constant L such that $|f(x) - f(y)| \leq L\|x-y\|_{\ell_2}$ for all x and y. The smallest such constant L is called the *Lipschitz constant* of the map f. If f is Lipschitz, it cannot vary too rapidly. In particular, note that if f is differentiable and Lipschitz, then L is a bound on the norm of the gradient of f. The following theorem states that the deviations of a Lipschitz function applied to a Gaussian random variable have Gaussian tails.

Theorem 5 *Let $x \in \mathbb{R}^D$ be a normally distributed random vector with zero-mean variance equal to the identity. Let $f : \mathbb{R}^D \to \mathbb{R}$ be a function with Lipschitz constant L. Then*

$$\mathbb{P}[|f(x) - \mathbb{E}[f(x)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2L^2}\right) .$$

See [20] for a proof of this theorem with slightly weaker constants and a list of several references to more complicated proofs that give rise to this concentration inequality. The following lemma bounds the Lipschitz constant of interest

Lemma 4 *For $i = 1, \ldots, M$, let $X_i \in \mathbb{R}^{D_1 \times D_2}$ and $Y_i \in \mathbb{R}^{D_3 \times D_4}$ with $D_1 \leq D_2$ and $D_3 \leq D_4$. Define the function*

$$F_I(X_1, \ldots, X_M, Y_1, \ldots, Y_M) = \inf_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^M v_i X_i \right\|_* - \left\| \sum_{i=1}^M v_i Y_i \right\|_* .$$

Then the Lipschitz constant of F_I is at most $\sqrt{D_1 + D_3}$.

The proof of this lemma is straightforward and can be found in the Appendix. Using Theorem 5 and Lemmas 3 and 4, we can now bound

$$\mathbb{P}\left[\inf_{\|v\|_{\ell_2}=1} \left\| (I - P_{X_0}) \left(\sum_{i=1}^M v_i G_i \right) (I - Q_{X_0}) \right\|_* \right.$$
$$\left. - \left\| P_{X_0} \left(\sum_{i=1}^M v_i G_i \right) Q_{X_0} \right\|_* \leq t n_2^{3/2} \right] \leq \exp\left(-\frac{u_w^2 n_2^2}{2\gamma} + o(n_2^2) \right) \tag{14}$$

with

$$u_w = \varphi\left(\frac{\gamma - \beta\gamma}{1 - \beta\gamma}\right)(1 - \beta\gamma)^{3/2} - \frac{8}{3\pi}\gamma^{3/2}\beta^{3/2} - \gamma\sqrt{1-\mu} - t\,.$$

Here, we use $D_1 = D_2 = n_1 - r$ and $D_3 = D_4 = r$ in Lemma 4. Setting $t = 0$, we see that u_w is non-negative as long as the triple (β, γ, μ) satisfies (4). This completes the proof of Theorem 1. We will use the concentration inequality (14) with a non-zero t to prove the Strong Bound.

2.3 Proof of the Strong Bound

The proof of Theorem 2 is similar to that of Theorem 1 except we prove that (12) holds for *all* operators P and Q that project onto r-dimensional subspaces. Our proof will require an ϵ-net for the projection operators. By an ϵ-net, we mean a finite set Ω consisting of pairs of r-dimensional projection operators such that for any P and Q that project onto r-dimensional subspaces, there exists $(P', Q') \in \Omega$

with $\|P-P'\|+\|Q-Q'\| \le \epsilon$. We will show that if a slightly stronger bound than (12) holds on the ϵ-net, then (12) holds for all choices of row and column spaces.

Let us first examine how (12) changes when we perturb P and Q. Let P, Q, P' and Q' all be projection operators onto r-dimensional subspaces of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ respectively. Let W be some $n_1 \times n_2$ matrix and observe that

$$\begin{aligned}
&\|(I-P)W(I-Q)\|_* - \|PWQ\|_* - (\|(I-P')W(I-Q')\|_* - \|P'WQ'\|_*)\\
\le&\|(I-P)W(I-Q) - (I-P')W(I-Q')\|_* + \|PWQ - P'WQ'\|_*\\
\le&\|(I-P)W(I-Q) - (I-P')W(I-Q)\|_*\\
&\qquad + \|(I-P')W(I-Q) - (I-P')W(I-Q')\|_*\\
&\qquad + \|PWQ - P'WQ\|_* + \|P'WQ - P'WQ'\|_*\\
\le&\|P-P'\|\|W\|_*\|I-Q\| + \|I-P'\|\|W\|_*\|Q-Q'\|\\
&\qquad + \|P-P'\|\|W\|_*\|Q\| + \|P'\|\|W\|_*\|Q-Q'\|\\
\le&2(\|P-P'\|+\|Q-Q'\|)\|W\|_*\,.
\end{aligned}$$

Here, the first and second inequalities follow from the triangle inequality, the third inequality follows because $\|AB\|_* \le \|A\|\|B\|_*$, and the fourth inequality follows because P, P', Q, and Q' are all projection operators. Rearranging this inequality gives

$$\begin{aligned}
\|(I-P)W(I-Q)\|_* - \|PWQ\|_* \ge \|(I-P')W(I-Q')\|_* - \|P'WQ'\|_*\\
- 2(\|P-P'\|+\|Q-Q'\|)\|W\|_*\,.
\end{aligned} \tag{15}$$

Let us now suppose that with overwhelming probability

$$\|(I-P')W(I-Q')\|_* - \|P'WQ'\|_* - 4\epsilon\|W\|_* \ge 0 \tag{16}$$

for all (P',Q') in our ϵ-net Ω. Then by (15), this means that $\|(I-P)W(I-Q)\|_* - \|PWQ\|_* \ge 0$ for any arbitrary pair of projection operators onto r-dimensional subspaces. Thus, if we can show that (16) holds for all (P',Q') in an ϵ-net and for all W in the null space of $\mathcal{A}$, then we will have proven the Strong Bound.

To proceed, we need to know the size of an ϵ-net. The following bound on such a net is due to Szarek.

Theorem 6 (Szarek [35]) *Consider the space of all projection operators on $\mathbb{R}^n$ projecting onto r dimensional subspaces endowed with the metric*

$$d(P,P') = \|P-P'\|\,.$$

Then there exists an ϵ-net in this metric of cardinality at most $\left(\frac{3\pi}{2\epsilon}\right)^{r(n-r/2-1/2)}$.

With this covering number in hand, we now calculate the probability that for a given P and Q in the ϵ-net,

$$\begin{aligned}
\inf_{\|v\|_{\ell_2}=1}\left\{\left\|(I-P)\left(\sum_{i=1}^M v_iG_i\right)(I-Q)\right\|_* - \left\|P\left(\sum_{i=1}^M v_iG_i\right)Q\right\|_*\right\}\\
\ge 4\epsilon \sup_{\|v\|_{\ell_2}=1}\left\|\sum_{i=1}^M v_iG_i\right\|_*\,.
\end{aligned} \tag{17}$$

As we will show in Section 2.4, we can upper bound the right hand side of this inequality using a similar bound as in Lemma 3.

Lemma 5 *For $i = 1, \ldots, M$ let G_i be sampled from $\mathfrak{G}(\gamma n, n)$ with $\gamma \in (0, 1]$. Then*

$$\mathbb{E}\left[\sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^{M} v_i G_i \right\|_*\right] \leq (\varphi(\gamma) + o(1))\, n^{3/2} + \sqrt{\gamma M n}\,. \tag{18}$$

Moreover, we prove the following in the appendix.

Lemma 6 *For $i = 1, \ldots, M$, let $X_i \in \mathbb{R}^{D_1 \times D_2}$ with $D_1 \leq D_2$ and define the function*

$$F_S(X_1, \ldots, X_M) = \sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^{M} v_i X_i \right\|_* .$$

Then the Lipschitz constant of F_S is at most $\sqrt{D_1}$.

Using Lemmas 5 and 6 combined with Theorem 5, we have that

$$\mathbb{P}\left[\sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^{M} v_i G_i \right\|_* \geq \frac{t n_2^{3/2}}{4\epsilon}\right] \leq \exp\left(-\frac{\left(\varphi(\gamma) - \gamma\sqrt{1-\mu} - \frac{t}{4\epsilon} + o(1)\right)^2 n_2^2}{2\gamma}\right). \tag{19}$$

Here, $D_1 = n_1$ and $D_2 = n_2$ when applying Lemma 6. Let t_0 be such that the exponents of (14) and (19) equal each other. Then we find after some algebra and the union bound

$$\begin{aligned}
&\mathbb{P}\left[(17) \text{ holds for fixed } P \text{ and } Q\right] \\
&\geq 1 - \mathbb{P}\left[\inf_{\|v\|_{\ell_2}=1} \left\|(I-P)\left(\sum_{i=1}^{M} v_i G_i\right)(I-Q)\right\|_* - \left\|P\left(\sum_{i=1}^{M} v_i G_i\right)Q\right\|_* < t_0 n_2^{3/2}\right] \\
&\qquad - \mathbb{P}\left[4\epsilon \sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^{M} v_i G_i \right\|_* > t_0 n_2^{3/2}\right] \\
&\geq 1 - 2\exp\left(-\frac{u_s^2 n_2^2}{2\gamma} + o(n_2^2)\right)
\end{aligned}$$

with

$$u_s = \frac{\varphi\left(\frac{\gamma - \beta\gamma}{1-\beta\gamma}\right)(1-\beta\gamma)^{3/2} - \frac{8}{3\pi}\gamma^{3/2}\beta^{3/2} - 4\epsilon\varphi(\gamma)}{1+4\epsilon} - \gamma\sqrt{1-\mu}\,.$$

Now, let Ω be an ϵ-net for the set of pairs of projection operators (P, Q) such that P (resp. Q) projects $\mathbb{R}^{n_1}$ (resp. $\mathbb{R}^{n_2}$) onto an r-dimensional subspace. By Theorem 6, we may assume $|\Omega| \leq \left(\frac{3\pi}{\epsilon}\right)^{r(n_1+n_2-r)}$. Again by the union bound, we have that

$$\begin{aligned}
&\mathbb{P}\left[(17) \text{ holds } \forall (P, Q) \in \Omega\right] \\
&\geq 1 - 2\exp\left(-\tfrac{1}{2\gamma}\left\{\left(f(\beta,\gamma,\epsilon) - \gamma\sqrt{1-\mu}\right)^2 - g(\beta,\gamma,\epsilon)^2\right\} n_2^2 + o(n_2^2)\right)
\end{aligned}$$

where

$$f(\gamma,\beta,\epsilon) = \frac{\varphi\left(\frac{\gamma-\beta\gamma}{1-\beta\gamma}\right)(1-\beta\gamma)^{3/2} - \frac{8}{3\pi}\gamma^{3/2}\beta^{3/2} - 4\epsilon\varphi(\gamma)}{1+4\epsilon}$$
$$g(\gamma,\beta,\epsilon) = \sqrt{2\beta\gamma(1+\gamma-\beta\gamma)\log\left(\frac{3\pi}{\epsilon}\right)}\,.$$

We have already shown that if (17) holds for all pairs in Ω, it holds for all pairs of projection operators projecting onto subspaces of dimension at most r. Thus, finding the parameters μ, β, γ, and ϵ that make the terms multiplying n_2^2 negative completes the proof of the Strong Bound.

2.4 Comparison Theorems for Gaussian Processes and the Proofs of Lemmas 3 and 5

Both of the two following Comparison Theorems provide sufficient conditions for when the expected supremum or infimum of one Gaussian process is greater to that of another. Elementary proofs of both of these theorems and several other Comparison Theorems can be found in §3.3 of [20].

Theorem 7 (Slepian's Lemma [32]) *Let X and Y be Gaussian random vectors in $\mathbb{R}^N$ such that*

$$\begin{cases}\mathbb{E}[X_iX_j] \le \mathbb{E}[Y_iY_j] & \textit{for all } i \neq j\\ \mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2] & \textit{for all } i\end{cases}$$

Then

$$\mathbb{E}[\max_i Y_i] \le \mathbb{E}[\max_i X_i]\,.$$

Theorem 8 (Gordan [18,19]) *Let $X = (X_{ij})$ and $Y = (Y_{ij})$ be Gaussian random matrices in $\mathbb{R}^{N_1\times N_2}$ such that*

$$\begin{cases}\mathbb{E}[X_{ij}X_{ik}] \le \mathbb{E}[Y_{ij}Y_{ik}] & \textit{for all } i,j,k\\ \mathbb{E}[X_{ij}X_{lk}] \ge \mathbb{E}[Y_{ij}Y_{lk}] & \textit{for all } i\neq l \textit{ and } j,k\\ \mathbb{E}[X_{ij}^2] = \mathbb{E}[X_{ij}^2] & \textit{for all } i,j\end{cases}$$

Then

$$\mathbb{E}[\min_i \max_j Y_{ij}] \le \mathbb{E}[\min_i \max_j X_{ij}]\,.$$

The next two lemmas follow from applications of these Comparison Theorems. We prove them in more generality than necessary for the current work because both lemmas are interesting in their own right. Let $\|\cdot\|_p$ be any norm on $D_1 \times D_2$ matrices and let $\|\cdot\|_d$ be its associated dual norm (See Section 1.3). Again without loss of generality, we assume $D_1 \le D_2$. This first lemma is now a straightforward consequence of Slepian's Lemma

Lemma 7 *Let $\Delta > 0$, $\sigma_d \ge \sup_{\|Z\|_d=1}\|Z\|_F$, and let g be a Gaussian random vector in $\mathbb{R}^M$. Let $G, G_1, \ldots, G_M$ be sampled i.i.d. from $\mathfrak{G}(D_1,D_2)$. Then*

$$\mathbb{E}\left[\sup_{\|v\|_{\ell_2}=1}\ \sup_{\|Y\|_d=1} \Delta\langle g,v\rangle + \left\langle \sum_{i=1}^M v_iG_i, Y\right\rangle\right] \le \mathbb{E}[\|G\|_p] + \sqrt{M(\Delta^2+\sigma_d^2)}\,.$$

Proof We follow the strategy used to prove Theorem 3.20 in [20]. Let $G, G_1, \ldots, G_M$ be sampled i.i.d. from $\mathfrak{G}(D_1, D_2)$ and $g \in \mathbb{R}^M$ be a Gaussian random vector and let γ be a zero-mean, unit-variance Gaussian random variable. For $v \in \mathbb{R}^M$ and $Y \in \mathbb{R}^{D_1 \times D_2}$ define

$$\begin{aligned} Q_L(v, Y) &= \Delta \langle g, v \rangle + \left\langle \sum_{i=1}^M v_i G_i, Y \right\rangle + \sigma_d \gamma \\ Q_R(v, Y) &= \langle G, Y \rangle + \sqrt{\Delta^2 + \sigma_d^2} \langle g, v \rangle \,. \end{aligned}$$

Now observe that for any M-dimensional unit vectors v, $\hat{v}$ and any $D_1 \times D_2$ matrices Y, $\hat{Y}$ with dual norm 1

$$\begin{aligned} & \mathbb{E}[Q_L(v, Y) Q_L(\hat{v}, \hat{Y})] - \mathbb{E}[Q_R(v, Y) Q_R(\hat{v}, \hat{Y})] \\ =& \Delta^2 \langle v, \hat{v} \rangle + \langle v, \hat{v} \rangle \langle Y, \hat{Y} \rangle + \sigma_d^2 - \langle Y, \hat{Y} \rangle - (\Delta^2 + \sigma_d^2) \langle v, \hat{v} \rangle \\ =& (\sigma_d^2 - \langle Y, \hat{Y} \rangle)(1 - \langle v, \hat{v} \rangle) \,. \end{aligned}$$

The first quantity is always non-negative because $\langle Y, \hat{Y} \rangle \le \max(\|Y\|_F^2, \|\hat{Y}\|_F^2) \le \sigma_d^2$ by definition. The difference in expectation is thus equal to zero if $v = \hat{v}$ and is greater than or equal to zero if $v \neq \hat{v}$. Hence, by Slepian's Lemma and a compactness argument (see Proposition 1 in the Appendix),

$$\mathbb{E}\left[\sup_{\|v\|_{\ell_2}=1} \sup_{\|Y\|_d=1} Q_L(v, Y) \right] \le \mathbb{E}\left[\sup_{\|v\|_{\ell_2}=1} \sup_{\|Y\|_d=1} Q_R(v, Y) \right]$$

which proves the lemma.

The following lemma can be proved in a similar fashion

Lemma 8 *Let $\|\cdot\|_p$ be a norm on $\mathbb{R}^{D_1 \times D_2}$ with dual norm $\|\cdot\|_d$ and let $\|\cdot\|_b$ be a norm on $\mathbb{R}^{D_3 \times D_4}$. Let $\sigma_d \ge \sup_{\|Z\|_d=1} \|Z\|_F$. Let g be a Gaussian random vector in $\mathbb{R}^M$. Let $G_0, G_1, \ldots, G_M$ be sampled i.i.d. from $\mathfrak{G}(D_1, D_2)$ and $G_1', \ldots, G_M'$ be sampled i.i.d. from $\mathfrak{G}(D_3, D_4)$. Then*

$$\begin{aligned} & \mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1} \inf_{\|Y\|_b=1} \sup_{\|Z\|_d=1} \left\langle \sum_{i=1}^M v_i G_i, Z \right\rangle + \left\langle \sum_{i=1}^M v_i G_i', Y \right\rangle \right] \\ & \qquad \ge \mathbb{E}\left[\|G_0\|_p \right] - \mathbb{E}\left[\sup_{\|v\|_{\ell_2}=1} \sup_{\|Y\|_b=1} \sigma_d \langle g, v \rangle + \left\langle \sum_{i=1}^M v_i G_i', Y \right\rangle \right] . \end{aligned}$$

Proof Let η be a normally distributed random variable and define the functionals

$$\begin{aligned} P_L(v, Y, Z) &= \left\langle \sum_{i=1}^M v_i G_i, Z \right\rangle + \left\langle \sum_{i=1}^M v_i G_i', Y \right\rangle + \eta \sigma_d \\ P_R(v, Y, Z) &= \langle G_0, Z \rangle + \sigma_d \langle g, v \rangle + \left\langle \sum_{i=1}^M v_i G_i', Y \right\rangle . \end{aligned}$$

Let v and $\hat{v}$ be unit vectors in $\mathbb{R}^M$, Y and $\hat{Y}$ be $D_3 \times D_4$ matrices with $\|Y\|_b = \|\hat{Y}\|_b = 1$, and Z and $\hat{Z}$ be $D_1 \times D_2$ matrices with $\|Z\|_d = \|\hat{Z}\|_d = 1$. Then we have

$$\begin{aligned}
&\mathbb{E}[P_L(v,Y,Z)P_L(\hat{v},\hat{Y},\hat{Z})] - \mathbb{E}[P_R(v,Y,Z)P_L(\hat{v},\hat{Y},\hat{Z})]\\
=&\langle v,\hat{v}\rangle\langle Z,\hat{Z}\rangle + \langle v,\hat{v}\rangle\langle Y,\hat{Y}\rangle + \sigma_d^2 - \langle Z,\hat{Z}\rangle - \sigma_d^2\langle v,\hat{v}\rangle - \langle v,\hat{v}\rangle\langle Y,\hat{Y}\rangle\\
=&(\sigma_d^2 - \langle Z,\hat{Z}\rangle)(1-\langle v,\hat{v}\rangle)\,.
\end{aligned}$$

Just as was the case in the proof of Lemma 7, the first quantity is always nonnegative. Hence, the difference in expectations is greater than or equal to zero and equal to zero when $v = \hat{v}$ or $Y = \hat{Y}$. Hence, by Gordan's Lemma and a compactness argument,

$$\mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1}\inf_{\|Y\|_b=1}\sup_{\|Z\|_d=1} Q_L(v,Y,Z)\right] \geq \mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1}\inf_{\|Y\|_b=1}\sup_{\|Z\|_d=1} Q_R(v,Y,Z)\right]$$

completing the proof.

Together with Lemmas 7 and 8, we can prove Lemma 3.

Proof (of Lemma 3) For $i = 0,\dots,M$, let $G_i \in \mathfrak{G}(\gamma n_2, n_2)$. Since the Gaussian distribution and the nuclear norm are rotationally invariant, we may perform a change of coordinates such that

$$P = \begin{bmatrix} I_r & 0\\ 0 & 0\end{bmatrix} \quad\text{and}\quad Q = \begin{bmatrix} I_r & 0\\ 0 & 0\end{bmatrix}$$

where I_r denotes the $r \times r$ identity matrix. Under such a transformation, we may make the identifications

$$PG_iQ = \begin{bmatrix} G_i' & 0\\ 0 & 0\end{bmatrix} \quad\text{and}\quad (I-P)G_i(I-Q) = \begin{bmatrix} 0 & 0\\ 0 & \hat{G}_i\end{bmatrix}$$

$\hat{G}_i \in \mathfrak{G}((1-\beta)\gamma n_2, (1-\beta\gamma)n_2)$ and $G_i' \in \mathfrak{G}(\gamma\beta n_2, \gamma\beta n_2)$. Note that G_i' and $\hat{G}_i$ are independent for all i.

Recall that the dual norm of the nuclear norm is the operator norm. Moreover, if W is a $D_1 \times D_2$ matrix with $D_1 \leq D_2$, then $\sup_{\|Z\|=1}\|Z\|_F = \sqrt{D_1}$. We now apply Lemmas 8 and 7 to find

$$\begin{aligned}
&\mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1}\left\|(I-P)\left(\sum_{i=1}^M v_iG_i\right)(I-Q)\right\|_* - \left\|P\left(\sum_{i=1}^M v_iG_i\right)Q\right\|_*\right]\\
&= \mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1}\left\|\sum_{i=1}^M v_i\hat{G}_i\right\|_* - \left\|\sum_{i=1}^M v_iG_i'\right\|_*\right]\\
&= \mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1}\inf_{\|Y\|=1}\sup_{\|Z\|=1}\left\langle\sum_{i=1}^M v_i\hat{G}_i, Z\right\rangle + \left\langle\sum_{i=1}^M v_iG_i', Y\right\rangle\right]\\
&\geq \mathbb{E}\left[\|\hat{G}_0\|_*\right] - \mathbb{E}\left[\sup_{\|v\|_{\ell_2}=1}\sup_{\|Y\|=1}\sqrt{(1-\beta)\gamma n_2}\langle g,v\rangle + \left\langle\sum_{i=1}^M v_iG_i', Y\right\rangle\right]\\
&\geq \mathbb{E}\left[\|\hat{G}_0\|_*\right] - \mathbb{E}\left[\|G_0'\|_*\right] - \sqrt{M}\sqrt{(1-\beta)\gamma n_2 + \gamma\beta n_2}\\
&= \mathbb{E}\left[\|\hat{G}_0\|_*\right] - \mathbb{E}\left[\|G_0'\|_*\right] - \sqrt{M}\sqrt{\gamma n_2}
\end{aligned}$$

where the first inequality follows from Lemma 8, and the second inequality follows from Lemma 7. We use $\sigma_d = \sqrt{(1-\beta)\gamma n_2}$ when applying Lemma 8 and $\sigma_d = \sqrt{\beta\gamma n_2}$ when applying Lemma 7.

Now we only need to plug in the asymptotic expected value of the nuclear norm which may be asymptotically approximated using a classical result of Marčenko and Pastur. Let G be sampled from $\mathfrak{G}(D_1, D_2)$. Then

$$\mathbb{E}\|G\|_* = \varphi\left(\frac{D_1}{D_2}\right) D_2^{3/2} + q(D_2) \tag{20}$$

where $\varphi(\cdot)$ is is defined by the integral in Equation (3) (see, e.g., [25,4]) and $q(D_2)/D_2^{3/2} = o(1)$. Note that $\varphi(1)$ can be computed in closed form:

$$\varphi(1) = \frac{1}{2\pi}\int_0^4 \sqrt{4-t}\, dt = \frac{8}{3\pi} \approx 0.85\,.$$

Plugging these values in with the appropriate dimensions completes the proof.

Proof (of Lemma 5) This lemma immediately follows from applying Lemma 7 with $\Delta = 0$ and from the calculations at the end of the proof above. It is also an immediate consequence of Lemma 3.21 from [20].

3 Numerical Experiments

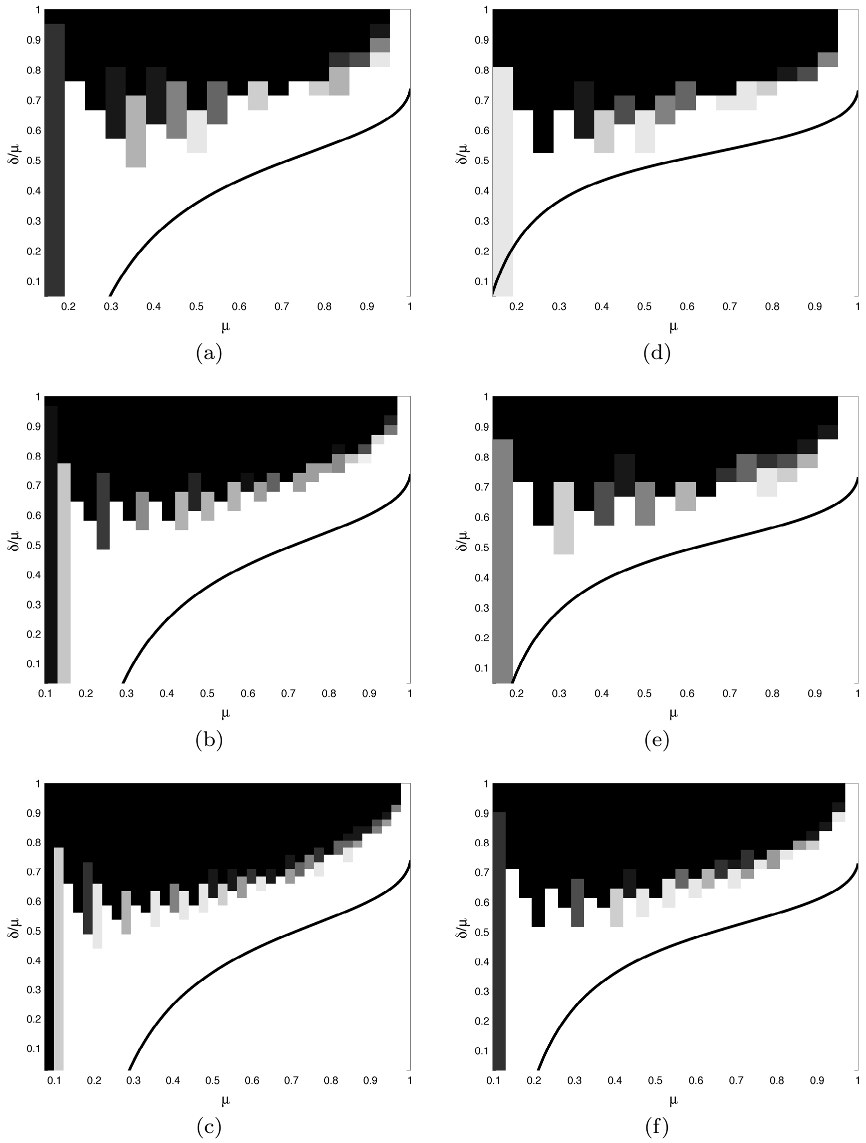
We now show that these asymptotic estimates hold even for moderately sized matrices. We conducted a series of experiments for a variety of the matrix sizes n, ranks r, aspect ratios γ, and numbers of measurements m. As in the previous section, we let $\beta = \frac{r}{n}$ and $\mu = \frac{m}{n^2}$. For a fixed n, we constructed random recovery scenarios for low-rank $\gamma n \times n$ matrices. For each n, we varied μ between 0 and 1 where the matrix is completely determined. For a fixed n, γ, and μ, we generated all possible ranks such that $\gamma\beta(1-\gamma-\beta\gamma) \leq \mu$. This cutoff was chosen because the quantity on the left hand side is the number of parameters of a rank r matrix of size $\gamma n \times n$. Beyond this value of β, there would be an infinite set of matrices of rank r satisfying the m equations.

For each (n, μ, β, γ) tuple, we repeated the following procedure 10 times. A matrix of rank r was generated by choosing two random factors Y_L and Y_R (of size $\gamma n \times r$ and $n \times r$ respectively) with i.i.d. random entries and setting $Y_0 = Y_L Y_R^*$. A matrix $\mathbf{A}$ was sampled from the Gaussian ensemble with m rows and γn^2 columns. Then the nuclear norm minimization

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \mathbf{A}\,\mathrm{vec}\,X = \mathbf{A}\,\mathrm{vec}\,Y_0 \end{array}$$

was solved using the freely available software SeDuMi [34] using the semidefinite programming formulation described in [29]. On a 2.0 GHz Laptop, each semidefinite program could be solved in less than two minutes for 40×40 dimensional X. We declared Y_0 to be recovered if $\|X - Y_0\|_F/\|Y_0\|_F < 10^{-3}$.

Figure 3 displays the results of these experiments for six settings of the parameters. The color of the cell in the figures reflects the empirical recovery rate of the 10 runs (scaled between 0 and 1). White denotes perfect recovery in all experiments, and black denotes failure for all experiments. We observe that the Weak Bound falls completely within the white region in all of our experiments and is an good approximation of the boundary between success and failure for large μ.



(a) (d)

(b) (e)

(c) (f)

Fig. 3 Random rank recovery experiments for matrices of size (a) 20 × 20, (b) 30 × 30, (c) 40 × 40, (d) 20 × 30, (e) 20 × 40, and (f) 30 × 40. The color of each cell reflects the empirical recovery rate. White denotes perfect recovery in all experiments, and black denotes failure for all experiments. The axes are the same as in FIgure 1, and the y-axis denotes the ratio of the model size to the number of measurements. In all frames, we plot the Weak Bound (4), showing that the predicted recovery regions are contained within the empirical regions, and the boundary between success and failure is well approximated for large values of μ.

4 Discussion and Future Work

Future work should investigate if the probabilistic analysis that provides the bounds in Theorems 1 and 2 can be further tightened at all. There are two particular regions where the bounds can be improved. First, when $\beta = 0$, μ should also equal zero. However, in our Weak Bound, $\gamma = 1$ and $\beta = 0$ tells us that μ must be greater than or equal to 0.2795. In order to provide estimates of the behavior for small values of μ, we will need to find a different lower bound than (13). When μ is small, M in (13) is very large causing the bound on the expected value to be negative. This suggests that a different parametrization of the null space of $\mathcal{A}$ could be the key to a better bound for small values of β. It also may be fruitful to investigate if some of the techniques in [13,14] on neighborly polytopes can be generalized to yield tighter approximations of the recovery region. It would also be of interest to construct a *necessary* condition, parallel to the sufficient condition of Section 2.1, and apply a similar probabilistic analysis to yield an upper bound for the phase transition.

The comparison theorem techniques in this paper add a novel set of tools to the behavior of the nuclear norm heuristic, and they may be very useful in the study of other rank minimization scenarios. For example, the structured problems that arise in Control Theory can be formulated in the form of (1) with a very structured $\mathcal{A}$ operator (see, e.g., [30]). It would be of interest to see if these structured problems can also be analyzed within the null space framework. Using the particular structure of the null space of $\mathcal{A}$ in these specialized problems may provide sharper bounds for these cases. Along these lines, a problem of great interest is the Matrix Completion Problem where we would like to reconstruct a low-rank matrix from a small subset of its entries. In this scenario, the operator $\mathcal{A}$ reveals a few of the entries of the unknown low-rank matrix, and the null space of $\mathcal{A}$ is simply the set of matrices that are zero in the specified set. The Gaussian comparison theorems studied above cannot be directly applied to this problem, but it is possible that generalizations exist that could be applied to the Matrix Completion problem and could possibly tighten the bounds provided in [8].

References

1. Ames, B.P.W., Vavasis, S.A.: Nuclear norm minimization for the planted clique and biclique problems (2009). Submitted to Mathematical Programming. Preprint available at `http://arxiv.org/abs/0901.3348v1`
2. Amit, Y., Fink, M., Srebro, N., Ullman, S.: Uncovering shared structures in multiclass classification. In: Proceedings of the International Conference of Machine Learning (2007)
3. Argyriou, A., Micchelli, C.A., Pontil, M.: Convex multi-task feature learning. Machine Learning (2008). Published online first at `http://www.springerlink.com/`
4. Bai, Z.D.: Methodologies in spectral analysis of large dimensional random matrices. Statistica Sinica **9**(3), 611–661 (1999)
5. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constructive Approximation (2008). To Appear. Preprint available at `http://dsp.rice.edu/cs/jlcs-v03.pdf`
6. Beck, C., D'Andrea, R.: Computational study and comparisons of LFT reducibility methods. In: Proceedings of the American Control Conference (1998)
7. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion (2008). To appear in *SIAM J. on Optimization.* Preprint available at `http://arxiv.org/abs/0810.3286`

8. Candès, E., Recht, B.: Exact matrix completion via convex optimization. Foundations of Computational Mathematics **9**(6), 717–772 (2009)
9. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory **52**(2), 489–509 (2006)
10. Candès, E.J., Tao, T.: Decoding by linear programming. IEEE Transactions on Information Theory **51**(12), 4203–4215 (2005)
11. Donoho, D.: High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. Discrete and Computational Geometry **35**(4), 617–652 (2006)
12. Donoho, D., Huo, X.: Uncertainty principles and ideal atomic decomposition. IEEE Transactions on Information Theory **47**(7), 2845–2862 (2001)
13. Donoho, D.L., Tanner, J.: Neighborliness of randomly projected simplices in high dimensions. Proc. Natl. Acad. Sci. USA **102**(27), 9452–9457 (2005)
14. Donoho, D.L., Tanner, J.: Sparse nonnegative solution of underdetermined linear equations by linear programming. Proc. Natl. Acad. Sci. USA **102**(27), 9446–9451 (2005)
15. Fazel, M.: Matrix rank minimization with applications. Ph.D. thesis, Stanford University (2002)
16. Fazel, M., Hindi, H., Boyd, S.: A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American Control Conference (2001)
17. El Ghaoui, L., Gahinet, P.: Rank minimization under LMI constraints: A framework for output feedback problems. In: Proceedings of the European Control Conference (1993)
18. Gordan, Y.: Some inequalities for Gaussian processes and applications. Israel Journal of Math **50**, 265–289 (1985)
19. Gordan, Y.: Gaussian processes and almost spherical sections of convex bodies. Annals of Probability **16**, 180–188 (1988)
20. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer-Verlag, Berlin (1991)
21. Lee, K., Bresler, Y.: Efficient and guaranteed rank minimization by atomic decomposition. In: IEEE International Symposium on Information Theory (2009)
22. Linial, N., London, E., Rabinovich, Y.: The geometry of graphs and some of its algorithmic applications. Combinatorica **15**, 215–245 (1995)
23. Liu, Z., Vandenberghe, L.: Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications **31**(3), 1235–1256 (2009)
24. Ma, S., Goldfarb, D., Chen, L.: Fixed point and Bregman iterative methods for matrix rank minimization (2008). Preprint available at `http://www.optimization-online.org/DB_HTML/2008/11/2151.html`
25. Marčenko, V.A., Pastur, L.A.: Distributions of eigenvalues for some sets of random matrices. Math. USSR-Sbornik **1**, 457–483 (1967)
26. Meka, R., Jain, P., Caramanis, C., Dhillon, I.S.: Rank minimization via online learning. In: Proceedings of the International Conference on Machine Learning (2008)
27. Mesbahi, M., Papavassilopoulos, G.P.: On the rank minimization problem over a positive semidefinite linear matrix inequality. IEEE Transactions on Automatic Control **42**(2), 239–243 (1997)
28. Parrilo, P.A., Khatri, S.: On cone-invariant linear matrix inequalities. IEEE Trans. Automat. Control **45**(8), 1558–1563 (2000)
29. Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. SIAM Review (2007). To appear. Preprint Available at `http://pages.cs.wisc.edu/~brecht/publications.html`
30. Recht, B., Xu, W., Hassibi, B.: Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization. In: Proceedings of the 47th IEEE Conference on Decision and Control (2008)
31. Rennie, J.D.M., Srebro, N.: Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of the International Conference of Machine Learning (2005)
32. Slepian, D.: The one-sided barrier problem for Gaussian noise. Bell System Technical Journal **41**, 463–501 (1962)
33. Stojnic, M., Xu, W., Hassibi, B.: Compressed sensing - probabilistic analysis of a null-space characterization. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2008)
34. Sturm, J.F.: Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Optimization Methods and Software **11-12**, 625–653 (1999)

35. Szarek, S.J.: Metric entropy of homogeneous spaces. In: Quantum probability (Gdańsk, 1997), *Banach Center Publ.*, vol. 43, pp. 395–410. Polish Acad. Sci., Warsaw (1998). Preprint available at `arXiv:math/9701213v1`
36. Weinberger, K.Q., Saul, L.K.: Unsupervised learning of image manifolds by semidefinite programming. International Journal of Computer Vision **70**(1), 77–90 (2006)
37. Yuan, M., Ekici, A., Lu, Z., Monteiro, R.: Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society: Series B **69**, 329–346 (2007)
38. Zhang, Y.: A simple proof for recoverability of ℓ_1 minimization: go over or under? Tech. Rep. TR05-09, Rice CAAM Department (2005)

A Appendix

A.1 Proof of Theorem 3

We begin by proving the converse. Assume the condition of part 1 is violated, i.e., there exists some Y, such that $\mathcal{A}(Y) = 0$, $Y = Y_1 + Y_2$, $\operatorname{rank}(Y_2) > \operatorname{rank}(Y_1) = r$, yet $\|Y_1\|_* > \|Y_2\|_*$. Now take $X_0 = Y_1$ and $b = \mathcal{A}(X_0)$. Clearly, $\mathcal{A}(-Y_2) = b$ (since Y is in the null space) and so we have found a matrix of higher rank, but lower nuclear norm.

For the other direction, assume the null space property of part 1 holds. Note that this property implies that there are no matrices in the kernel of $\mathcal{A}$ with rank less than or equal to r. To see this, let W have rank less than or equal to r and choose an appropriate basis such that

$$W = \begin{bmatrix} W_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

where W_{11} is $r \times r$. Let Δ be any matrix such that

$$\Delta = \begin{bmatrix} 0 & \Delta_{12} \\ \Delta_{21} & 0 \end{bmatrix} .$$

in this same basis with Δ_{12} and Δ_{21} both non-zero. Let I denote the $r \times r$ identity matrix. Then there exists an arbitrarily small $\delta > 0$ such that $W_{11} + \delta I$ is invertible. For such a δ,

$$S = \begin{bmatrix} W_{11} + \delta I & \delta \Delta_{12} \\ \delta \Delta_{21} & \delta^2 \Delta_{21}(W_{11} + \delta I)^{-1} \Delta_{12} \end{bmatrix}$$

has rank r and $W - S$ has rank strictly greater than r. Certainly $S + (W - S)$ is in the null space of $\mathcal{A}$, but, for δ sufficiently small, we will have $\|S\|_* > \|W - S\|_*$. This violates our assumption about the null space of $\mathcal{A}$.

To complete the proof, we now proceed again by contradiction. Let X_0 denote the minimum rank solution and X_* denote the minimum nuclear norm solution, and suppose that $X_0 \neq X_*$. Then, in an appropriate basis, we may write

$$X_0 = \begin{bmatrix} X_{11} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad X_* - X_0 = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \tag{21}$$

where X_{11} and Y_{11} are $r \times r$, and X_{11} is a diagonal matrix with nonnegative diagonal entries. Note that either Y_{22} is non-zero or both Y_{12} and Y_{21} are nonzero as otherwise, $X_* - X_0$ would have rank less than or equal to r. Therefore, it is always possible to find an $\epsilon > 0$ such that $Y_{11} - \epsilon I$ has full rank and $Y_{22} - Y_{21}(Y_{11} - \epsilon I)^{-1} Y_{12} \neq 0$. Define the matrices

$$Z_1 := \begin{bmatrix} \epsilon I & 0 \\ 0 & Y_{22} - Y_{21}(Y_{11} + \epsilon I)^{-1} Y_{12} \end{bmatrix}$$
$$Z_2 := \begin{bmatrix} Y_{11} - \epsilon I & Y_{12} \\ Y_{21} & Y_{21}(Y_{11} - \epsilon I)^{-1} Y_{12} \end{bmatrix}$$

with I denoting the $r \times r$ identity matrix. Then $Z_1 + Z_2 = X_* - X_0$ is nonzero and lies in the null space of $\mathcal{A}$. Moreover, since $Y_1 - \epsilon I$ has full rank, the rank of Z_2 is r. On the other hand,

$Y_{22} - Y_{21}(Y_{11}+\epsilon I)^{-1}Y_{12} \neq 0$ implies that the rank of Z_1 is strictly greater than r. So by the assumption of Condition 1 of the theorem, we must have that $\|Z_1\|_* > \|Z_2\|_*$. It now follows that

$$\begin{aligned}\|X_*\|_* &= \|X_0 + X_* - X_0\|_* \\ &\geq \|X_0 + Z_1\|_* - \|Z_2\|_* \\ &= \|X_{11}+\epsilon I\|_* + \left\|Y_{22} - Y_{21}(Y_{11}+\epsilon I)^{-1}Y_{12}\right\|_* - \|Z_2\|_* \\ &= \|X_0\|_* + \|\epsilon I\|_* + \left\|Y_{22} - Y_{21}(Y_{11}+\epsilon I)^{-1}Y_{12}\right\|_* - \|Z_2\|_* \\ &= \|X_0\|_* + \|Z_1\|_* - \|Z_2\|_* \\ &> \|X_0\|_* \,.\end{aligned}$$

Here, the first inequality follows from the triangle inequality and the definitions of Z_1 and Z_2. The next equality holds by the partitioning of Equation (21) and because the nuclear norm of a block diagonal matrix is equal to the sum of the nuclear norms of each block. The subsequent inequality follows because the nuclear norm of the sum of two nonnegative diagonal matrices is equal to the sum of the nuclear norms of the individual summands. The next line again follows because Z_1 is block diagonal. The final inequality is strict and follows because, as discussed above, $\|Z_1\|_* > \|Z_2\|_*$. But X_* is the minimum nuclear norm solution, so we have arrived at a contradiction. Consequently, this means that X_0 must equal X_*.

A.2 Lipschitz Constants of F_I and F_S

We begin with the proof of Lemma 6 and then use this to estimate the Lipschitz constant in Lemma 4.

Proof (of Lemma 6) Note that the function F_S is convex as we can write as a supremum of a collection of convex functions

$$F_S(X_1,\dots,X_M) = \sup_{\|v\|_{\ell_2}=1}\ \sup_{\|Z\|<1} \langle \sum_{i=1}^{M} v_i X_i, Z\rangle \,. \tag{22}$$

The Lipschitz constant L is bounded above by the maximal norm of a subgradient of this convex function. That is,

$$L \leq \sup_{\bar{X}}\ \sup_{\bar{Z}\in\partial F_S(\bar{X})} \left(\sum_{i=1}^{M} \|Z_i\|_F^2\right)^{1/2} \,.$$

where $\bar{X} := (X_1,\dots,X_M)$ and $\bar{Z} := (Z_1,\dots,Z_M)$. Now, by (22), a subgradient of F_S at $\bar{X}$ is given of the form $(v_1 Z, v_2 Z, \dots, v_M Z)$ where v has norm 1 and Z has operator norm 1. For any such subgradient

$$\sum_{i=1}^{M} \|v_i Z\|_F^2 = \|Z\|_F^2 \leq D_1$$

bounding the Lipschitz constant as desired.

Proof (of Lemma 4) For $i=1,\dots,M$, let $X_i, \hat{X}_i \in \mathbb{R}^{D_1\times D_2}$, and $Y_i, \hat{Y}_i \in \mathbb{R}^{D_3\times D_4}$. Let

$$w^* = \arg\min_{\|w\|_{\ell_2}=1} \|\sum_{i=1}^{M} w_i \hat{X}_i\|_* - \|\sum_{i=1}^{M} w_i \hat{Y}_i\|_* \,.$$

Then we have that

$$
\begin{aligned}
&F_I(X_1,\dots,X_M,Y_1,\dots,Y_M) - F_I(\hat{X}_1,\dots,\hat{X}_M,\hat{Y}_1,\dots,\hat{Y}_M)\\
=&\left(\inf_{\|v\|_{\ell_2}=1}\|\sum_{i=1}^M v_i X_i\|_* - \|\sum_{i=1}^M v_i Y_i\|_*\right) - \left(\inf_{\|w\|_{\ell_2}=1}\|\sum_{i=1}^M w_i \hat{X}_i\|_* - \|\sum_{i=1}^M w_i \hat{Y}_i\|_*\right)\\
\leq&\|\sum_{i=1}^M w_i^* X_i\|_* - \|\sum_{i=1}^M w_i^* Y_i\|_* - \|\sum_{i=1}^M w_i^* \hat{X}_i\|_* + \|\sum_{i=1}^M w_i^* \hat{Y}_i\|_*\\
\leq&\|\sum_{i=1}^M w_i^*(X_i-\hat{X}_i)\|_* + \|\sum_{i=1}^M w_i^*(Y_i-\hat{Y}_i)\|_*\\
\leq&\sup_{\|w\|_{\ell_2}=1}\|\sum_{i=1}^M w_i(X_i-\hat{X}_i)\|_* + \|\sum_{i=1}^M w_i(Y_i-\hat{Y}_i)\|_* = \sup_{\|w\|_{\ell_2}=1}\|\sum_{i=1}^M w_i\tilde{X}_i\|_* + \|\sum_{i=1}^M w_i\tilde{Y}_i\|_*
\end{aligned}
$$

where $\tilde{X}_i = X_i - \hat{X}_i$ and $\tilde{Y}_i = Y_i - \hat{Y}_i$. This last expression is a convex function of $\tilde{X}_i$ and $\tilde{Y}_i$,

$$
\sup_{\|w\|_{\ell_2}=1}\|\sum_{i=1}^M w_i\tilde{X}_i\|_* + \|\sum_{i=1}^M w_i\tilde{Y}_i\|_* = \sup_{\|w\|_{\ell_2}=1}\sup_{\|Z_X\|<1}\sup_{\|Z_Y\|<1}\langle\sum_{i=1}^M w_i\tilde{X}_i, Z_X\rangle + \langle\sum_{i=1}^M w_i\tilde{Y}_i Z_Y\rangle
$$

with Z_X $D_1 \times D_2$ and Z_Y $D_3 \times D_4$. Using an identical argument as the one presented in the proof of Lemma 6, we have that a subgradient of this expression is of the form

$$
(w_1 Z_X, w_2 Z_X, \dots, w_M Z_X, w_1 Z_Y, w_2 Z_Y, \dots, w_M Z_Y)
$$

where w has norm 1 and Z_X and Z_Y have operator norms 1, and thus

$$
\sum_{i=1}^M \|w_i Z_X\|_F^2 + \|w_i Z_Y\|_F^2 = \|Z_X\|_F^2 + \|Z_Y\|_F^2 \leq D_1 + D_3
$$

completing the proof.

A.3 Compactness Argument for Comparison Theorems

Proposition 1 *Let Ω be a compact metric space with distance function ρ. Suppose that f and g are real-valued function on Ω such that f is continuous and for any finite subset $X \subset \Omega$*

$$
\max_{x\in X} f(x) \leq \max_{x\in X} g(x)\,.
$$

Then

$$
\sup_{x\in\Omega} f(x) \leq \sup_{x\in\Omega} g(x)\,.
$$

Proof Let $\epsilon > 0$. Since f is continuous and Ω is compact, f is uniformly continuous on Ω. That is, there exists a $\delta > 0$ such that for all $x, y \in \Omega$, $\rho(x,y) < \delta$ implies $|f(x) - f(y)| < \epsilon$. Let X_δ be a δ-net for Ω. Then, for any $x \in \Omega$, there is a y in the δ-net with $\rho(x,y) < \delta$ and hence

$$
f(x) \leq f(y) + \epsilon \leq \sup_{z\in X_\delta} f(z) + \epsilon \leq \sup_{z\in X_\delta} g(z) + \epsilon \leq \sup_{z\in\Omega} g(z) + \epsilon\,.
$$

Since this holds for all $x \in \Omega$ and $\epsilon > 0$, this completes the proof.