Null Space Conditions and Thresholds for Rank Minimization

Benjamin Recht*, Weiyu Xu†, and Babak Hassibi‡

April 21, 2009

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 constraints. Our starting point is the characterization of a necessary and sufficient condition that determines when this heuristic finds the minimum rank solution. We then obtain conditions, as a function of the matrix dimensions and rank and the number of constraints, such that our conditions for success are satisfied for almost all linear constraint sets 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.

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

*Center for the Mathematics of Information, California Institute of Technology, 1200 E California Blvd, Pasadena, CA brecht@ist.caltech.edu

†Electrical Engineering, California Institute of Technology, 1200 E California Blvd, Pasadena, CA weiyu@systems.caltech.edu

‡Electrical Engineering, California Institute of Technology, 1200 E California Blvd, Pasadena, CA bhassibi@systems.caltech.edu

a matrix—equal to the sum of the singular values—can be optimized in polynomial time. This initial 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 [9], 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, 14, 16], and dimensionality reduction [22, 37, 38].

Numerical experiments in [29] suggested that the nuclear norm heuristic significantly outperformed 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 constrained the values of a subset of the entries of the matrix [9].

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 present 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 is akin to the one in compressed sensing [33], first reported in [30]. The condition characterizes a particular property of the null-space of the linear map which defines the affine space. 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 dimensions of the equality constraints are of appropriate size. 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

$$\begin{array}{ll} \text{minimize} & \text{rank}(X) \\ \text{subject to} & \mathcal{A}(X) = b\,. \end{array} \tag{1.1}$$

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 (1.1) coincides with the optimal solution of

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \mathcal{A}(X) = b\,. \end{array} \tag{1.2}$$

This optimization 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 affine rank minimization problem (1.1) by three dimensionless parameters that take values in $(0, 1]$: the *aspect ratio* γ, the *constraint ratio* μ, the *rank ratio* β. Without loss of generality, we will assume throughout that we are dealing with matrices with fewer rows than columns. The aspect ratio is such that the number of rows is equal to $n_1 = \gamma n_2$. The constraint ratio is the ratio of the number of constraints to the number of parameters needed to fully specify an $n_1 \times n_2$ matrix. That is, the number of measurements is equal to $\mu\gamma n_2^2$. Generically, in the case that $\mu \geq 1$, the linear system describing the constraints is overdetermined and hence the minimum rank solution can be found by least-squares. The rank ratio is the ratio of the number of rows to the rank of the matrix so that the rank is equal to $\beta n_1 = \beta\gamma n_2$. 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 $\beta(1 + \gamma - \beta\gamma)n_2^2$. We will focus our attention to determining for which triples (β, γ, μ) the problem (1.2) has the same optimal solution as the rank minimization problem (1.1).

Whenever $\mu < 1$, the null space of $\mathcal{A}$, that is the set of Y such that $\mathcal{A}(Y) = 0$, is not empty. Note that X is the unique optimal solution for (1.2) if and only if for every Y in the null-space of $\mathcal{A}$

$$\|X + Y\|_* > \|X\|_*\,. \tag{1.3}$$

The following theorem 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. Our first result is the following

Theorem 1.1 *Let X_0 be the optimal solution of (1.1) and assume that X_0 has rank $r < n_1/2$. Then*

1. *If 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

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

then X_0 is the unique minimizer of (1.2).

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.*

This result is of interest for multiple reasons. First, it gives a necessary and sufficient condition on the mapping $\mathcal{A}$ such that *all* sufficiently low rank X_0 are recoverable from (1.2). Second, as shown in [30], a variety of the rank minimization problems, including those with inequality and semidefinite cone constraints, can be reformulated in the form of (1.1). Finally, we now present a family of random equality constraints under which the nuclear norm heuristic succeeds with overwhelming probability. We prove both of the following two theorems by showing that $\mathcal{A}$ obeys the null-space criteria of Equation (1.3) and Theorem 1.1 respectively with overwhelming probability.

Note that for a linear map $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, we can always find an $m \times n_1 n_2$ matrix $\mathbf{A}$ such that

$$\mathcal{A}(X) = \mathbf{A} \operatorname{vec} X \,. \tag{1.4}$$

In the case where $\mathbf{A}$ has entries sampled independently from a zero-mean, unit-variance Gaussian distribution, then the null space characterization of Theorem 1.1 holds with overwhelming probability provided m is large enough. 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. We also denote $\mathfrak{G}(d, d)$ by $\mathfrak{G}(d)$.

In order to state our results, we need to define a function $\varphi : [0, 1] \to \mathbb{R}$ that specifies the asymptotic mean of the nuclear norm of a matrix sampled from $\mathfrak{G}(d_1, d_2)$ $(d_1 \leq d_2)$.

$$\varphi(\gamma) := \frac{1}{2\pi} \int_{(1-\sqrt{\gamma})^2}^{(1+\sqrt{\gamma})^2} \sqrt{\frac{(z - s_1)(s_2 - z)}{z}} dz \tag{1.5}$$

The origins of this formula will be described in Section 3.4. We can now state our main threshold theorems. The first result characterizes when a particular low-rank matrix can be recovered from a random linear system via nuclear norm minimization.

Theorem 1.2 (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 $\mathbf{A}$ is sampled from $\mathfrak{G}(\mu n_1 n_2, n_1 n_2)$. Then whenever

$$\mu \geq 1 - \left(\varphi\left(\frac{\gamma - \beta\gamma}{1 - \beta\gamma} \right) \frac{(1-\beta)^{3/2}}{\gamma} - \frac{8}{3\pi} \gamma^{1/2} \beta^{3/2} \right)^2 \tag{1.6}$$

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 (1.6), 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 (1.6) provides a lower-bound on the empirical phase transition observed in [29]. Note that this theorem only depends on the null-space of $\mathcal{A}$ being selected from the uniform distribution of subspaces. 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 (1.6). A particular case of interest is the case of square matrices ($\gamma = 1$). In this case, the Weak Bound (1.6) takes the elegant closed form:

$$\mu \geq 1 - \frac{64}{9\pi^2}\left((1-\beta)^{3/2} - \beta^{3/2}\right)^2 . \tag{1.7}$$

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

Theorem 1.3 (Strong Bound) *Let $\mathcal{A}$ be defined as in Theorem 1.2. Define the two functions*

$$f(\gamma, \beta, \epsilon) = \frac{\varphi\left(\frac{\gamma-\beta\gamma}{1-\beta\gamma}\right)\gamma^{-1}(1-\beta)^{3/2} - \frac{8}{3\pi}\gamma^{1/2}\beta^{3/2} - 4\epsilon\varphi(\gamma)}{1+4\epsilon} \tag{1.8}$$

$$g(\gamma, \beta, \epsilon) = \sqrt{2\beta\gamma(1+\gamma-\beta\gamma)\log\left(\frac{3\pi}{2\epsilon}\right)} . \tag{1.9}$$

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

$$\mu \geq 1 - \sup_{\substack{\epsilon>0 \\ f(\beta,\epsilon)-g(\beta,\epsilon)>0}} \left(f(\beta,\epsilon) - g(\beta,\epsilon)\right)^2 . \tag{1.10}$$

In particular, if β, γ, and μ satisfy (1.10), 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 plots the bound from Theorems 1.2 and 1.3 with $\gamma = 1$. We call (1.6) the *Weak Bound* because it is a condition that depends on the optimal solution of (1.1). On the other hand, we call (1.10) the *Strong Bound* as it guarantees the nuclear norm heuristic succeeds, *no matter what the optimal solution*, as long as the minimum of the rank minimization problem is sufficiently small. The Weak Bound is the only bound that can be tested experimentally, and, in Section 4, 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 in of itself and results in a much better bound than what was available from previous results (c.f., [29]).

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

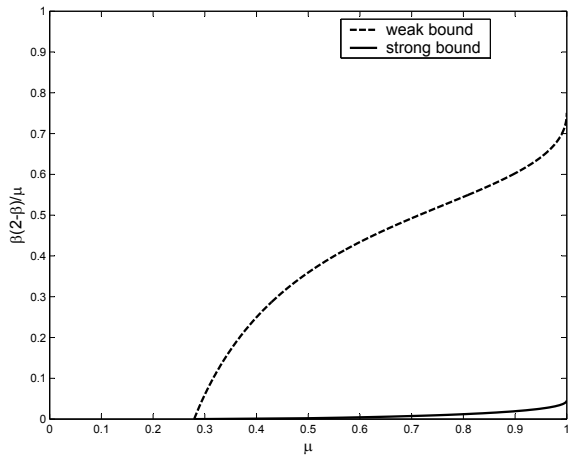


Figure 1: The Weak Bound (1.6) versus the Strong Bound (1.10).

with partial information [31] and Multi-task learning [3]. In control theory, problems in controller design [14, 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 [37].

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 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, 8, 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 matrix low rank matrices can be recovered from a sampling of on the order of $(n^{1.2}r)$ of the matrices entries [9] 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, 13, 33]. In these papers, the authors study properties of the null space of the linear operator that gives rise to the affine constraints. In [11, 12], the authors think of the constraint set as a k-neighborly polytope. It turns out that this characterization of the matrix A is in fact a necessary and sufficient condition for the ℓ_1 minimization to produce the sparest solution. Furthermore, using the results of [36], it can be shown that if the matrix A has i.i.d. zero-mean Gaussian entries with overwhelming probability it also constitutes a k-neighborly poly-tope. The precise relation between m and k in order for this to happen is characterized in [11] as well. It should also be noted that for a given value m i.e. for a given value of the constant α, the sparsity bound is significantly better in [11, 12] than in [10]. Furthermore, the values of sparsity thresholds obtained for different values of α in [11] approach the ones obtained by simulation as $n \longrightarrow \infty$. Our null-space criteria generalizes the concept of the same name in Compressed Sensing.

Unfortunately, the polyhedral analysis of Donoho and Tanner does not extend to the space of 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

$$\left\{ X \ : \ X = \begin{bmatrix} x & y \\ y & z \end{bmatrix}, \ \|X\|_* \leq 1 \right\}. \tag{1.11}$$

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.

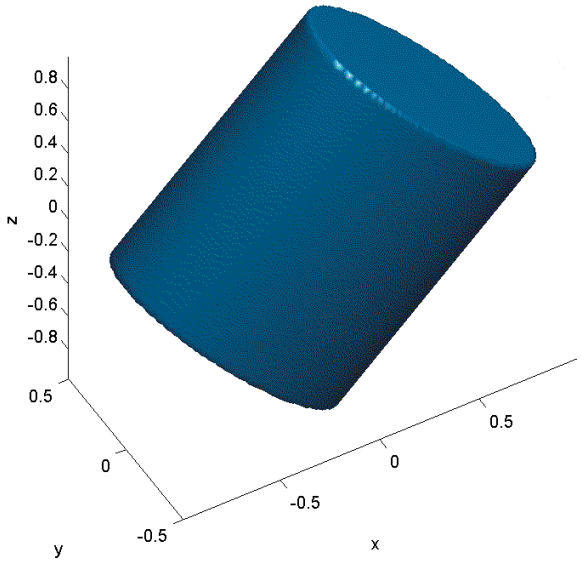


Figure 2: The unit ball of the nuclear norm. The figure depicts the set of all matrices of the form of equation (1.11) with nuclear norm less than one.

1.3 Notation and Preliminaries

For a rectangular matrix $X \in \mathbb{R}^{n_1 \times n_2}$, X^* denotes the transpose of X. $\mathrm{vec}(X)$ denotes the vector in $\mathbb{R}^{n_1 n_2}$ with the columns of X stacked on top of one and other.

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 $\langle X, Y \rangle := \mathrm{trace}(X^* Y) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} X_{ij} Y_{ij}$. 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:

$$||X|| \leq ||X||_F \leq ||X||_* \leq \sqrt{r}||X||_F \leq r||X||. \tag{1.12}$$

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 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.

2 Necessary and Sufficient Conditions

We first prove our necessary and sufficient condition for success of the nuclear norm heuristic. We will need the following two technical lemmas. The first is an easily verified fact.

Lemma 2.1 *Suppose X and Y are $n_1 \times n_2$ matrices such that $X^*Y = 0$ and $XY^* = 0$. Then $\|X + Y\|_* = \|X\|_* + \|Y\|_*$.*

Indeed, if $X^*Y = 0$ and $XY^* = 0$, we can find a coordinate system in which

$$X = \left\| \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} \right\|_* \quad \text{and} \quad Y = \left\| \begin{bmatrix} 0 & 0 \\ 0 & B \end{bmatrix} \right\|_*$$

from which the lemma trivially follows. The next Lemma allows us to exploit Lemma 2.1 in our proof.

Lemma 2.2 *Let X be an $n_1 \times n_2$ matrix with rank $r < \frac{n_1}{2}$ and Y be an arbitrary $n_1 \times n_2$ matrix. Let P_X^c and P_X^r be the matrices that project onto the column and row spaces of X respectively. Then if $P_X^c Y P_X^r$ has full rank, Y can be decomposed as*

$$Y = Y_1 + Y_2,$$

where Y_1 has rank r, and

$$\|X + Y_2\|_* = \|X\|_* + \|Y_2\|_*.$$

Proof Without loss of generality, we can write X as

$$X = \begin{bmatrix} X_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

where X_{11} is $r \times r$ and full rank. Accordingly, Y becomes

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

where Y_{11} is full rank since $P_X^r Y P_X^c$ is. The decomposition is now clearly

$$Y = \underbrace{\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{21}Y_{11}^{-1}Y_{12} \end{bmatrix}}_{Y_1} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & Y_{22} - Y_{21}Y_{11}^{-1}Y_{12} \end{bmatrix}}_{Y_2}.$$

That Y_1 has rank r follows from the fact that the rank of a block matrix is equal to the rank of a diagonal block plus the rank of its Schur complement (see, e.g., [19, §2.2]). That $\|X_1 + Y_2\|_* = \|X_1\|_* + \|Y_2\|*$ follows from Lemma 2.1. ■

We can now provide a proof of Theorem 1.1.

Proof We begin by proving the converse. Assume the condition of part 1 is violated, i.e., there exists some Y, such that $\mathcal{A}(Y) = 0$, $Y = Y_1 + Y_2$, $\text{rank}(Y_2) > \text{rank}(Y_1) = r$, 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 condition of part 1 holds. Now use Lemma 2.2 with $X = X_0$ and $Y = X_* - X_0$. That is, let P_X^c and P_X^r be the matrices that project onto the column and row spaces of X_0 respectively and assume that $P_{X_0}^c(X_* - X_0)P_{X_0}^r$ has full rank. Write $X_* - X_0 = Y_1 + Y_2$ where Y_1 has rank r and $\|X_0 + Y_2\|_* = \|X_0\|_* + \|Y_2\|_*$. Assume further that Y_2 has rank larger than r (recall $r < n/2$). We will consider the case where $P_{X_0}^c(X_* - X_0)P_{X_0}^r$ does not have full rank and/or Y_2 has rank less than or equal to r in the appendix. We now have:

$$\begin{aligned} \|X_*\|_* &= \|X_0 + X_* - X_0\|_* \\ &= \|X_0 + Y_1 + Y_2\|_* \\ &\geq \|X_0 + Y_2\|_* - \|Y_1\|_* \\ &= \|X_0\|_* + \|Y_2\|_* - \|Y_1\|_* \quad \text{by Lemma 2.2.} \end{aligned}$$

But $\mathcal{A}(Y_1 + Y_2) = 0$, so $\|Y_2\|_* - \|Y_1\|_*$ non-negative and therefore $\|X_*\|_* \geq \|X_0\|_*$. Since X_* is the minimum nuclear norm solution, implies that $X_0 = X_*$. ■

For the interested reader, the argument for the case where $P_{X_0}^r(X_* - X_0)P_{X_0}^c$ does not have full rank or Y_2 has rank less than or equal to r can be found in the Appendix.

3 Proofs of the Probabilistic Bounds

We now turn to the proofs of the probabilistic bounds (1.6) and (1.10). We first provide a sufficient condition which implies the necessary and sufficient null-space conditions. 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 sufficient conditions are met. This will require the introduction of two useful auxiliary functions whose actions on Gaussian processes are explored in Section 3.4.

3.1 Sufficient Condition for Null-space Characterizations

The following theorem gives us a new condition that implies our necessary and sufficient condition.

Theorem 3.1 *Let $\mathcal{A}$ be a linear map of $n_1 \times n_2$ matrices into $\mathbb{R}^m$. Suppose that for every Y in the null-space of $\mathcal{A}$ and any projection operators P and Q onto r-dimensional subspaces of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ respectively that*

$$\|(I - P)Y(I - Q)\|_* \geq \|PYQ\|_* . \tag{3.1}$$

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 $\mathcal{A}$. In particular, if (3.1) holds for every pair of projection operators P and Q, 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 3.2 *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:

$$\|X\|_* = \sup \left\{ \left\langle \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}, \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right\rangle \quad \middle| \quad \left\| \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \right\| = 1 \right\} . \tag{3.2}$$

and similarly

$$\|A\|_* + \|D\|_* = \sup \left\{ \left\langle \begin{bmatrix} Z_{11} & 0 \\ 0 & Z_{22} \end{bmatrix}, \begin{bmatrix} A & B \\ C & D \end{bmatrix} \right\rangle \quad \middle| \quad \left\| \begin{bmatrix} Z_{11} & 0 \\ 0 & Z_{22} \end{bmatrix} \right\| = 1 \right\} . \tag{3.3}$$

Since (3.2) is a supremum over a larger set that (3.3), the claim follows. ■

Theorem 3.1 now trivially follows

Proof [of Theorem 3.1] 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 3.2,

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

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. ■

3.2 Proof of the Weak Bound

Now we can turn to the proof of Theorem 1.2. The key observation in proving this lemma is the following characterization of the null-space of $\mathcal{A}$ provided by Stojnic et al [33]

Lemma 3.3 *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 3.1 and Lemma 3.3, we need to show that for all $v \in \mathbb{R}^M$,

$$\left\| (I - P_{X_0}) \left(\sum_{i=1}^{M} v_i G_i \right) (I - Q_{X_0}) \right\|_* \geq \left\| P_{X_0} \left(\sum_{i=1}^{M} v_i G_i \right) Q_{X_0} \right\|_* . \tag{3.4}$$

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 3.1. 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 3.4 *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*

$$\begin{aligned} \mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1} \left\| (I-P) \left(\sum_{i=1}^{M} v_i G_i \right) (I-Q) \right\|_* - \left\| P \left(\sum_{i=1}^{M} v_i G_i \right) Q \right\|_* \right] \\ \geq \left(\left(\varphi\left(\frac{\gamma - \beta\gamma}{1-\beta\gamma} \right) + o(1) \right) (1-\beta)^{3/2} - (\varphi(1) + o(1)) \gamma^{3/2} \beta^{3/2} \right) n_2^{3/2} - \sqrt{M n_1} \end{aligned} \tag{3.5}$$

where φ is defined as in (1.5).

We will prove this Lemma and a similar inequality required for the proof of the Strong Bound in Section 3.4 below. But we now show how using this Lemma and a concentration of measure argument, we can prove Theorem 1.2.

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 (3.5) is non-negative if (1.6) holds. To bound the probability that (3.4) 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 *Lipshitz* 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 *Lipshitz constant* of the map f. If f is Lipshitz, it cannot vary too rapidly. In particular, note that if f is differentiable and Lipshitz, then L is a bound on the norm of the gradient of f. The following theorem states that the deviations of a Lipshitz function applied to a Gaussian random variable have Gaussian tails.

Theorem 3.5 *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 Lipshitz 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 Lipshitz constant of interest

Lemma 3.6 *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 Lipshitz 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 3.5 and Lemmas 3.4 and 3.6, we can now bound

$$\begin{aligned} \mathbb{P}\left[\inf_{\|v\|_{\ell_2}=1} \left\| (I - P_{X_0}) \left(\sum_{i=1}^M v_i G_i \right) (I - Q_{X_0}) \right\|_* - \left\| P_{X_0} \left(\sum_{i=1}^M v_i G_i \right) Q_{X_0} \right\|_* \leq t n_2^{3/2} \right] \\ \leq \exp\left(-\tfrac{1}{2} \left\{ \varphi\left(\frac{\gamma - \beta\gamma}{1 - \beta\gamma} \right) \frac{(1-\beta)^{3/2}}{\gamma} - \frac{8}{3\pi} \gamma^{1/2} \beta^{3/2} - \sqrt{1-\mu} - \frac{t}{\gamma} \right\} n_2^2 + o(n_2^2) \right). \end{aligned} \tag{3.6}$$

Setting $t = 0$ completes the proof of Theorem 1.2. We will use this concentration inequality with a non-zero t to prove the Strong Bound.

3.3 Proof of the Strong Bound

The proof of the Strong Bound is similar to that of the Weak Bound except we prove that (3.4) 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 $\|P - P'\| + \|Q - Q'\| \leq \epsilon$. We will show that if a slightly stronger bound than (3.4) holds on the ϵ-net, then (3.4) holds for all choices of row and column spaces.

Let us first examine how (3.4) 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

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

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

$$\begin{aligned}
\|(I-P)W(I-Q)\|_* - \|PWQ\|_* \geq \|(I-P')W(I-Q')\|_* - \|P'WQ'\|_* \\
- 2(\|P-P'\| + \|Q-Q'\|)\|W\|_* .
\end{aligned} \tag{3.7}$$

Let us now suppose that with overwhelming probability

$$\|(I-P')W(I-Q')\|_* - \|P'WQ'\|_* - 4\epsilon\|W\|_* \geq 0 \tag{3.8}$$

for all (P', Q') in our ϵ-net Ω. Then by (3.7), this means that $\|(I-P)W(I-Q)\|_* - \|PWQ\|_* \geq 0$ for any arbitrary pair of projection operators onto r-dimensional subspaces. Thus, if we can show that (3.8) holds on an ϵ-net, we will have proved 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 3.7 (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 space with 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,

$$\inf_{\|v\|_{\ell_2}=1} \left\| (I-P)\left(\sum_{i=1}^M v_i G_i\right)(I-Q) \right\|_* - \left\| P\left(\sum_{i=1}^M v_i G_i\right) Q \right\|_* \geq 4\epsilon \sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^M v_i G_i \right\|_* . \tag{3.9}$$

As we will show in Section 3.4, we can upper bound the right hand side of this inequality using a similar bound as in Lemma 3.4.

Lemma 3.8 *For $i = 1, \ldots, M$ let G_i be sampled from $\mathfrak{G}(\gamma n, n)$ with $\gamma \in (0, 1]$. Then*

$$\mathbb{E}\left[\sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^M v_i G_i \right\|_* \right] \leq (\varphi(\gamma) + o(1))\, n^{3/2} + \sqrt{\gamma M n}\,. \tag{3.10}$$

Moreover, we prove the following in the appendix.

Lemma 3.9 *For $i = 1, \dots, M$, let $X_i \in \mathbb{R}^{D_1 \times D_2}$ with $D_1 \leq D_2$ and define the function*

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

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

Using Lemmas 3.8 and 3.9 combined with Theorem 3.5, we have that

$$\mathbb{P}\left[4\epsilon \sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^M v_i G_i \right\|_* \geq t n_2^{3/2} \right] \leq \exp\left(-\tfrac{1}{2} \left(\frac{\varphi(\gamma)}{\gamma} - \sqrt{1-\mu} - \frac{t}{4\epsilon\gamma} \right)^2 n_2^2 + o(n_2^2) \right) . \quad (3.11)$$

Let t_0 be such that the exponents of (3.6) and (3.11) equal to each other. Then we find after some algebra and the union bound

$$\begin{aligned}
&\mathbb{P}\left[\inf_{\|v\|_{\ell_2}=1} \left\| (I-P)\left(\sum_{i=1}^M v_i G_i \right)(I-Q) \right\|_* - \left\| P\left(\sum_{i=1}^M v_i G_i \right) Q \right\|_* \geq 4\epsilon \sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^M v_i G_i \right\|_* \right] \\
\geq &\mathbb{P}\left[\inf_{\|v\|_{\ell_2}=1} \left\| (I-P)\left(\sum_{i=1}^M v_i G_i \right)(I-Q) \right\|_* - \left\| P\left(\sum_{i=1}^M v_i G_i \right) Q \right\|_* > t_0 n_2^{3/2} > 4\epsilon \sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^M v_i G_i \right\|_* \right] \\
\geq &1 - \mathbb{P}\left[\inf_{\|v\|_{\ell_2}=1} \left\| (I-P)\left(\sum_{i=1}^M v_i G_i \right)(I-Q) \right\|_* - \left\| P\left(\sum_{i=1}^M v_i G_i \right) Q \right\|_* < t_0 n_2^{3/2} \right] \\
&\qquad - \mathbb{P}\left[4\epsilon \sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^M v_i G_i \right\|_* > t_0 n_2^{3/2} \right] \\
\geq &1 - 2\exp\left(-\tfrac{1}{2} \left(\frac{\varphi\left(\frac{\gamma-\beta\gamma}{1-\beta\gamma}\right)\gamma^{-1}(1-\beta)^{3/2} - \frac{8}{3\pi}\gamma^{1/2}\beta^{3/2} - 4\epsilon\varphi(\gamma)}{1+4\epsilon} - \sqrt{1-\mu} \right)^2 n_2^2 + o(n_2^2) \right) .
\end{aligned}$$

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. Again by the union bound, we have that

$$\begin{aligned}
&\mathbb{P}\left[\forall P, Q \inf_{\|v\|_{\ell_2}=1} \left\| (I-P)\left(\sum_{i=1}^M v_i G_i \right)(I-Q) \right\|_* - \left\| P\left(\sum_{i=1}^M v_i G_i \right) Q \right\|_* \geq 4\epsilon \sup_{\|v\|_{\ell_2}=1} \left\| \sum_{i=1}^M v_i G_i Q \right\|_* \right] \\
&\leq 1 - 2\exp\left(-\left\{ \tfrac{1}{2}\left(f(\beta,\gamma,\epsilon) - \sqrt{1-\mu} \right)^2 - \tfrac{1}{2} g(\beta,\gamma,\epsilon)^2 \right\} n_2^2 + o(n_2^2) \right)
\end{aligned} \quad (3.12)$$

where

$$f(\gamma,\beta,\epsilon) = \frac{\varphi\left(\frac{\gamma-\beta\gamma}{1-\beta\gamma}\right)\gamma^{-1}(1-\beta)^{3/2} - \frac{8}{3\pi}\gamma^{1/2}\beta^{3/2} - 4\epsilon\varphi(\gamma)}{1+4\epsilon} \quad (3.13)$$

$$g(\gamma,\beta,\epsilon) = \sqrt{2\beta\gamma(1+\gamma-\beta\gamma)\log\left(\frac{3\pi}{2\epsilon}\right)} . \quad (3.14)$$

Finding the parameters μ, β, γ, and ϵ that make the terms multiplying n_2^2 negative completes the proof of the Strong Bound.

3.4 Comparison Theorems for Gaussian Processes and the Proofs of Lemmas 3.4 and 3.8

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

$$\begin{cases} \mathbb{E}[X_i X_j] \leq \mathbb{E}[Y_i Y_j] & \text{for all } i \neq j \\ \mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2] & \text{for all } i \end{cases}$$

Then

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

Theorem 3.11 (Gordan [17, 18]) *Let $X = (X_{ij})$ and $Y = (Y_{ij})$ be Gaussian random matrices in $\mathbb{R}^{N_1 \times N_2}$ such that*

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

Then

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

The following 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 $D_1 \leq D_2$. Let us define the quantity $\sigma(\|\cdot\|_p)$ to be the maximum attainable Frobenius norm of an element in the unit ball of the dual norm. That is

$$\sigma(\|\cdot\|_p) = \sup_{\|Z\|_d = 1} \|Z\|_F \,, \tag{3.15}$$

and note that by this definition, we have for $G \in \mathfrak{G}(D_1, D_2)$

$$\sigma(\|\cdot\|_p) = \sup_{\|Z\|_d = 1} \mathbb{E}_G \left[\langle G, Z \rangle^2\right]^{1/2}$$

motivating the notation.

This first Lemma is now a straightforward consequence of Slepian's Lemma

Lemma 3.12 *Let $\Delta > 0$ 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*

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

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

$$Q_L(v, Y) = \Delta \langle g, v \rangle + \left\langle \sum_{i=1}^{M} v_i G_i, Y \right\rangle + \sigma(\|\cdot\|_p)\gamma$$
$$Q_R(v, Y) = \langle G, Y \rangle + \sqrt{\Delta^2 + \sigma(\|\cdot\|_p)^2}\langle g, v \rangle .$$

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

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

The first quantity is always non-negative because $\langle Y, \hat{Y} \rangle \leq \max(\|Y\|_F^2, \|\hat{Y}\|_F^2) \leq \sigma(\|\cdot\|_p)^2$ 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 A.1 in the Appendix),

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

which proves the Lemma. ■

The following lemma can be proved in a similar fashion

Lemma 3.13 *Let $\|\cdot\|_p$ be a norm on $\mathbb{R}^{D_1 \times D_1}$ with dual norm $\|\cdot\|_d$ and let $\|\cdot\|_b$ be a norm on $\mathbb{R}^{D_2 \times D_2}$. 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 $\mathfrak{G}(D_1)$ and $G'_1, \ldots, G'_M$ be sampled i.i.d. from $\mathfrak{G}(D_2)$. Then*

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

Proof Define the functionals

$$P_L(v, Y, Z) = \left\langle \sum_{i=1}^{M} v_i G_i, Z \right\rangle + \left\langle \sum_{i=1}^{M} v_i G'_i, Y \right\rangle + \gamma\sigma(\|\cdot\|_p)$$
$$P_R(v, Y, Z) = \langle G_0, Z \rangle + \sigma(\|\cdot\|_p)\langle g, v \rangle + \left\langle \sum_{i=1}^{M} v_i G'_i, Y \right\rangle .$$

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

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

Just as was the case in the proof of Lemma 3.12, the first quantity is always non-negative. Hence, the difference in expectations is greater than or equal to zero and equal to zero when $v = \hat{v}$ and $Y = \hat{Y}$. 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 3.12 and 3.13, we can prove the Lemma 3.4.

Proof [of Lemma 3.4] For $i = 1,\ldots,M$, let $G_i \in \mathfrak{G}((1-\beta)\gamma n_2, (1-\beta\gamma)n_2)$ and $G_i' \in \mathfrak{G}(\gamma\beta n_2, \gamma\beta n_2)$. Then

$$\begin{aligned}
&\mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1}\left\|\sum_{i=1}^M v_i G_i\right\|_* - \left\|\sum_{i=1}^M v_i G_i'\right\|_*\right] \\
&= \mathbb{E}\left[\inf_{\|v\|_{\ell_2}=1}\inf_{\|Y\|=1}\sup_{\|Z\|=1}\left\langle\sum_{i=1}^M v_i G_i, Z\right\rangle + \left\langle\sum_{i=1}^M v_i G_i', Y\right\rangle\right] \\
&\geq \mathbb{E}\left[\|G_0\|_*\right] - \mathbb{E}\left[\sup_{\|v\|_{\ell_2}=1}\sup_{\|Y\|=1}\sigma(\|\cdot\|_*)\langle g, v\rangle + \left\langle\sum_{i=1}^M v_i G_i', Y\right\rangle\right] \\
&\geq \mathbb{E}\left[\|G_0\|_*\right] - \mathbb{E}\left[\|G_0'\|_*\right] - \sqrt{M}\sqrt{\sigma(\|\cdot\|_*)^2 + \sigma(\|\cdot\|_*)^2}
\end{aligned}$$

where the first inequality follows from Lemma 3.13, and the second inequality follows from Lemma 3.12.

Now we only need to plug in the asymptotic expected value of the nuclear norm and the quantity $\sigma(\|\cdot\|_*)$. Let G be sampled from $\mathfrak{G}(D_1, D_2)$. Then

$$\mathbb{E}\|G\|_* = D_1\mathbb{E}\sigma_i = \varphi\left(\frac{D_1}{D_2}\right)D_2^{3/2} + q(D_2) \tag{3.16}$$

where $\varphi(\cdot)$ is found by integrating the Marčenko-Pastur distribution (see, e.g., [25, 4]):

$$\begin{aligned}
\varphi(\gamma) &= \frac{1}{2\pi}\int_{s_1}^{s_2}\sqrt{\frac{(z-s_1)(s_2-z)}{z}}dz \\
s_1 &= (1-\sqrt{\gamma})^2 \\
s_2 &= (1+\sqrt{\gamma})^2\,.
\end{aligned}$$

and $q(D_2)/D_2^{3/2} = o(1)$. Note that $\varphi(1)$ can be computed in closed form:

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

For $\sigma(\|\cdot\|_*)$, a straightforward calculation reveals

$$\sigma(\|\cdot\|_*) = \sup_{\|H\|\leq 1} \|G\|_F = \sqrt{D_1}\,.$$

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

Proof [of Lemma 3.8] This lemma immediately follows from applying Lemma 3.12 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]. ■

4 Numerical Experiments

We now show that these asymptotic estimates hold even for moderately sized matrices. For simplicity of presentation, we restrict our attention in this section to square matrices with $n = n_1 = n_2$ (i.e., $\gamma = 1$). We conducted a series of experiments for a variety of the matrix sizes n, ranks r, 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 $n \times n$ matrices. For each n, we varied μ between 0 and 1 where the matrix is completely determined. For a fixed n and μ, we generated all possible ranks such that $\beta(2-\beta) \leq \mu$. This cutoff was chosen because beyond that point there would be an infinite set of matrices of rank r satisfying the m equations.

For each (n, μ, β) triple, we repeated the following procedure 10 times. A matrix of rank r was generated by choosing two random $n \times r$ factors Y_L and Y_R 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 n^2 columns. Then the nuclear norm minimization

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

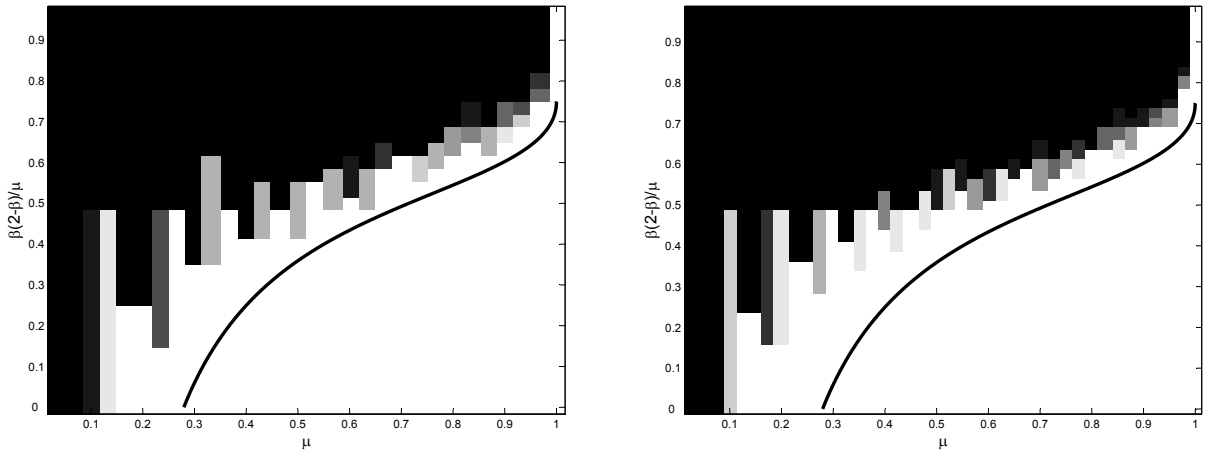
was solved using the freely available software SeDuMi [34] using the semidefinite programming formulation described in [29]. On a 2.0 GHz Laptop, each semidefinite program could be solved in less than two minutes for 40×40 dimensional X. We declared Y_0 to be recovered if

$$\|X - Y_0\|_F / \|Y_0\|_F < 10^{-3}\,.$$

Figure 3 displays the results of these experiments for $n = 30$ and 40. 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. It is remarkable to note that not only are the plots very similar for $n = 30$ and $n = 40$, but that the Weak Bound falls completely within the white region and is an excellent approximation of the boundary between success and failure for large β.

5 Discussion and Future Work

Future work should investigate if the probabilistic analysis that provides the bounds in Theorems 1.2 and 1.3 can be further tightened at all. There are two particular regions where the bounds can be



(a) (b)

Figure 3: Random rank recovery experiments for (a) $n = 30$ and (b) $n = 40$. The color of each cell reflects the empirical recovery rate. White denotes perfect recovery in all experiments, and black denotes failure for all experiments. In both frames, we plot the Weak Bound (1.6), 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 β.

improved. First, when $\beta = 0$, μ should also equal zero. However, in our Weak Bound, $\beta = 0$ 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 (3.5). When μ is small, M in (3.5) 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. For large values of β, the bound is a rather good approximation of empirical results, and it might not be possible to further tighten this bound. However, it is still worth looking to see if some of the techniques in [12, 13] 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 3.1, 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 (1.1) 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. For example, 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 [9].

References

[1] B. P. W. Ames and S. A. Vavasis, "Nuclear norm minimization for the planted clique and biclique problems," 2009.

[2] F. M. S. N. . U. S. Amit, Y., "Uncovering shared structures in multiclass classification," in *Proceedings of the International Conference of Machine Learning*, 2007.

[3] A. Argyriou, C. A. Micchelli, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, 2008, published online first at http://www.springerlink.com/.

[4] Z. D. Bai, "Methodologies in spectral analysis of large dimensional random matrices," *Statistica Sinica*, vol. 9, no. 3, pp. 611–661, 1999.

[5] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "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] C. Beck and R. D'Andrea, "Computational study and comparisons of LFT reducibility methods," in *Proceedings of the American Control Conference*, 1998.

[7] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," Tech. Rep., 2008, preprint available at http://arxiv.org/abs/0810.3286.

[8] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[9] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, 2009, to Appear.

[10] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[11] D. Donoho, "High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension," *Discrete and Computational Geometry*, vol. 35, no. 4, pp. 617–652, 2006.

[12] D. L. Donoho and J. Tanner, "Neighborliness of randomly projected simplices in high dimensions," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 27, pp. 9452–9457, 2005.

[13] ——, "Sparse nonnegative solution of underdetermined linear equations by linear programming," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 27, pp. 9446–9451, 2005.

[14] L. El Ghaoui and P. Gahinet, "Rank minimization under LMI constraints: A framework for output feedback problems," in *Proceedings of the European Control Conference*, 1993.

[15] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford University, 2002.

[16] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American Control Conference*, 2001.

[17] Y. Gordan, "Some inequalities for gaussian processes and applications," *Israel Journal of Math*, vol. 50, pp. 265–289, 1985.

[18] ——, "Gaussian processes and almost spherical sections of convex bodies," *Annalys of Probability*, vol. 16, pp. 180–188, 1988.

[19] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. New York: Cambridge University Press, 1991.

[20] M. Ledoux and M. Talagrand, *Probability in Banach Spaces*. Berlin: Springer-Verlag, 1991.

[21] K. Lee and Y. Bresler, "Efficient and guaranteed rank minimization by atomic decomposition," 2009, submitted to ISIT2009. Preprint available at http://arxiv.org/abs/0901.1898v1.

[22] N. Linial, E. London, and Y. Rabinovich, "The geometry of graphs and some of its algorithmic applications," *Combinatorica*, vol. 15, pp. 215–245, 1995.

[23] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," 2008, submitted.

[24] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," Tech. Rep., 2008.

[25] V. A. Marčenko and L. A. Pastur, "Distributions of eigenvalues for some sets of random matrices," *Math. USSR-Sbornik*, vol. 1, pp. 457–483, 1967.

[26] R. Meka, P. Jain, C. Caramanis, and I. S. Dhillon, "Rank minimization via online learning," in *Proceedings of the International Conference on Machine Learning*, 2008.

[27] M. Mesbahi and G. P. Papavassilopoulos, "On the rank minimization problem over a positive semidefinite linear matrix inequality," *IEEE Transactions on Automatic Control*, vol. 42, no. 2, pp. 239–243, 1997.

[28] P. A. Parrilo and S. Khatri, "On cone-invariant linear matrix inequalities," *IEEE Trans. Automat. Control*, vol. 45, no. 8, pp. 1558–1563, 2000.

[29] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization," Submitted. Preprint Available at http://www.ist.caltech.edu/~brecht/publications.html.

[30] B. Recht, W. Xu, and B. Hassibi, "Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization," in *Prooceedings of the 47th IEEE Conference on Decision and Control*, 2008.

[31] J. D. M. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proceedings of the International Conference of Machine Learning*, 2005.

[32] D. Slepian, "The one-sided barrier problem for gaussian noise," *Bell System Technical Journal*, vol. 41, pp. 463–501, 1962.

[33] M. Stojnic, W. Xu, and B. Hassibi, "Compressed sensing - probabilistic analysis of a null-space characterization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[34] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999.

[35] S. J. Szarek, "Metric entropy of homogeneous spaces," in *Quantum probability (Gdańsk, 1997)*, ser. Banach Center Publ. Warsaw: Polish Acad. Sci., 1998, vol. 43, pp. 395–410, preprint available at arXiv:math/9701213v1.

[36] A. M. Vershik and P. V. Sporyshev, "Asymptotic behavior of the number of faces of random polyhedra and the neighborliness problem," *Selecta Mathematica Sovietica*, vol. 11, no. 2, pp. 181–201, 1992.

[37] K. Q. Weinberger and L. K. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 77–90, 2006.

[38] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, "Dimension reduction and coecient estimation in multivariate linear regression," *Journal of the Royal Statistical Society: Series B*, vol. 69, pp. 329–346, 2007.

A Appendix

A.1 Rank-deficient case of Theorem 1.1

As promised above, here is the completion of the proof of Theorem 1.1

Proof In an appropriate basis, we may write

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

If Y_{11} and $Y_{22} - Y_{21}Y_{11}^{-1}Y_{12}$ have full rank, then all our previous arguments apply. Thus, assume that at least one of them is not full rank. Nonetheless, it is always possible to find an *arbitrarily small* $\epsilon > 0$ such that

$$Y_{11} + \epsilon I \quad \text{and} \quad \begin{bmatrix} Y_{11} + \epsilon I & Y_{12} \\ Y_{21} & Y_{22} + \epsilon I \end{bmatrix}$$

are full rank. This, of course, is equivalent to having $Y_{22} + \epsilon I - Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12}$ full rank. We can write

$$\begin{aligned}
\|X_*\|_* &= \|X_0 + X_* - X_0\|_* \\
&= \left\| \begin{bmatrix} X_{11} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \right\|_* \\
&\geq \left\| \begin{bmatrix} X_{11} - \epsilon I & 0 \\ 0 & Y_{22} - Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} \right\|_* - \left\| \begin{bmatrix} Y_{11} + \epsilon I & Y_{12} \\ Y_{21} & Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} \right\|_* \\
&= \|X_{11} - \epsilon I\|_* + \left\| \begin{bmatrix} 0 & 0 \\ 0 & Y_{22} - Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} \right\|_* - \left\| \begin{bmatrix} Y_{11} + \epsilon I & Y_{12} \\ Y_{21} & Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} \right\|_* \\
&\geq \|X_0\|_* - r\epsilon + \left\| \begin{bmatrix} \epsilon I - \epsilon I & 0 \\ 0 & Y_{22} - Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} \right\|_* - \left\| \begin{bmatrix} Y_{11} + \epsilon I & Y_{12} \\ Y_{21} & Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} \right\|_* \\
&\geq \|X_0\|_* - 2r\epsilon \quad + \left\| \begin{bmatrix} -\epsilon I & 0 \\ 0 & Y_{22} - Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} \right\|_* - \left\| \begin{bmatrix} Y_{11} + \epsilon I & Y_{12} \\ Y_{21} & Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} \right\|_* \\
&\geq \|X_0\|_* - 2r\epsilon,
\end{aligned}$$

where the last inequality follows from the condition of part 1 and noting that

$$X_0 - X_* = \begin{bmatrix} -\epsilon I & 0 \\ 0 & Y_{22} - Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix} + \begin{bmatrix} Y_{11} + \epsilon I & Y_{12} \\ Y_{21} & Y_{21}(Y_{11} + \epsilon I)^{-1}Y_{12} \end{bmatrix},$$

lies in the null space of $\mathcal{A}(\cdot)$ and the first matrix above has rank more than r. But, since ϵ can be arbitrarily small, this implies that $X_0 = X_*$. ■

A.2 Lipshitz Constants of F_I and F_S

We begin with the proof of Lemma 3.9 and then use this to estimate the Lipshitz constant in Lemma 3.6.

Proof [of Lemma 3.9] Note that the function F_S is convex as we can write as a supremum of a collection of convex functions

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

The Lipshitz constant L is bounded above by the maximal norm of a subgradient of this convex function. That is, if we denote $\bar{X} := (X_1, \ldots, X_M)$, then we have

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

Now, by (A.1), a subgradient of F_S at $\bar{X}$ is given of the form $(v_1 Z, v_2 Z, \ldots, v_M Z)$ 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 Lipshitz constant as desired. ■

Proof [of Lemma 3.6] For $i = 1, \ldots, M$, 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

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

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 3.9, we have that a subgradient of this expression is of the form

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

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 A.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. ■