Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization

Benjamin Recht [*] Maryam Fazel [†] Pablo A. Parrilo [‡]

August 6, 2008

Abstract

The affine rank minimization problem consists of finding a matrix of minimum rank that satisfies a given system of linear equality constraints. Such problems have appeared in the literature of a diverse set of fields including system identification and control, Euclidean embedding, and collaborative filtering. Although specific instances can often be solved with specialized algorithms, the general affine rank minimization problem is NP-hard, because it contains vector cardinality minimization as a special case.

In this paper, we show that if a certain restricted isometry property holds for the linear transformation defining the constraints, the minimum rank solution can be recovered by solving a convex optimization problem, namely the minimization of the nuclear norm over the given affine space. We present several random ensembles of equations where the restricted isometry property holds with overwhelming probability, provided the codimension of the subspace is sufficiently large.

The techniques used in our analysis have strong parallels in the compressed sensing framework. We discuss how affine rank minimization generalizes this pre-existing concept and outline a dictionary relating concepts from cardinality minimization to those of rank minimization. We also discuss several algorithmic approaches to solving the nuclear norm minimization relaxation, and illustrate our results with numerical examples.

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

Keywords. rank, convex optimization, matrix norms, random matrices, compressed sensing, semidefinite programming.

1 Introduction

Notions such as order, complexity, or dimensionality can often be expressed by means of the rank of an appropriate matrix. For example, a low-rank matrix could correspond to a low-degree statistical model for a random process (e.g., factor analysis), a low-order realization of a linear system [38], a low-order controller for a plant [30], or a low-dimensional embedding of data in Euclidean space [47].

[*] Center for the Mathematics of Information, California Institute of Technology

[†] Electrical Engineering Department, University of Washington

[‡] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology

If the set of models that satisfy the desired constraints is convex, then choosing the simplest model can be cast as a *rank minimization problem*,

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & X \in \mathcal{C} \end{array} \tag{1.1}$$

where $X \in \mathbb{R}^{m \times n}$ is the decision variable, and $\mathcal{C}$ is some given convex constraint set. This problem arises in various application areas; see, for example, [37, 52]. In certain instances with very special structure, the rank minimization problem can be solved by using the singular value decomposition, or can be exactly reduced to the solution of linear systems [54, 60]. In general, however, the problem (1.1) is a challenging non-convex optimization problem, and even when $\mathcal{C}$ is an affine subspace seems to require at least exponential running time in both theory and practice. For the general case, a variety of heuristic algorithms based on local optimization, including alternating projections and its variations [42, 58], alternating matrix inequalities [68], linearization [31], and augmented Lagrangian methods [34] have been proposed.

A recent heuristic introduced by Fazel *et al.* in [37, 38] minimizes the *nuclear norm*, or the sum of the singular values of the matrix, over the constraint set. The nuclear norm is a convex function, can be optimized efficiently, and is the best convex approximation of the rank function over the set of matrices with spectral norm less than or equal to one. When the matrix variable is symmetric and positive semidefinite, this heuristic is equivalent to the trace heuristic sometimes used by the systems and control community (e.g., [6, 54]). The nuclear norm heuristic has been observed to produce very low-rank solutions in practice, but a theoretical characterization of when it produces the minimum rank solution has not been previously available. This paper provides the first such mathematical characterization.

In this paper, we focus on the scenario where the set of feasible models or designs is affine in the matrix variable, and consider the *affine rank minimization problem*,

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & \mathcal{A}(X) = b, \end{array} \tag{1.2}$$

where $X \in \mathbb{R}^{m \times n}$ is the decision variable, and the linear map $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^p$ and vector $b \in \mathbb{R}^p$ are given. Our work is built upon a large body of literature on a related optimization problem. When the matrix variable is constrained to be diagonal, the affine rank minimization problem reduces to finding the *sparsest vector* in an affine subspace. This problem is commonly referred to as *cardinality minimization*, since one seeks the vector whose support has the smallest cardinality, and is known to be NP-hard [56]. Just as in the case of rank minimization, a variety of heuristic algorithms have been proposed to solve cardinality minimization problems including Projection Pursuit [40, 51, 62] and Orthogonal Matching Pursuit [19, 24, 61].

For diagonal matrices, the sum of the singular values is equal to the sum of the absolute values (i.e., the ℓ_1 norm) of the diagonal elements. Minimization of the ℓ_1 norm is a well-known heuristic for the cardinality minimization problem, employed as early as the 1970s by geophysicists attempting to deconvolve seismic activity [21, 74]. Since then, ℓ_1 minimization has been applied to a variety of cardinality minimization problems including image denoising [65], model selection in statistics [76], sparse approximation [20], portfolio optimization with fixed transaction costs [49], design of sparse interconnect wiring in circuits [80], and design of sparse feedback gains in control systems [43].

Recently, stunning results pioneered by Candès and Tao [12] and Donoho [25] have characterized a vast set of instances for which the ℓ_1 heuristic can be *a priori* guaranteed to yield the optimal solution. These techniques provide the foundations of the recently developed *compressed sensing* or *compressive sampling* frameworks for measurement, coding, and signal estimation. As has been shown by a number of research groups (e.g., [4, 14, 15, 16]), the ℓ_1 heuristic for cardinality minimization provably recovers the sparsest solution whenever the sensing matrix has certain "basis incoherence" properties, and in particular, when it is randomly chosen according to certain specific ensembles.

The fact that the ℓ_1 heuristic is a special case of the nuclear norm heuristic suggests that these results from the compressed sensing literature might be extended to provide guarantees about the nuclear norm heuristic for the more general rank minimization problem. In this paper, we show that this is indeed the case, and the parallels are surprisingly strong. Following the program laid out in the work of Candès and Tao, our main contribution is the development of a restricted isometry property (RIP), under which the nuclear norm heuristic can be *guaranteed* to produce the minimum-rank solution. Furthermore, as in the case for the ℓ_1 heuristic, we provide several specific examples of matrix ensembles for which RIP holds with overwhelming probability. Our results considerably extend the compressed sensing machinery in a so far undeveloped direction, by allowing a much more general notion of parsimonious models that rely on low-rank assumptions instead of cardinality restrictions.

To make the parallels as clear as possible, we begin by establishing a dictionary between the matrix rank and nuclear norm minimization problems and the vector sparsity and ℓ_1 norm problems in Section 2. In the process of this discussion, we present a review of many useful properties of the matrices and matrix norms necessary for the main results. We then generalize in Section 3 the notion of Restricted Isometry to matrices, and show that when linear mappings are Restricted Isometries, recovering low-rank solutions of underdetermined systems can be achieved by nuclear norm minimization. In Section 4, we present several families of random linear maps that are restricted isometries with overwhelming probability when the dimensions are sufficiently large. In Section 5, we briefly discuss three different algorithms designed for solving the nuclear norm minimization problem and their relative strengths and weaknesses: interior point methods, gradient projection methods, and a low-rank factorization technique. In Section 6, we present several numerical experiments, and demonstrate that in practice nuclear-norm minimization recovers the lowest rank solutions of affine sets with even fewer constraints than those guaranteed by our mathematical analysis. Finally, in Section 7, we list a number of possible directions for future research.

1.1 When are random constraints interesting for rank minimization?

As in the case of compressed sensing, the conditions we derive to guarantee properties about the nuclear norm heuristic are deterministic, but they are at least as difficult to check as solving the rank minimization problem itself. We are only able to guarantee that the nuclear norm heuristic recovers the minimum rank solution of $\mathcal{A}(X) = b$ when $\mathcal{A}$ is sampled from specific ensembles of random maps. The constraints appearing in many of the applications mentioned above, such as low-order control system design, are typically not random at all and have structured demands according to the specifics of the design problem. Furthermore, in many of these applications, the problem is formulated as minimizing rank subject to some more general convex constraints than the linear equalities we are considering. It thus behooves us to present several examples where random affine constraints manifest themselves in practical scenarios for which no practical solution procedure is known.

Minimum order linear system realization Rank minimization forms the basis of many model reduction and low-order system identification problems for linear time-invariant systems. The following example illustrates how random constraints might arise in this context. Consider the problem of finding the minimum order discrete-time LTI system that is consistent with a set of time-domain observations. In particular, suppose our observations are the system output sampled at a fixed time N, after a random Gaussian input signal is applied from $t = 0$ to $t = N$. Suppose we make such measurements for p different input signals, that is, we observe $y_i(N) = \sum_{t=0}^{N} a_i(N-t)h(t)$ for $i = 1, \dots, p$, where a_i, the ith input signal, is a zero-mean Gaussian random variable with the same variance for $t = 0, \dots, N$, and $h(t)$ denotes the impulse response. We can write this compactly as $y = Ah$, where $h = [h(0), \dots, h(N)]'$, and $A_{ij} = a_i(N - j)$.

From linear system theory, the order of the minimal realization for such a system is given by the rank of the following Hankel matrix (see, e.g., [39, 69])

$$\operatorname{hank}(h) := \begin{bmatrix} h(0) & h(1) & \cdots & h(N) \\ h(1) & h(2) & \cdots & h(N+1) \\ \vdots & \vdots & & \vdots \\ h(N) & h(N+1) & \cdots & h(2N) \end{bmatrix}.$$

Therefore the problem can be expressed as

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\operatorname{hank}(h)) \\ \text{subject to} & Ah = y \end{array}$$

where the optimization variables are $h(0), \dots, h(2N)$, and the matrix A consists of i.i.d. zero-mean Gaussian entries.

Low-rank matrix completion In the matrix completion problem, we are given a random subset of entries of a matrix and would like to fill in the missing entries such that the resulting matrix has the lowest possible rank. This problem arises in machine learning scenarios where we are given partially observed examples of a process with a low-rank covariance matrix and would like to estimate the missing data. Such models are ubiquitous in Factor Analysis, Collaborative Filtering, and Latent Semantic Analysis [63, 70]. In many of these settings, some prior probability distribution (such as a Bernoulli model or uniform distribution on subsets) is assumed to generate the set of available entries.

Suppose we are presented with a set of triples $(I(i), J(i), S(i))$ for $i = 1, \dots, p$ and wish to find the matrix with $S(i)$ in the entry corresponding to row $I(i)$ and column $J(i)$ for all i. The matrix completion problem seeks to

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(Y) \\ \text{subject to} & Y_{I(i),J(i)} = S(i), \qquad i = 1, \dots, K \end{array}$$

which is a special case of the affine rank minimization problem.

Low-dimensional Euclidean embedding problems A problem that arises in a variety of fields is the determination of configurations of points in low-dimensional Euclidean spaces subject to some given distance information. In computational chemistry, these problems arise in inferring the three-dimensional structure of a molecule (molecular conformation) from information about interatomic distances [78]. In manifold learning, one may be given high dimensional data with low-dimensional structure that can be recovered by searching for a low-dimensional embedding of the data preserving local distance information [64, 75].

A symmetric matrix $D \in \mathcal{S}^n$ is called a *Euclidean distance matrix* (EDM) if there exist points $x_1, \ldots, x_n$ in $\mathbb{R}^d$ such that $D_{ij} = \|x_i - x_j\|^2$. Let $V := I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ be the projection matrix onto the hyperplane $\{v \in \mathbb{R}^n \ : \ \mathbf{1}^T v = 0\}$. A classical result by Schoenberg states that D is a Euclidean distance matrix of n points in $\mathbb{R}^d$ if and only if $D_{ii} = 0$, the matrix VDV is negative semidefinite, and $\operatorname{rank}(VDV)$ is less than or equal to d [67]. If the matrix D is known exactly, the corresponding configuration of points (up to a unitary transform) is obtained by simply taking a matrix square root of $-\frac{1}{2}VDV$. However, in many cases, only a random sampling collection of the distances are available. The problem of finding a valid EDM consistent with the known inter-point distances and with the smallest embedding dimension can be expressed as the rank optimization problem

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(VDV) \\ \text{subject to} & VDV \preceq 0 \\ & \mathcal{A}(D) = b, \end{array}$$

where $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^p$ is a random sampling operator as discussed in the matrix completion problem.

This problem involves a Linear Matrix Inequality (LMI) and appears to be more general than the equality constrained rank minimization problem. However, general LMIs can equivalently be expressed as rank constraints on an appropriately defined block matrix. The rank of a block symmetric matrix is equal to the rank of a diagonal block plus the rank of its Schur complement (see, e.g., [45, §2.2]). Given a function f that maps matrices into $q \times q$ symmetric matrices, the condition that $f(X)$ is positive semidefinite can be equivalently expressed through a rank constraint as

$$f(X) \succeq 0 \qquad \Leftrightarrow \qquad \operatorname{rank}\left(\begin{bmatrix} I_q & B \\ B' & f(X) \end{bmatrix} \right) \leq q, \qquad \text{for some } B \in \mathbb{R}^{q \times q}.$$

That is, if there exists a matrix B satisfying the inequality above, then $f(X) = B'B \succeq 0$. Using this equivalent representation allows us to rewrite problem (1.1) with general LMI constraints as an affine rank minimization problem.

Image compression A simple and well-known method to compress two-dimensional images can be obtained by using the singular value decomposition (e.g., [3]). The basic idea is to associate to the given grayscale image a rectangular matrix M, with the entries M_{ij} corresponding to the gray level of the (i, j) pixel. The best rank-k approximation of M is given by

$$X^* := \arg \min_{\operatorname{rank}(X) \leq k} ||M - X||,$$

where $||\cdot||$ is any unitarily invariant norm. By the classical Eckart-Young-Mirsky theorem ([28, 55]), the optimal approximant is given by a truncated singular value decomposition of M, i.e., if $M = U\Sigma V^T$, then $X^* = U\Sigma_k V^T$, where the first k diagonal entries of Σ_k are the largest k singular values, and the rest of the entries are zero. If for a given rank k, the approximation error $||M - X^*||$ is small enough, then the amount of data needed to encode the information about the image is $k(m+n-k)$ real numbers, which can be much smaller than the mn required to transmit the values of all the entries.

parsimony concept	cardinality	rank
Hilbert Space norm	Euclidean	Frobenius
sparsity inducing norm	ℓ_1	nuclear
dual norm	ℓ_∞	operator
norm additivity	disjoint support	orthogonal row and column spaces
convex optimization	linear programming	semidefinite programming

Table 1: A dictionary relating the concepts of cardinality and rank minimization.

Consider a given image, whose associated matrix M has low-rank, or can be well-approximated by a low-rank matrix. As proposed by Wakin *et al.* [81], a single-pixel camera would ideally produce measurements that are random linear combinations of all the pixels of the given image. Under this situation, the image reconstruction problem boils down exactly to affine rank minimization, where the constraints are given by the random linear functionals.

It should be remarked that the simple SVD image compression scheme described has certain deficiencies that more sophisticated techniques do not share (in particular, the lack of invariance of the description length under rotations). Nevertheless, due to its simplicity and relatively good practical performance, this method is particularly popular in introductory treatments and numerical linear algebra textbooks.

2 From compressed sensing to rank minimization

As discussed above, when the matrix variable is constrained to be diagonal, the affine rank minimization problem (1.2) reduces to the cardinality minimization problem of finding the element in the affine space with the fewest number of nonzero components. In this section we will establish a dictionary between the concepts of rank and cardinality minimization. The main elements of this correspondence are outlined in Table 2. With these elements in place, the existing proofs of sparsity recovery provide a template for the more general case of low-rank recovery.

In establishing our dictionary, we will provide a review of useful facts regarding matrix norms and their characterization as convex optimization problems. We will show how computing both the operator norm and the nuclear norm of a matrix can be cast as semidefinite programming (SDP) problems. We also establish suitable optimality conditions for the minimization of the nuclear norm subject to affine equality constraints, the main convex optimization problem studied in this article. Our discussion of matrix norms and their connections to semidefinite programming and convex optimization will mostly follow the discussion in [9, 37, 79] where extensive lists of references are provided.

Matrix vs. vector norms The three vector norms that play significant roles in the compressed sensing framework are the ℓ_1, ℓ_2, and ℓ_∞ norms, denoted by $\|x\|_1$, $\|x\|$ and $\|x\|_\infty$ respectively. These norms have natural generalizations to matrices, inheriting many appealing properties from the vector case. In particular, there is a parallel duality structure.

For a rectangular matrix $X \in \mathbb{R}^{m\times n}$, $\sigma_i(X)$ denotes the i-th largest singular value of X and is equal to the square-root of the i-th largest eigenvalue of XX'. The rank of X will usually be denoted by r, and is equal to the number of nonzero singular values. For matrices X and Y of the same dimensions, we define the inner product in $\mathbb{R}^{m\times n}$ as $\langle X, Y\rangle := \mathrm{Tr}(X'Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij}$. 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.,

$$\|X\|_F := \sqrt{\langle X, X\rangle} = \sqrt{\mathrm{Tr}(X'X)} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{r}\sigma_i^2\right)^{\frac{1}{2}}.$$

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.,

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

and is alternatively known by several other names including the Schatten 1-norm, the Ky Fan r-norm, and the trace class norm. Since the singular values are all positive, the nuclear norm is also equal to the ℓ_1 norm of the vector of singular values. These three norms are related by the following inequalities which hold for any matrix X of rank at most r:

$$||X|| \le ||X||_F \le ||X||_* \le \sqrt{r}||X||_F \le r||X||. \tag{2.1}$$

Dual norms For any given norm $\|\cdot\|$ in an inner product space, there exists a dual norm $\|\cdot\|_d$ defined as

$$\|X\|_d := \sup_Y\{\langle X, Y\rangle \,:\, \|Y\| \le 1\}. \tag{2.2}$$

Furthermore, the norm dual to the norm $||\cdot||_d$ is again the original norm $||\cdot||$.

In the case of vector norms in $\mathbb{R}^n$, it is well-known that the dual norm of the ℓ_p norm (with $1 < p < \infty$) is the ℓ_q norm, where $\frac{1}{p} + \frac{1}{q} = 1$. This fact is essentially equivalent to Hölder's inequality. Similarly, the dual norm of the ℓ_∞ norm of a vector is the ℓ_1 norm. These facts also extend to the matrix norms we have defined. For instance, the dual norm of the Frobenius norm is the Frobenius norm. This can be verified by simple calculus (or Cauchy-Schwarz), since

$$\sup_Y\{\mathrm{Tr}(X'Y) \,:\, \mathrm{Tr}(Y'Y) \le 1\}$$

is equal to $\|X\|_F$, with the maximizing Y being equal to $X/\|X\|_F$. Similarly, as shown below, the dual norm of the operator norm is the nuclear norm. The proof of this fact will also allow us to present variational characterizations of each of these norms as semidefinite programs.

Proposition 2.1 *The dual norm of the operator norm $||\cdot||$ in $\mathbb{R}^{m\times n}$ is the nuclear norm $||\cdot||_*$.*

Proof First consider an $m \times n$ matrix Z. The fact that Z has operator norm less than t can be expressed as a linear matrix inequality:

$$\|Z\| \leq t \quad \Longleftrightarrow \quad t^2 I_m - ZZ' \succeq 0 \quad \Longleftrightarrow \quad \begin{bmatrix} tI_m & Z \\ Z' & tI_n \end{bmatrix} \succeq 0, \tag{2.3}$$

where the last implication follows from a Schur complement argument. As a consequence, we can give a semidefinite optimization characterization of the operator norm, namely

$$\|Z\| = \inf_t t \qquad \text{subject to} \quad \begin{bmatrix} tI_m & Z \\ Z' & tI_n \end{bmatrix} \succeq 0. \tag{2.4}$$

Now let $X = U\Sigma V'$ be a singular value decomposition of an $m \times n$ matrix X, where U is an $m \times r$ matrix, V is an $n \times r$ matrix, Σ is a $r \times r$ diagonal matrix and r is the rank of X. Let $Y := UV'$. Then $\|Y\| = 1$ and $\mathrm{Tr}(XY') = \sum_{i=1}^r \sigma_i(X) = ||X||_*$, and hence the dual norm is greater than or equal to the nuclear norm.

To provide an upper bound on the dual norm, we appeal to semidefinite programming duality. From the characterization in (2.3), the optimization problem

$$\begin{array}{ll} \text{maximize} & \mathrm{Tr}(X'Y) \\ \text{subject to} & \|Y\| \leq 1 \end{array}$$

is equivalent to the semidefinite program

$$\begin{array}{ll} \text{maximize} & \mathrm{Tr}(X'Y) \\ \text{subject to} & \begin{bmatrix} I_m & Y \\ Y' & I_n \end{bmatrix} \succeq 0. \end{array} \tag{2.5}$$

The dual of this SDP (after an inconsequential rescaling) is given by

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}(\mathrm{Tr}(W_1) + \mathrm{Tr}(W_2)) \\ \text{subject to} & \begin{bmatrix} W_1 & X \\ X' & W_2 \end{bmatrix} \succeq 0. \end{array} \tag{2.6}$$

Set $W_1 := U\Sigma U'$ and $W_2 := V\Sigma V'$. Then the triple (W_1, W_2, X) is feasible for (2.6) since

$$\begin{bmatrix} W_1 & X \\ X' & W_2 \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} \Sigma \begin{bmatrix} U \\ V \end{bmatrix}' \succeq 0.$$

Furthermore, we have $\mathrm{Tr}(W_1) = \mathrm{Tr}(W_2) = \mathrm{Tr}(\Sigma)$, and thus the objective function satisfies $(\mathrm{Tr}(W_1) + \mathrm{Tr}(W_2))/2 = \mathrm{Tr}\,\Sigma = \|X\|_*$. Since any feasible solution of (2.6) provides an upper bound for (2.5), we have that the dual norm is less than or equal to the nuclear norm of X, thus proving the proposition. ■

Notice that the argument given in the proof above further shows that the nuclear norm $||X||_*$ can be computed using either the SDP (2.5) or its dual (2.6), since there is no duality gap between them. Alternatively, this could have also been proven using a Slater-type interior point condition since both (2.5) and (2.6) admit strictly feasible solutions. Interested readers can find an in-depth discussion of Slater conditions for semidefinite programming in Chapter 4 of [83].

Convex envelopes of rank and cardinality functions Let $\mathcal{C}$ be a given convex set. The *convex envelope* of a (possibly nonconvex) function $f : \mathcal{C} \to \mathbb{R}$ is defined as the largest convex function g such that $g(x) \leq f(x)$ for all $x \in \mathcal{C}$ (see, e.g., [44]). This means that among all convex functions, g is the best pointwise approximation to f. In particular, if g can be conveniently described, it can serve as an approximation to f that can be minimized efficiently.

By the chain of inequalities in (2.1), we have that $\operatorname{rank}(X) \geq \|X\|_*/\|X\|$ for all X. For all matrices with $\|X\| \leq 1$, we must have that $\operatorname{rank}(X) \geq \|X\|_*$, so the nuclear norm is a convex lower bound of the rank function on the unit ball in the operator norm. In fact, it can be shown that this is the tightest convex lower bound.

Theorem 2.2 ([37]) *The convex envelope of* $\operatorname{rank}(X)$ *on the set* $\{X \in \mathbb{R}^{m\times n} : \|X\| \leq 1\}$ *is the nuclear norm* $\|X\|_*$.

The proof of this statement is given in [37]. It relies on a basic result from convex analysis that establishes that, under certain technical conditions, the biconjugate of a function is its convex envelope [44].

Theorem 2.2 provides the following interpretation of the nuclear norm heuristic for the affine rank minimization problem. Suppose X_0 is the minimum rank solution of $\mathcal{A}(X) = b$, and $M = \|X_0\|$. The convex envelope of the rank on the set $\mathcal{C} = \{X \in \mathbb{R}^{m\times n} : \|X\| \leq M\}$ is $\|X\|_*/M$. Let X_* be the minimum nuclear norm solution of $\mathcal{A}(X) = b$. Then we have

$$\|X_*\|_*/M \leq \operatorname{rank}(X_0) \leq \operatorname{rank}(X_*)$$

providing an upper and lower bound on the optimal rank when the norm of the optimal solution is known. Furthermore, this is the tightest lower bound among all convex lower bounds of the rank function on the set $\mathcal{C}$.

For vectors, we have a similar inequality. Let $\operatorname{card}(x)$ denote the cardinality function which counts the number of non-zero entries in the vector x. Then we have $\operatorname{card}(x) \geq \|x\|_1/\|x\|_\infty$. Not surprisingly, the ℓ_1 norm is also the convex envelope of the cardinality function over the set $\{x \in \mathbb{R}^n : \|x\|_\infty \leq 1\}$. This result can be either proven directly or can be seen as a special case of the above theorem.

Additivity of rank and nuclear norm A function f mapping a linear space into $\mathbb{R}$ is called *subadditive* if $f(x+y) \leq f(x) + f(y)$. It is *additive* if $f(x+y) = f(x) + f(y)$. In the case of vectors, both the cardinality function and the ℓ_1 norm are subadditive. That is, if x and y are sparse vectors, then it always holds that the number of non-zeros in $x + y$ is less than or equal to the number of non-zeros in x plus the number of non-zeros of y; furthermore (by the triangle inequality) $\|x+y\|_1 \leq \|x\|_1 + \|y\|_1$. In particular, the cardinality function is additive exactly when the vectors x and y have disjoint support. In this case, the ℓ_1 norm is also additive, in the sense that $\|x+y\|_1 = \|x\|_1 + \|y\|_1$.

For matrices, the rank function is subadditive. For the rank to be additive, it is necessary and sufficient that the row and column spaces of the two matrices intersect only at the origin, since in this case they operate in essentially disjoint spaces (see, e.g., [53]). As we will show below, a related condition that ensures that the nuclear norm is additive, is that the matrices A and B have row and column spaces that are *orthogonal.* In fact, a compact sufficient condition for the additivity of the nuclear norm will be that $AB' = 0$ and $A'B = 0$. This is a stronger requirement than the aforementioned condition for rank additivity, as orthogonal subspaces only intersect at the origin. The disparity arises because the nuclear norm of a linear map depends on the choice of the inner products on the spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ on which the matrix acts, whereas the rank is independent of such a choice.

Lemma 2.3 *Let A and B be matrices of the same dimensions. If $AB' = 0$ and $A'B = 0$ then $\|A + B\|_* = \|A\|_* + \|B\|_*$.*

Proof Partition the singular value decompositions of A and B to reflect the zero and non-zero singular vectors

$$A = \begin{bmatrix} U_{A1} & U_{A2} \end{bmatrix} \begin{bmatrix} \Sigma_A & \\ & 0 \end{bmatrix} \begin{bmatrix} V_{A1} & V_{A2} \end{bmatrix}' \qquad B = \begin{bmatrix} U_{B1} & U_{B2} \end{bmatrix} \begin{bmatrix} \Sigma_B & \\ & 0 \end{bmatrix} \begin{bmatrix} V_{B1} & V_{B2} \end{bmatrix}' .$$

The condition $AB' = 0$ implies that $V_{A1}'V_{B1} = 0$, and similarly, $A'B = 0$ implies that $U_{A1}'U_{B1} = 0$. Hence, there exist matrices U_C and V_C such that $[U_{A1}\, U_{B1}\, U_C]$ and $[V_{A1}\, V_{B1}\, V_C]$ are orthogonal matrices. Thus, the following are valid singular value decompositions for A and B:

$$A = \begin{bmatrix} U_{A1} & U_{B1} & U_C \end{bmatrix} \begin{bmatrix} \Sigma_A & & \\ & 0 & \\ & & 0 \end{bmatrix} \begin{bmatrix} V_{A1} & V_{B1} & V_C \end{bmatrix}'$$

$$B = \begin{bmatrix} U_{A1} & U_{B1} & U_C \end{bmatrix} \begin{bmatrix} 0 & & \\ & \Sigma_B & \\ & & 0 \end{bmatrix} \begin{bmatrix} V_{A1} & V_{B1} & V_C \end{bmatrix}' .$$

In particular, we have that

$$A + B = \begin{bmatrix} U_{A1} & U_{B1} \end{bmatrix} \begin{bmatrix} \Sigma_A & \\ & \Sigma_B \end{bmatrix} \begin{bmatrix} V_{A1} & V_{B1} \end{bmatrix}' .$$

This shows that the singular values of $A+B$ are equal to the union (with repetition) of the singular values of A and B. Hence, $\|A + B\|_* = \|A\|_* + \|B\|_*$ as desired. ■

Corollary 2.4 *Let A and B be matrices of the same dimensions. If the row and column spaces of A and B are orthogonal, then $\|A + B\|_* = \|A\|_* + \|B\|_*$.*

Proof It suffices to show that if the row and column spaces of A and B are orthogonal, then $AB' = 0$ and $A'B = 0$. But this is immediate: if the columns of A are orthogonal to the columns of B, we have $A'B = 0$. Similarly, orthogonal row spaces imply that $AB' = 0$ as well. ■

Nuclear norm minimization Let us turn now to the study of equality-constrained norm minimization problems where we are searching for a matrix $X \in \mathbb{R}^{m\times n}$ of minimum nuclear norm belonging to a given affine subspace. In our applications, the subspace is usually described by linear equations of the form $\mathcal{A}(X) = b$, where $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ is a linear mapping. This problem admits the primal-dual convex formulation

$$\begin{array}{llcll} \text{minimize} & \|X\|_* & \qquad & \text{maximize} & b'z \\ \text{subject to} & \mathcal{A}(X) = b & & \text{subject to} & ||\mathcal{A}^*(z)|| \leq 1, \end{array} \tag{2.7}$$

where $\mathcal{A}^* : \mathbb{R}^p \to \mathbb{R}^{m\times n}$ is the adjoint of $\mathcal{A}$. By a primal-dual pair, we mean that each optimization problem is the Lagrangian dual of the other, and hence that the minimum of the first optimization problem is equal to the maximum of the second. This notion of duality generalizes the well-known case of linear programming, and is in fact applicable to all convex optimization problems; see e.g. [8, 9].

The formulation (2.7) is valid for any norm minimization problem, by replacing the norms appearing above by any dual pair of norms. In particular, if we replace the nuclear norm with the ℓ_1 norm and the operator norm with the ℓ_∞ norm, we obtain a primal-dual pair of optimization problems, that can be reformulated in terms of linear programming.

Using the SDP characterizations of the nuclear and operator norms given in (2.5)-(2.6) above allows us to rewrite (2.7) as the primal-dual pair of semidefinite programs

$$\begin{array}{llcll} \text{minimize} & \frac{1}{2}(\mathrm{Tr}(W_1) + \mathrm{Tr}(W_2)) & \qquad & \text{maximize} & b'z \\ \text{subject to} & \left[\begin{array}{cc} W_1 & X \\ X' & W_2 \end{array}\right] \succeq 0 & & \text{subject to} & \left[\begin{array}{cc} I_m & \mathcal{A}^*(z) \\ \mathcal{A}^*(z)' & I_n \end{array}\right] \succeq 0. \\ & \mathcal{A}(X) = b & & & \end{array} \tag{2.8}$$

Optimality conditions In order to describe the optimality conditions for the norm minimization problem (2.7), we must first characterize the set of all subgradients (i.e., the subdifferential) of the nuclear norm. Recall that for a convex function $f : \mathbb{R}^n \to \mathbb{R}$, the subdifferential of f at $x \in \mathbb{R}^n$ is the compact convex set

$$\partial f(x) := \{d \in \mathbb{R}^n \,:\, f(y) \geq f(x) + \langle d, y - x\rangle \quad \forall y \in \mathbb{R}^n\}.$$

Let X be an $m \times n$ matrix with rank r and let $X = U\Sigma V'$ be a singular value decomposition where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$ and Σ is an $r \times r$ diagonal matrix. The subdifferential of the nuclear norm at X is then given by (e.g., [46, 82])

$$\partial\|X\|_* = \{UV' + W \,:\, W \text{ and } X \text{ have orthogonal row and column spaces, and } \|W\| \leq 1\}. \tag{2.9}$$

For comparison, recall the case of the ℓ_1 norm, where T denotes the support of the n-vector x, T^c is the complement of T in the set $\{1, \dots, n\}$, and

$$\partial\|x\|_1 = \{d \in \mathbb{R}^n \,:\, d_i = \mathrm{sign}(x) \text{ for } i \in T, \quad |d_i| \leq 1 \text{ for } i \in T^c\}. \tag{2.10}$$

The similarity between (2.9) and (2.10) is particularly transparent if we recall the *polar decomposition* of a matrix into a product of orthogonal and positive semidefinite matrices (see, e.g., [45]). The "angular" component of the matrix X is exactly given by UV'. Thus, these subgradients always have the form of an "angle" (or sign), plus possibly a contraction in an orthogonal direction if the norm is not differentiable at the current point.

We can now write concise optimality conditions for the optimization problem (2.7). A matrix X is an optimal solution for (2.7) if there exists a vector $z \in \mathbb{R}^p$ such that

$$\mathcal{A}(X) = b, \qquad \mathcal{A}^*(z) \in \partial\|X\|_*. \tag{2.11}$$

The first condition in (2.11) requires feasibility of the linear equations, and the second one guarantees that there is no feasible direction of improvement. Indeed, since $\mathcal{A}^*(z)$ is in the subdifferential at X, for any Y in the primal feasible set of (2.7) we have

$$||Y||_* \geq ||X||_* + \langle \mathcal{A}^*(z), Y - X\rangle = ||X||_* + \langle z, \mathcal{A}(Y - X)\rangle = ||X||_*,$$

where the last step follows from the feasibility of X and Y. As we can see, the optimality conditions (2.11) for the nuclear norm minimization problem exactly parallel those of the ℓ_1 optimization case.

These optimality conditions can be used to check and certify whether a given candidate X is indeed a minimum nuclear norm solution. For this, it is sufficient (and necessary) to find a vector $z \in \mathbb{R}^p$ in the subdifferential of the norm, i.e., such that the left- and right-singular spaces of $\mathcal{A}^*(z)$ are aligned with those of X, and is a contraction in the orthogonal complement.

3 Restricted isometry and recovery of low-rank matrices

Let us now turn to the central problem analyzed in this paper. Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ be a linear map and let X_0 be a matrix of rank r. Set $b := \mathcal{A}(X_0)$, and define

$$X^* := \arg\min_X \|X\|_* \qquad \text{s.t.} \quad \mathcal{A}(X) = b. \tag{3.1}$$

That is, X^* is the member of the affine space defined by $\mathcal{A}$ and b with smallest nuclear norm. In this section, we will characterize specific cases when we can *a priori* guarantee that $X^* = X_0$. The key conditions will be determined by the values of a sequence of parameters δ_r that quantify the behavior of the linear map $\mathcal{A}$ when restricted to the set of matrices of rank r. The following definition is the natural generalization of the Restricted Isometry Property from vectors to matrices.

Definition 3.1 *Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ be a linear map. Without loss of generality, assume $m \leq n$. For every integer r with $1 \leq r \leq m$, define the r-restricted isometry constant to be the smallest number $\delta_r(\mathcal{A})$ such that*

$$(1 - \delta_r(\mathcal{A}))\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1 + \delta_r(\mathcal{A}))\|X\|_F \tag{3.2}$$

holds for all matrices X of rank at most r.

Note that by definition, $\delta_r(\mathcal{A}) \leq \delta_{r'}(\mathcal{A})$ for $r \leq r'$.

The Restricted Isometry Property for sparse vectors was developed by Candès and Tao in [16], and requires (3.2) to hold with the Euclidean norm replacing the Frobenius norm and rank being replaced by cardinality. Since for diagonal matrices, the Frobenius norm is equal to the Euclidean norm of the diagonal, this definition reduces to the original Restricted Isometry Property of [16] in the diagonal case.[1]

Unlike the case of "standard" compressed sensing, our RIP condition for low-rank matrices cannot be interpreted as guaranteeing all sub-matrices of the linear transform $\mathcal{A}$ of a certain size are well conditioned. Indeed, the set of matrices X for which (3.2) must hold is *not* a finite union of subspaces, but rather a certain "generalized Stiefel manifold," which is also an algebraic variety (in fact, it is the rth-secant variety of the variety of rank-one matrices). Surprisingly, we are still able to derive analogous recovery results for low-rank solutions of equations when $\mathcal{A}$ obeys this RIP condition. Furthermore, we will see in Section 4 that many ensembles of random matrices have the Restricted Isometry Property with δ_r quite small with high probability for reasonable values of m,n, and p.

[1] In [16], the authors define the restricted isometry properties with squared norms. We note here that the analysis is identical modulo some algebraic rescaling of constants. We choose to drop the squares as it greatly simplifies the analysis in Section 4.

The following two recovery theorems will characterize the power of the restricted isometry constants. Both theorems are more or less immediate generalizations from the sparse case to the low-rank case and use only minimal properties of the rank of matrices and the nuclear norm. The first theorem generalizes Lemma 1.3 in [16] to low-rank recovery.

Theorem 3.2 *Suppose that $\delta_{2r} < 1$ for some integer $r \geq 1$. Then X_0 is the only matrix of rank at most r satisfying $\mathcal{A}(X) = b$.*

Proof Assume, on the contrary, that there exists a rank r matrix X satisfying $\mathcal{A}(X) = b$ and $X \neq X_0$. Then $Z := X_0 - X$ is a nonzero matrix of rank at most $2r$, and $\mathcal{A}(Z) = 0$. But then we would have $0 = \|\mathcal{A}(Z)\| \geq (1 - \delta_{2r})\|Z\|_F > 0$ which is a contradiction. ■

The proof of the preceding theorem is identical to the argument given by Candès and Tao and is an immediate consequence of our definition of the constant δ_r. No adjustment is necessary in the transition from sparse vectors to low-rank matrices. The key property used is the sub-additivity of the rank.

Next, we state a simple condition which guarantees $X^* = X_0$. The proof follows the approach in [14], but a few details need to be adjusted when switching from vectors to matrices.

Theorem 3.3 *Suppose that $r \geq 1$ is such that $\delta_{5r} < 1/10$. Then $X^* = X_0$.*

We will need the following technical lemma that shows for any two matrices A and B, we can decompose B as the sum of two matrices B_1 and B_2 such that $\operatorname{rank}(B_1)$ is not too large and B_2 satisfies the conditions of Lemma 2.3. This will be the key decomposition for proving Theorem 3.3.

Lemma 3.4 *Let A and B be matrices of the same dimensions. Then there exist matrices B_1 and B_2 such that*

1. $B = B_1 + B_2$
2. $\operatorname{rank}(B_1) \leq 2\operatorname{rank}(A)$
3. $AB_2' = 0$ *and* $A'B_2 = 0$
4. $\langle B_1, B_2 \rangle = 0$

Proof Consider a full singular value decomposition of A

$$A = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V',$$

and let $\hat{B} := U'BV$. Partition $\hat{B}$ as

$$\hat{B} = \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{21} & \hat{B}_{22} \end{bmatrix}.$$

Defining now

$$B_1 := U \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{21} & 0 \end{bmatrix} V', \qquad B_2 := U \begin{bmatrix} 0 & 0 \\ 0 & \hat{B}_{22} \end{bmatrix} V',$$

it can be easily verified that B_1 and B_2 satisfy the conditions (1)–(4). ■

We now proceed to a proof of Theorem 3.3.

Proof [of Theorem 3.3] By optimality of X^*, we have $\|X_0\|_* \geq \|X^*\|_*$. Let $R := X^* - X_0$. Applying Lemma 3.4 to the matrices X_0 and R, there exist matrices R_0 and R_c such that $R = R_0 + R_c$, $\operatorname{rank}(R_0) \leq 2\operatorname{rank}(X_0)$, and $X_0 R_c' = 0$ and $X_0' R_c = 0$. Then,

$$\|X_0\|_* \geq \|X_0 + R\|_* \geq \|X_0 + R_c\|_* - \|R_0\|_* = \|X_0\|_* + \|R_c\|_* - \|R_0\|_*, \tag{3.3}$$

where the middle assertion follows from the triangle inequality and the last one from Lemma 2.3. Rearranging terms, we can conclude that

$$\|R_0\|_* \geq \|R_c\|_*. \tag{3.4}$$

Next we partition R_c into a sum of matrices $R_1, R_2, \ldots$, each of rank at most $3r$. Let $R_c = U \operatorname{diag}(\sigma) V'$ be the singular value decomposition of R_c. For each $i \geq 1$ define the index set $I_i = \{3r(i-1)+1, \ldots, 3ri\}$, and let $R_i := U_{I_i} \operatorname{diag}(\sigma_{I_i}) V_{I_i}'$ (notice that $\langle R_i, R_j \rangle = 0$ if $i \neq j$). By construction, we have

$$\sigma_k \leq \frac{1}{3r} \sum_{j \in I_i} \sigma_j \qquad \forall\, k \in I_{i+1}, \tag{3.5}$$

which implies $\|R_{i+1}\|_F^2 \leq \frac{1}{3r}\|R_i\|_*^2$. We can then compute the following bound

$$\sum_{j\geq 2} \|R_j\|_F \leq \frac{1}{\sqrt{3r}} \sum_{j \geq 1} \|R_j\|_* = \frac{1}{\sqrt{3r}} \|R_c\|_* \leq \frac{1}{\sqrt{3r}} \|R_0\|_* \leq \frac{\sqrt{2r}}{\sqrt{3r}} \|R_0\|_F\,, \tag{3.6}$$

where the last inequality follows from (2.1) and the fact that $\operatorname{rank}(R_0) \leq 2r$. Finally, note that the rank of $R_0 + R_1$ is at most $5r$, so we may put this all together as

$$\begin{aligned} \|\mathcal{A}(R)\| &\geq \|\mathcal{A}(R_0 + R_1)\| - \sum_{j \geq 2} \|\mathcal{A}(R_j)\| \\ &\geq (1 - \delta_{5r})\, \|R_0 + R_1\|_F - (1 + \delta_{3r}) \sum_{j \geq 2} \|R_j\|_F \\ &\geq \left((1 - \delta_{5r}) - \sqrt{\tfrac{2}{3}}(1 + \delta_{3r}) \right) \|R_0\|_F \\ &\geq \big((1 - \delta_{5r}) - \tfrac{9}{11}(1 + \delta_{3r}) \big)\, \|R_0\|_F. \end{aligned} \tag{3.7}$$

By assumption $\mathcal{A}(R) = \mathcal{A}(X^* - X_0) = 0$, so if the factor on the right-hand side is strictly positive, $R_0 = 0$, which further implies $R_c = 0$ by (3.4), and thus $X^* = X_0$. Simple algebra reveals that the right-hand side is positive when $9\delta_{3r} + 11\delta_{5r} < 2$. Since $\delta_{3r} \leq \delta_{5r}$, we immediately have that $X^* = X_0$ if $\delta_{5r} < 1/10$. ■

The rational number (9/11) in the proof of the theorem is chosen for notational simplicity and is clearly not optimal. A slightly tighter bound can be achieved working directly with the second to last line in (3.7). The most important point is that our recovery condition on δ_{5r} is an absolute constant, independent of m, n, r, and p.

We have yet to demonstrate any specific linear mappings $\mathcal{A}$ for which $\delta_r < 1$. We shall show in the next section that random linear transformations sampled from several distributions of matrices with appropriately chosen dimensions have this property with overwhelming probability. The analysis is again similar to the compressive sampling literature, but several details specific to the rank recovery problem need to be employed.

4 Nearly isometric random matrices

In this section, we will demonstrate that when we sample linear maps from a class of probability distributions obeying certain large deviation inequalities, then they will obey the Restricted Isometry Property (3.2) as p, m, and n tend to infinity at appropriate rates. The following definition characterizes this family of random linear transformations.

Definition 4.1 *Let $\mathcal{A}$ be a random variable that takes values in linear maps from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^p$. We say that $\mathcal{A}$ is* nearly isometrically distributed *if for all $X \in \mathbb{R}^{m\times n}$*

$$\mathbf{E}[\|\mathcal{A}(X)\|^2] = \|X\|_F^2 \tag{4.1}$$

and for all $0 < \epsilon < 1$ we have,

$$\mathbf{P}(|\|\mathcal{A}(X)\|^2 - \|X\|_F^2| \geq \epsilon\|X\|_F^2) \leq 2\exp\left(-\frac{p}{2}(\epsilon^2/2 - \epsilon^3/3)\right) \tag{4.2}$$

and for all $t > 0$, we have

$$\mathbf{P}\left(\|\mathcal{A}\| \geq 1 + \sqrt{\frac{mn}{p}} + t\right) \leq \exp(-\gamma p t^2) \tag{4.3}$$

for some constant $\gamma > 0$.

There are two ingredients for a random linear map to be nearly isometric. First, it must be isometric in expectation. Second, the probability of large distortions of length must be exponentially small. The exponential bound in (4.2) guarantees union bounds will be small even for rather large sets. This concentration is the typical ingredient required to prove the Johnson-Lindenstrauss Lemma (cf. [2, 22]).

It is often simpler to describe nearly isometric random maps in terms of random matrices. For a linear map $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, we can always write its matrix representation as

$$\mathcal{A}(X) = \mathbf{A}\,\mathrm{vec}(X)\,, \tag{4.4}$$

where $\mathrm{vec}(X)$ denotes the vector of X with its columns stacked in order on top of one another, and $\mathbf{A}$ is a $p \times mn$ matrix. We now give several examples of nearly isometric random variables in this matrix representation. The most well known is the ensemble with independent, identically distributed (i.i.d.) Gaussian entries [22]

$$A_{ij} \sim \mathcal{N}(0, \frac{1}{p})\,. \tag{4.5}$$

We also mention two ensembles of matrices described in [2]. One has entries sampled from an i.i.d. symmetric Bernoulli distribution

$$A_{ij} = \begin{cases} \sqrt{\frac{1}{p}} & \text{with probability } \frac{1}{2} \\ -\sqrt{\frac{1}{p}} & \text{with probability } \frac{1}{2} \end{cases}, \tag{4.6}$$

and the other has zeros in two-thirds of the entries

$$A_{ij} = \begin{cases} \sqrt{\frac{3}{p}} & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -\sqrt{\frac{3}{p}} & \text{with probability } \frac{1}{6} \end{cases}. \tag{4.7}$$

The fact that the top singular value of the matrix $\mathbf{A}$ is concentrated around $1 + \sqrt{D/p}$ for all of these ensembles follows from the work of Yin, Bai, and Krishnaiah, who showed that whenever the entries A_{ij} are i.i.d. with zero mean and finite fourth moment, then the maximum singular value of $\mathbf{A}$ is almost surely $1 + \sqrt{D/p}$ for D sufficiently large [84]. El Karoui uses this result to prove the concentration inequality (4.3) for all such distributions [32]. The result for Gaussians is rather tight with $\gamma = 1/2$ (see, e.g., [23]).

Finally, note that a random projection also obeys all of the necessary concentration inequalities. Indeed, since the norm of a random projection is exactly $\sqrt{D/p}$, (4.3) holds trivially. The concentration inequality (4.2) is proven in [22].

The main result of this section is the following:

Theorem 4.2 *Fix $0 < \delta < 1$. If $\mathcal{A}$ is a nearly isometric random variable, then for every $1 \leq r \leq m$, there exist positive constants c_0, c_1 depending only on δ such that, with probability at least $1 - \exp(-c_1 p)$, $\delta_r(\mathcal{A}) \leq \delta$ whenever $p \geq c_0 r(m+n)\log(mn)$.*

The proof will make use of standard techniques in concentration of measure. We first extend the concentration results of [4] to subspaces of matrices. We will show that the distortion of a subspace by a linear map is robust to perturbations of the subspace. Finally, we will provide an epsilon net over the set of all subspaces and, using a union bound, will show that with overwhelming probability, nearly isometric random variables will obey the Restricted Isometry Property (3.2) as the size of the matrices tend to infinity.

The following lemma characterizes the behavior of a nearly isometric random mapping $\mathcal{A}$ when restricted to an arbitrary subspace of matrices U of dimension d.

Lemma 4.3 *Let $\mathcal{A}$ be a nearly isometric linear map and let U be an arbitrary subspace of $m \times n$ matrices with $d = \dim(U) \leq p$. Then for any $0 < \delta < 1$ we have*

$$(1-\delta)\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1+\delta)\|X\|_F \qquad \forall\, X \in U \tag{4.8}$$

with probability at least

$$1 - 2(12/\delta)^d \exp\left(-\frac{p}{2}(\delta^2/8 - \delta^3/24)\right). \tag{4.9}$$

Proof The proof of this theorem is identical to the argument in [4] where the authors restricted their attention to subspaces aligned with the coordinate axes. We will sketch the proof here as the argument is straightforward.

There exists a finite set Ω of at most $(12/\delta)^d$ points such that for every $X \in U$ with $\|X\|_F \leq 1$, there exists a $Q \in \Omega$ such that $\|X - Q\|_F \leq \delta/4$. By (4.2) and the standard union bound, we have that $(1-\delta/2)\|Q\|_F \leq \|\mathcal{A}(Q)\| \leq (1+\delta/2)\|Q\|_F$ for all $Q \in \Omega$ probability at least (4.9).

Let X be in $\{X \in U : \|X\|_F \leq 1\}$, and M be the maximum of $\|\mathcal{A}(X)\|$ on this set. Then there exists a $Q \in \Omega$ such that $\|X - Q\|_F \leq \delta/4$. We then have

$$\|\mathcal{A}(X)\| \leq \|\mathcal{A}(Q)\| + \|\mathcal{A}(X-Q)\| \leq 1 + \delta/2 + M\delta/4 \,.$$

Taking the supremum of both sides of this inequality yields $M \leq 1 + \delta/2 + M\delta/4$, and we thus have $M \leq 1 + \delta$. The lower bound is proven by the following chain of inequalities

$$\|\mathcal{A}(X)\| \geq \|\mathcal{A}(Q)\| - \|\mathcal{A}(X-Q)\| \geq 1 - \delta/2 - (1+\delta)\delta/4 \geq 1 - \delta.$$

■

The proof of preceding lemma revealed that the near isometry of a linear map is robust to small perturbations of the matrix on which the map is acting. We will now show that this behavior is robust with respect to small perturbations of the subspace U as well. This perturbation will be measured in the natural distance between two subspaces

$$\rho(T_1, T_2) := \|P_{T_1} - P_{T_2}\|, \tag{4.10}$$

where T_1 and T_2 are subspaces and P_{T_i} is the orthogonal projection associated with each subspace. This distance measures the operator norm of the difference between the corresponding projections, and is equal to the sine of the largest principal angle between T_1 and T_2 [1].

The set of all d-dimensional subspaces of $\mathbb{R}^D$ is commonly known as the Grassmannian manifold $\mathfrak{G}(D,d)$. We will endow it with the metric $\rho(\cdot,\cdot)$ given by (4.10), also known as the *projection 2-norm.* In the following lemma we characterize and quantify the change in the isometry constant δ as one smoothly moves through the Grassmannian.

Lemma 4.4 *Let U_1 and U_2 be d-dimensional subspaces of $\mathbb{R}^D$. Suppose that for all $X \in U_1$,*

$$(1-\delta)\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1+\delta)\|X\|_F \tag{4.11}$$

for some constant $0 < \delta < 1$. Then for all $Y \in U_2$

$$(1-\delta')\|Y\|_F \leq \|\mathcal{A}(Y)\| \leq (1+\delta')\|Y\|_F \tag{4.12}$$

with

$$\delta' = \delta + (1 + \|\mathcal{A}\|) \cdot \rho(U_1, U_2) \,. \tag{4.13}$$

Proof Consider any $Y \in U_2$. Then

$$\begin{aligned} \|\mathcal{A}(Y)\| &= \|\mathcal{A}\left(P_{U_1}(Y) - [P_{U_1} - P_{U_2}](Y)\right)\| \\ &\leq \|\mathcal{A}(P_{U_1}(Y))\| + \|\mathcal{A}\left([P_{U_1} - P_{U_2}](Y)\right)\| \\ &\leq (1+\delta)\|P_{U_1}(Y)\|_F + \|\mathcal{A}\|\|P_{U_1} - P_{U_2}\|\|Y\|_F \\ &\leq \left(1 + \delta + \|\mathcal{A}\|\|P_{U_1} - P_{U_2}\|\right)\|Y\|_F. \end{aligned} \tag{4.14}$$

Similarly, we have

$$\begin{aligned}\|\mathcal{A}(Y)\| &\geq \|\mathcal{A}(P_{U_1}(Y))\| - \|\mathcal{A}\left([P_{U_1} - P_{U_2}](Y)\right)\| \\ &\geq (1-\delta)\|P_{U_1}(Y)\|_F - \|\mathcal{A}\|\|P_{U_1} - P_{U_2}\|\|Y\|_F \\ &\geq (1-\delta)\|Y\|_F - (1-\delta)\|(P_{U_1} - P_{U_2})(Y)\|_F - \|\mathcal{A}\|\|P_{U_1} - P_{U_2}\|\|Y\|_F \\ &\geq \left[1 - \delta - (\|\mathcal{A}\|+1)\|P_{U_1} - P_{U_2}\|\right]\|Y\|_F,\end{aligned} \tag{4.15}$$

which completes the proof. ■

To apply these concentration results to low-rank matrices, we characterize the set of all matrices of rank at most r as a union of subspaces. Let $V \subset \mathbb{R}^m$ and $W \subset \mathbb{R}^n$ be fixed subspaces of dimension r. Then the set of all $m \times n$ matrices X whose row space is contained in W and column space is contained in V forms an r^2-dimensional subspace of matrices of rank less than or equal to r. Denote this subspace as $\Sigma(V,W) \subset \mathbb{R}^{m\times n}$. Any matrix of rank less than or equal to r is an element of some $\Sigma(V,W)$ for a suitable pair of subspaces, i.e., the set

$$\Sigma_{mnr} := \{\Sigma(V,W) \quad : \quad V \in \mathfrak{G}(m,r), \quad W \in \mathfrak{G}(n,r)\}.$$

We now characterize how many subspaces are necessary to cover this set to arbitrary resolution. The *covering number* $\mathfrak{N}(\epsilon)$ of Σ_{mnr} at resolution ϵ is defined to be the smallest number of subspaces (V_i, W_i) such that for any pair of subspaces (V,W), there is an i with $\rho(\Sigma(V,W),\Sigma(V_i,W_i)) \leq \epsilon$. That is, the covering number is the smallest cardinality of an ϵ-net. The following Lemma characterizes the cardinality of such a set.

Lemma 4.5 *The covering number $\mathfrak{N}(\epsilon)$ of Σ_{mnr} is bounded above by*

$$\mathfrak{N}(\epsilon) \leq \left(\frac{2C_0}{\epsilon}\right)^{r(m+n-2r)} \tag{4.16}$$

where C_0 is a constant independent of ϵ, m, n, and r.

Proof Note that the projection operator onto $\Sigma(V,W)$ can be written as $P_{\Sigma(V,W)} = P_V \otimes P_W$, so for a pair of subspaces (V_1,W_1) and (V_2,W_2), we have

$$\begin{aligned}\rho(\Sigma(V_1,W_1),\Sigma(V_2,W_2)) &= \|P_{\Sigma(V_1,W_1)} - P_{\Sigma(V_2,W_2)}\| \\ &= \|P_{V_1}\otimes P_{W_1} - P_{V_2}\otimes P_{W_2}\| \\ &= \|(P_{V_1} - P_{V_2})\otimes P_{W_1} + P_{V_2}\otimes(P_{W_1} - P_{W_2})\| \\ &\leq \|P_{V_1} - P_{V_2}\|\|P_{W_1}\| + \|P_{V_2}\|\|P_{W_1} - P_{W_2}\| \\ &= \rho(V_1,V_2) + \rho(W_1,W_2).\end{aligned} \tag{4.17}$$

The conditions $\rho(V_1,V_2) \leq \frac{\epsilon}{2}$ and $\rho(W_1,W_2) \leq \frac{\epsilon}{2}$ together imply that $\rho(\Sigma(V_1,W_1),\Sigma(V_2,W_2)) \leq \rho(V_1,V_2) + \rho(W_1,W_2) \leq \epsilon$. Let $V_1,\ldots,V_{N_1}$ cover the set of r-dimensional subspaces of $\mathbb{R}^m$ to resolution $\epsilon/2$ and $U_1,\ldots,U_{N_2}$ cover the r-dimensional subspaces of $\mathbb{R}^n$ to resolution $\epsilon/2$. Then for any (V,W), there exist i and j such that $\rho(V,V_i) \leq \epsilon/2$ and $\rho(W,W_j) \leq \epsilon/2$. Therefore, $\mathfrak{N}(\epsilon) \leq N_1 N_2$. By the work of Szarek on ϵ-nets of the Grassmannian ([72], [73, Th. 8]) there is a universal constant C_0, independent of ϵ, m, n, and r, such that

$$N_1 \leq \left(\frac{2C_0}{\epsilon}\right)^{r(m-r)} \qquad \text{and} \qquad N_2 \leq \left(\frac{2C_0}{\epsilon}\right)^{r(n-r)} \tag{4.18}$$

which completes the proof. ■

The exact value of the universal constant C_0 is not provided by Szarek in [73]. It takes the same value for any homogeneous space whose automorphism group is a subgroup of the orthogonal group, and is independent of the dimension of the homogeneous space. Hence, one might expect this constant to be quite large. However, it is known that for the sphere $C_0 \leq 3$ [50], and there is no indication that this constant is not similarly small for the Grassmannian.

We now proceed to the proof of the main result in this section. For this, we use a union bound to combine the probabilistic guarantees of Lemma 4.3 with the estimates of the covering number of $\Sigma(U, V)$.

Proof [of Theorem 4.2]

Let $\Omega = \{(V_i, W_i)\}$ be a finite set of subspaces that satisfies the conditions of Lemma 4.5 for $\epsilon > 0$, so $|\Omega| \leq \mathfrak{N}(\epsilon)$. For each pair (V_i, W_i), define the set of matrices

$$\mathcal{B}_i := \left\{ X \;\middle|\; \exists (V, W) \text{ such that } X \in \Sigma(V, W) \text{ and } \rho(\Sigma(V, W), \Sigma(V_i, W_i)) \leq \epsilon \right\}. \tag{4.19}$$

Since Ω is an ϵ-net, we have that the union of all the $\mathcal{B}_i$ is equal to Σ_{mnr}. Therefore, if for all i, $(1-\delta)\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1+\delta)\|X\|_F$ for all $X \in \mathcal{B}_i$, we must have that $\delta_r(\mathcal{A}) \leq \delta$ proving that

$$\begin{aligned} \mathbf{P}(\delta_r(\mathcal{A}) \leq \delta) &= \mathbf{P}\left[(1-\delta)\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1+\delta)\|X\|_F \quad \forall\, X \text{ s.t. } \operatorname{rank}(X) \leq r\right] \\ &\geq \mathbf{P}\left[\forall i \ (1-\delta)\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1+\delta)\|X\|_F \quad \forall\, X \in \mathcal{B}_i\right] \end{aligned} \tag{4.20}$$

Now note that if we have $(1 + \|\mathcal{A}\|)\epsilon \leq \delta/2$ and, for all $Y \in \Sigma(V_i, W_i)$, $(1-\delta/2)\|Y\|_F \leq \|\mathcal{A}(Y)\| \leq (1+\delta/2)\|Y\|_F$, Lemma 4.4 implies that $(1-\delta)\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1+\delta)\|X\|_F$ for all $X \in \mathcal{B}_i$. Therefore, using a union bound, (4.20) is greater than or equal to

$$\begin{aligned} 1 - \sum_{i=1}^{|\Omega|} \mathbf{P}&\left[\exists\, Y \in \Sigma(V_i, W_i) \quad \|\mathcal{A}(Y)\| < (1-\frac{\delta}{2})\|Y\|_F \quad \text{or} \quad \|\mathcal{A}(Y)\| > (1+\frac{\delta}{2})\|Y\|_F\right] \\ &- \mathbf{P}\left[\|\mathcal{A}\| \geq \frac{\delta}{2\epsilon} - 1\right]. \end{aligned} \tag{4.21}$$

We can bound these quantities separately. First we have by Lemmas 4.3 and 4.5

$$\begin{aligned} \sum_{i=1}^{|\Omega|} \mathbf{P}&\left[\exists\, Y \in \Sigma(V_i, W_i) \quad \|\mathcal{A}(Y)\| < (1-\frac{\delta}{2})\|Y\|_F \quad \text{or} \quad \|\mathcal{A}(Y)\| > (1+\frac{\delta}{2})\|Y\|_F\right] \\ &\leq 2\mathfrak{N}(\epsilon) \left(\frac{24}{\delta}\right)^{r^2} \exp\left(-\frac{p}{2}(\delta^2/32 - \delta^3/192)\right) \\ &\leq 2\left(\frac{2C_0}{\epsilon}\right)^{r(m+n-2r)} \left(\frac{24}{\delta}\right)^{r^2} \exp\left(-\frac{p}{2}(\delta^2/32 - \delta^3/192)\right). \end{aligned} \tag{4.22}$$

Secondly, since $\mathcal{A}$ is nearly isometric, there exists a constant γ such that

$$\mathbf{P}\left(\|\mathcal{A}\| \geq 1 + \sqrt{\frac{mn}{p}} + t\right) \leq \exp(-\gamma p t^2). \tag{4.23}$$

In particular,

$$\mathbf{P}\left(\|\mathcal{A}\| \geq \frac{\delta}{2\epsilon} - 1\right) \leq \exp\left(-\gamma p \left(\frac{\delta}{2\epsilon} - \sqrt{\frac{mn}{p}} - 2\right)^2\right) . \tag{4.24}$$

We now must pick a suitable resolution ϵ to guarantee that this probability is less than $\exp(-c_1 p)$ for a suitably chosen constant c_1. First note that if we choose $\epsilon < (\delta/4)(\sqrt{mn/p}+1)^{-1}$,

$$\mathbf{P}\left(\|\mathcal{A}\| \geq \frac{\delta}{2\epsilon} - 1\right) \leq \exp(-\gamma mn) , \tag{4.25}$$

which achieves the desired scaling because $mn > p$. With this choice of ϵ, the quantity in Equation (4.22) is less than or equal to

$$\begin{aligned} &2\left(\frac{8C_0(\sqrt{mn/p}+1)}{\delta}\right)^{r(m+n-2r)} (24/\delta)^{r^2} \exp\left(-\frac{p}{2}(\delta^2/32 - \delta^3/192)\right) \\ &= \exp\left(-pa(\delta) + r(m+n-2r)\log\left(\sqrt{\frac{mn}{p}}+1\right)\right. \\ &\qquad\qquad \left. + r(m+n-2r)\log\left(\frac{8C_0}{\delta}\right) + r^2 \log\left(\frac{24}{\delta}\right)\right) \end{aligned} \tag{4.26}$$

where $a(\delta) = \delta^2/64 - \delta^3/384$. Since $mn/p < mn$ for all $p > 1$, there exists a constant c_0 independent of m,n,p, and r, such that the sum of the last three terms in the exponent are bounded above by $(c_0/a(\delta))r(m+n)\log(mn)$. It follows that there exists a constant c_1 independent of m,n,p, and r such that $p \geq c_0 r(m+n)\log(mn)$ observations are sufficient to yield an RIP of δ with probability greater than $1 - e^{-c_1 p}$. ■

Intuitively, the scaling $p = O\left(r(m+n)\log(mn)\right)$ is very reasonable, since a rank r matrix has $r(m+n-r)$ degrees of freedom. This coarse tail bound only provides asymptotic estimates for recovery, and is quite conservative in practice. As we demonstrate in Section 6, minimum rank solutions can be determined from between $2r(m+n-r)$ to $4r(m+n-r)$ observations for many practical problems.

5 Algorithms for nuclear norm minimization

A variety of methods can be developed for the effective minimization of the nuclear norm over an affine subspace of matrices, and we do not have space for a fully comprehensive treatment here. Instead, in this section we focus on three specific methods, that illustrate and highlight the general trade-offs between theoretical guarantees, computational performance, and numerical accuracy of the resulting solution.

Directly solving the semidefinite characterization of the nuclear norm problem using primal-dual interior point methods is a numerically efficient method for small problems and can be used to yield accuracy up to floating-point precision. However, since interior point methods use second order information, the memory requirements for computing descent directions quickly becomes prohibitively large for most practical applications. Moreover, it is preferable to use methods that exploit, at least partially, the structure of the problem. This can be done at several levels, either by taking into account further information that may be available about the linear map $\mathcal{A}$ (e.g., the matrix completion problem discussed above) or by formulating algorithms that are specific to the nuclear norm problem. For the latter, we show how to apply subgradient methods to minimize the nuclear norm over an affine set. Such first-order methods cannot yield as high numerical precision as interior point methods, but much larger problems can be solved because no second-order information needs to be stored. For even larger problems, we discuss a semidefinite programming heuristic that explicitly works with a low-rank factorization of the decision variable. This method can be applied even when the matrix decision variable cannot fit into memory, but convergence guarantees are much less satisfactory than in the other two cases.

5.1 Interior point methods for semidefinite programming

For relatively small problems where a high-degree of numerical precision is required, interior point methods for semidefinite programming can be directly applied to solve the affine nuclear norm minimization problem. As we have seen in earlier sections, this optimization problem can be directly posed as a semidefinite program, via the standard form primal-dual pair (2.8). As written, the primal problem has one $(n+m) \times (n+m)$ semidefinite constraint and p affine constraints. Conversely, the dual problem has one $(n+m) \times (n+m)$ semidefinite constraint and p scalar decision variables. Thus, the total number of decision variables (primal and dual) is equal to $\binom{n+m+1}{2} + p$.

Modern interior point solvers for semidefinite programming generally use primal-dual methods, and compute an update direction for the current solution by solving a suitable Newton system. Depending on the structure of the linear mapping $\mathcal{A}$, this may entail constructing and solving a potentially large, dense linear system.

If the matrix dimensions n and m are not too large, then any good interior point SDP solver, such as SeDuMi [71] or SDPT3 [77], will quickly produce accurate solutions. In fact, as we will see in the next section, problems with n and m around 50 can be solved to machine precision in a few minutes on a desktop computer. However, solving such a primal-dual pair of programs with traditional interior point methods can prove to be quite challenging when the dimensions of the matrix X are much bigger than 100×100 and the number of equations is in the thousands. In this case, the corresponding Newton systems can become quite large, and, without any specific additional structure, the memory requirements of such dense systems quickly limit the size of problems that can be solved. This can be controlled to some extent by exploiting the problem structure when assembling and solving the Newton system, as in the recent work of Liu and Vandenberghe [48].

Perhaps the most important drawback of the direct SDP approach is that it completely ignores the possibility of efficiently computing the nuclear norm via a singular value decomposition, instead of the less efficient eigenvalue decomposition of a bigger matrix. The methods we discuss next will circumvent this obstacle, by directly working with subgradients of the nuclear norm.

5.2 Projected subgradient methods

The nuclear norm minimization (3.1) is a linearly constrained nondifferentiable convex problem. There are numerous techniques to approach this kind of problem depending on the specific nature of the constraints (e.g., dense vs. sparse) and the possibility of using first- or second-order information. In this section we describe a family of simple, easy to implement, subgradient projection methods to solve (3.1).

The standard Euclidean projected subgradient method computes a sequence of feasible points $\{X_k\}$, with iterates satisfying the update rule

$$X_{k+1} = \Pi(X_k - s_k Y_k), \qquad Y_k \in \partial ||X_k||_*,$$

where Π is the orthogonal projection onto the affine subspace defined by the linear constraints $\mathcal{A}(X) = b$, and $s_k > 0$ is a stepsize parameter. In other words, the method updates the current iterate X_k by taking a step along the direction of a subgradient at the current point and then projecting back onto the feasible set. Alternatively, since X_k is feasible, we can rewrite this as

$$X_{k+1} = X_k - s_k \Pi_{\mathcal{A}} Y_k,$$

where $\Pi_{\mathcal{A}}$ is the orthogonal projection onto the kernel of $\mathcal{A}$. Since the feasible set is an affine subspace, there are several options for the projection $\Pi_{\mathcal{A}}$. For small problems, one can precompute it using, for example, a QR decomposition of the matrix representation of $\mathcal{A}$ and store it. Alternatively, one can solve a least squares problem at each step by iterative methods such as conjugate gradients.

The subgradient-based method described above is extremely simple to implement, since only a subgradient evaluation is required at every step. The computation of the subgradient can be done using the formula given in (2.9) earlier, thus requiring only a singular value decomposition of the current point X_k.

A possible alternative to the computation of the SVD for the subgradient computation is to directly focus on the "angular" factor of the polar decomposition of X_k, using for instance the Newton-like methods developed by Gander in [41]. Specifically, for a given matrix X_k, the Halley-like iteration

$$X \to X(X'X + 3I)(3X'X + I)^{-1}$$

converges globally and quadratically to the polar factor of X, and thus yields an element of the subdifferential of the nuclear norm. This iterative method (suitably scaled) can be faster than a direct SVD computation, particularly if the singular values of the initial matrix are close to 1. This could be appealing since presumably only a very small number of iterations would be needed to update the polar factor of X_k, although the nonsmoothness of the subdifferential is bound to cause some additional difficulties.

Regarding convergence, for general nonsmooth problems, subgradient methods do not guarantee a decrease of the cost function at every iteration, even for arbitrarily small step sizes (see, e.g., [8, §6.3.1]), unless the minimum-norm subgradient is used. Instead, convergence is usually shown through the decrease (for small stepsize) of the distance from the iterates X_k to any optimal point. There are several possibilities for the choice of stepsize s_k. The simplest choice that can guarantee convergence is to use a diminishing stepsize with an infinite travel condition (i.e., such that $\lim_{k\to\infty} s_k = 0$ and $\sum_{k>0} s_k$ diverging).

More recently, several nonlinear projected subgradient methods, under the rubric of *mirror descent*, have been developed (e.g., by Nemirovski and Yudin [57]), followed by a subsequent rederivation and analysis by Beck and Teboulle [5]. These algorithms, and their accompanying analysis, provide improved theoretical guarantees and practical performance over the standard Euclidean projected subgradient method described above. It would be of great interest to analyze to what extent these methods can be applied to the nuclear norm minimization problem.

In many scenarios, even the computation of a singular value decomposition or Halley-like iteration can be too computationally expensive. The next section proposes a reduction of the size of the search space to alleviate such demands. We must give up guarantees of convergence for this convenience, but this may be an acceptable trade-off for very large-scale problems.

5.3 Low-rank parametrization

We now turn to a method that works with an explicit low-rank factorization of X. This algorithm not only requires less storage capacity and computational overhead than the previous methods, but for many problems does not even require one to be able to store the decision variable X in memory. This is the case, for example, in the matrix completion problem where $\mathcal{A}(X)$ is a subset of the entries of X.

Given observations of the form $\mathcal{A}(X) = b$ of an $m \times n$ matrix X of rank r, a possible algorithm to find a suitable X would be to find a factorization $X = LR'$, where L is an $m \times r$ matrix and R an $n \times r$ matrix, such that the equality constraints are satisfied. Since there are many possible such factorizations, we could search for one where the matrices L and R have Frobenius norm as small as possible, i.e., the solution of the optimization problem

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}(\|L\|_F^2 + \|R\|_F^2) \\ \text{subject to} & \mathcal{A}(LR') = b. \end{array} \tag{5.1}$$

Even though the cost function is convex, the constraint is not. Such a problem is a nonconvex quadratic program, and it is not evidently easy to optimize. Surprisingly, minimization of the nuclear norm subject to equality constraints is in fact equivalent to this rather natural heuristic optimization, as long as r is chosen to be sufficiently larger than the rank of the optimum of the nuclear norm problem.

Lemma 5.1 *Assume $r \geq \operatorname{rank}(X_0)$. The nonconvex quadratic optimization problem (5.1) is equivalent to the minimum nuclear norm relaxation (3.1).*

Proof Consider any feasible solution (L, R) of (5.1). Then, defining $W_1 := LL'$, $W_2 := RR'$, and $X := LR'$ yields a feasible solution of the primal SDP problem (2.8) that achieves the same cost. Since the SDP formulation is equivalent to the nuclear norm problem, we have that the optimal value of (5.1) is always greater than or equal to the nuclear norm heuristic.

For the converse, we can use an argument similar to the proof of Proposition 2.1. From the SVD decomposition $X^* = U\Sigma V'$ of the optimal solution of the nuclear norm relaxation (3.1), we can explicitly construct matrices $L := U\Sigma^{\frac{1}{2}}$ and $R := V\Sigma^{\frac{1}{2}}$ for (5.1) that yield exactly the same value of the objective. ■

The main advantage of this reformulation is to substantially decrease the number of primal decision variables from nm to $(n + m)r$. For large problems, this is quite a significant reduction that allows us to search for matrices of rank smaller than the order of 100, and $n + m$ in the hundreds of thousands on a desktop computer. However, the formulation (5.1) is nonconvex and thus potentially subject to local minima that are not globally optimal. This non-convexity does not pose as much of a problem as it could, for two reasons. First recall from Theorem 3.2 that if $\delta_{2r}(\mathcal{A}) < 1$, there is a unique X^* with rank at most r such that $\mathcal{A}(X^*) = b$. Since any local minimum (L^*, R^*) of (5.1) is feasible, we would have $X^* = L^* R^{*\prime}$ and we would have found the minimum rank solution. Second, we now present an algorithm that is guaranteed to converge to a local minimum for a judiciously selected r. We will also provide a sufficient condition for when we can construct an optimal solution of (2.8) from the solution computed by the method of multipliers.

Low-rank factorizations and the method of multipliers For general semidefinite programming problems, Burer and Monteiro have developed in [10, 11] a nonlinear programming approach that relies on a low-rank factorization of the matrix decision variable. We will adapt this idea to our problem, to provide a first-order Lagrangian minimization algorithm that efficiently finds a local minima of (5.1). As a consequence of the work in [11], it will follow that for values of r larger than the rank of the true optimal solution, the local minima of (5.1) can be transformed into global minima of (2.8) under the identification $W_1 = LL'$, $W_2 = RR'$ and $Y = LR'$ (cf. Lemma 5.1). We summarize below the details of this approach.

The algorithm employed is called the *method of multipliers*, a standard approach for solving equality constrained optimization problems [7]. The method of multipliers works with an augmented Lagrangian for (5.1)

$$\mathcal{L}_a(L, R; y, \sigma) := \tfrac{1}{2}(||L||_F^2 + ||R||_F^2) - y'(\mathcal{A}(LR') - b) + \tfrac{\sigma}{2}||\mathcal{A}(LR') - b||^2, \tag{5.2}$$

where the y_i are arbitrarily signed Lagrange multipliers and σ is a positive constant. A somewhat similar algorithm was proposed by Rennie *et al.* in [63] for applications in collaborative filtering. In this work, the authors minimize $\mathcal{L}_a$ with σ fixed and $y = 0$ to serve as a regularized algorithm for matrix completion. Remarkably, by deterministically varying σ and y, this method can be adapted into an algorithm for solving linearly constrained nuclear-norm minimization.

In the method of multipliers, one alternately minimizes the augmented Lagrangian with respect to the decision variables L and R, and then increases the value of the penalty coefficient σ and updates y. The augmented Lagrangian can be minimized using any local search technique, and the partial derivatives are particularly simple to compute. Let $\hat{y} := y - \sigma(\mathcal{A}(LR') - b)$. Then we have

$$\begin{aligned} \nabla_L \mathcal{L}_a &= L - \mathcal{A}^*(\hat{y})R \\ \nabla_R \mathcal{L}_a &= R - \mathcal{A}^*(\hat{y})'L. \end{aligned}$$

To calculate the gradients, we first compute the constraint violations $\mathcal{A}(LR') - b$, then form $\hat{y}$, and finally use the above equations to compute the gradients.

As the number of iterations tends to infinity, only feasible points will have finite values of $\mathcal{L}_a$, and for any feasible point, $\mathcal{L}_a(L, R)$ is equal to the original cost function $(\|L\|_F^2 + \|R\|_F^2)/2$. The method terminates when L and R are feasible, as in this case the Lagrangian is stationary and we are at a local minima of (5.1). Including the y multipliers improves the conditioning of each subproblem where $\mathcal{L}_a$ is minimized and enhances the rate of convergence. The following theorem shows that when the method of multipliers converges, it converges to a local minimum of (5.1).

Theorem 5.2 *Suppose we have a sequence $(L^{(k)}, R^{(k)}, y^{(k)})$ of local minima of the augmented Lagrangian at each step of the method of multipliers. Assume that our sequence of $\sigma^{(k)} \to \infty$ and that the sequence of $y^{(k)}$ is bounded. If $(L^{(k)}, R^{(k)})$ converges to (L^*, R^*) and the linear map*

$$\Lambda^{(k)}(y) := \begin{bmatrix} \mathcal{A}^*(y)R^{(k)} \\ \mathcal{A}^*(y)'L^{(k)} \end{bmatrix} \tag{5.3}$$

has kernel equal to the zero vector for all k, then there exists a vector y^* such that

(i) $\nabla \mathcal{L}_a(L^*, R^*; y^*) = 0$

(ii) $\mathcal{A}(L^* R^{*\prime}) = b$

Proof This proof is standard and follows the approach in [7]. As above, we define $\hat{y}^{(k)} := y^{(k)} - \sigma^{(k)}(\mathcal{A}(L^{(k)} R^{(k)\prime}) - b)$ for all k. Since $(L^{(k)}, R^{(k)})$ minimize the augmented Lagrangian at iteration k, we have

$$\begin{aligned} 0 &= L^{(k)} - \mathcal{A}^*(\hat{y}^{(k)}) R^{(k)} \\ 0 &= R^{(k)} - \mathcal{A}^*(\hat{y}^{(k)})' L^{(k)}, \end{aligned} \tag{5.4}$$

which we may rewrite as

$$\Lambda^{(k)}(\hat{y}^{(k)}) = \left[\begin{array}{c} L^{(k)} \\ R^{(k)} \end{array} \right] . \tag{5.5}$$

Since we have assumed that there is no non-zero y with $\Lambda^{(k)}(y) = 0$, there exists a left-inverse and we can solve for $\hat{y}^{(k)}$.

$$\hat{y}^{(k)} = \Lambda^{(k)\dagger} \left(\left[\begin{array}{c} L^{(k)} \\ R^{(k)} \end{array} \right] \right) . \tag{5.6}$$

Everything on the right-hand side is bounded, and $L^{(k)}$ and $R^{(k)}$ converge. Therefore, we must have that $\hat{y}^{(k)}$ converges to some $\hat{y}^*$. Taking the limit of (5.4) proves (i). To prove (ii), note that we must have $\hat{y}^{(k)}$ is bounded. Since $y^{(k)}$ is also bounded, we find that $\sigma^{(k)}(\mathcal{A}(L^{(k)} R^{(k)\prime}) - b)$ is also bounded. But $\sigma^{(k)} \to \infty$ implies that $\mathcal{A}(L^* R^{*\prime}) = b$, completing the proof. ■

Suppose the decision variables are chosen to be of size $m \times r_d$ and $n \times r_d$. A necessary condition for $\Lambda^k(y)$ to be full rank is for the number of decision variables $r_d(m+n)$ to be greater than the number of equalities p. In particular, this means that we must choose $r_d \geq p/(m+n)$ in order to have any hopes of satisfying the conditions of Theorem 5.2.

We close this section by relating the solution found by the method of multipliers to the optimal solution of the nuclear norm minimization problem. We have already shown that when the low-rank algorithm converges, it converges to a low-rank solution of $\mathcal{A}(X) = b$. If we additionally find that $\mathcal{A}^*(y^*)$ has norm less than or equal to one, then it is dual feasible. One can check using straightforward algebra that $(L^* R^{*\prime}, L^* L^{*\prime}, R^* R^{*\prime})$ and y^* form an optimal primal-dual pair for (2.8). This analysis proves the following theorem.

Theorem 5.3 *Let (L^*, R^*, y^*) satisfy (i)-(ii) in Theorem 5.2 and suppose $\|\mathcal{A}^*(y^*)\| \leq 1$. Then $(L^* R^{*\prime}, L^* L^{*\prime}, R^* R^{*\prime})$ is an optimal primal solution and y^* is an optimal dual solution of (2.8).*

6 Numerical experiments

To illustrate the scaling of low-rank recovery for a particular matrix M, consider the MIT logo presented in Figure 1. The image has a total of 46 rows and 81 columns (total 3726 elements), with three distinct non-zero numerical values corresponding to the colors white, red, and grey. Since the logo only has 5 distinct rows, it has rank 5. For each of the ensembles discussed in Section 4,

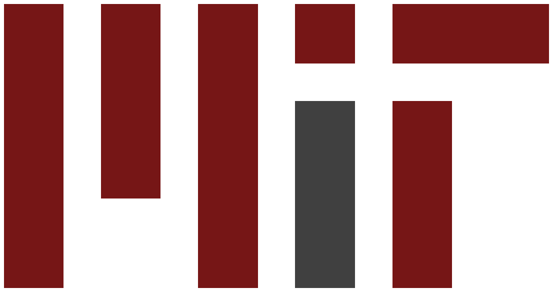


Figure 1: The MIT logo image. The associated matrix has dimensions 46 × 81 and has rank 5.

we sampled constraint matrices with p ranging between 700 and 1500, and solved the semidefinite program (2.6) using the freely available software SeDuMi [71]. On a 2.0 GHz laptop computer, each semidefinite program could be solved in less than four minutes. We chose to use this interior point method because it yielded the highest accuracy in the shortest amount of time, and we were interested in characterizing precisely when the nuclear norm heuristic succeeded and failed.

Figure 2 plots the Frobenius norm of the difference between the optimal solution of the semidefinite program and the true image. We observe a sharp transition to perfect recovery near 1200 constraints, which is approximately equal to $2r(m+n-r)$. In Figure 3, we display the recovered solutions for various values of p, under the Gaussian ensemble.

To demonstrate the average behavior of low-rank recovery, we conducted a series of experiments for a variety of the matrix sizes n, ranks r, and numbers of constraints p. For a fixed n, we constructed random recovery scenarios for low-rank $n \times n$ matrices. For each n, we varied p between 0 and n^2 where the matrix is completely determined. For a fixed n and p, we generated all possible ranks such that $r(2n-r) \leq p$. This cutoff was chosen because beyond that point there would be an infinite set of matrices of rank r satisfying the p equations.

For each (n, p, r) 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 p rows and n^2 columns. Then the nuclear norm minimization problem

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \mathbf{A}\,\text{vec}(X) = \mathbf{A}\,\text{vec}(Y_0) \end{array} \tag{6.1}$$

was solved using the SDP solver SeDuMi on the formulation (2.6). Again, we chose to use SeDuMi because we wanted to precisely distinguish between success and failure of the heuristic. We declared Y_0 to be recovered if $\|X - Y_0\|_F / \|Y_0\|_F < 10^{-3}$. Figure 4 shows 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.

These experiments demonstrate that the logarithmic factors and constants present in our scaling results are somewhat conservative. For example, as one might expect, low-rank matrices are per-

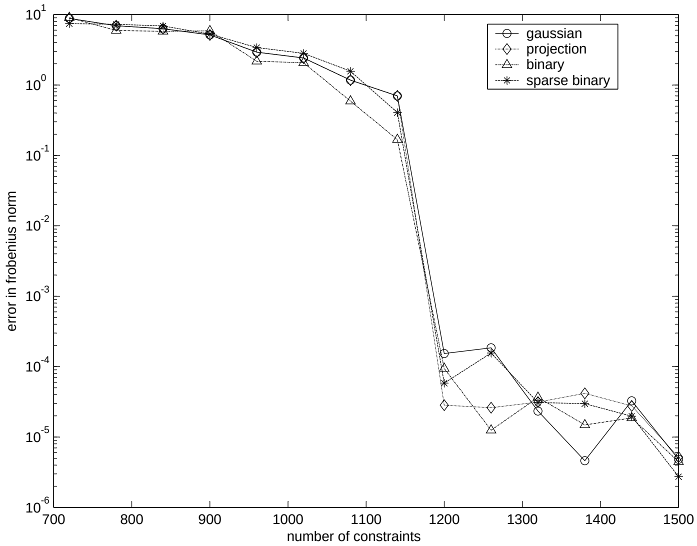
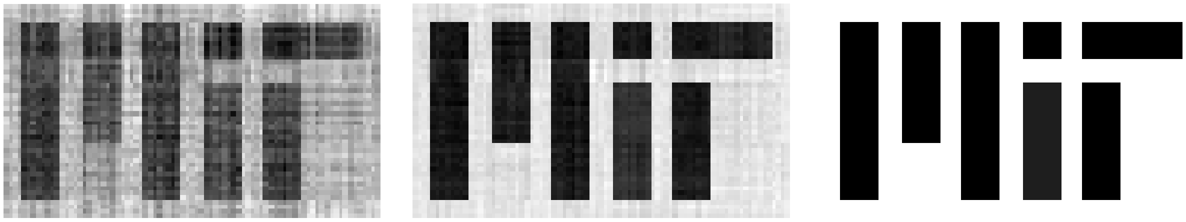
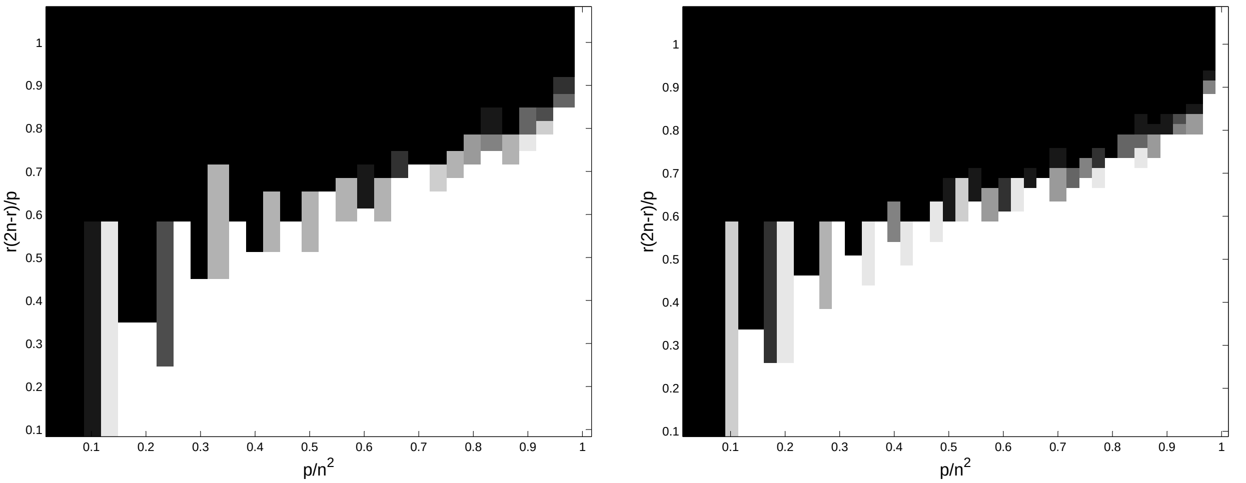


Figure 2: (a) Error, as measured by the Frobenius norm, between the recovered image and the ground truth. Observe that there is a sharp transition to near zero error at around 1200 constraints.



(a) (b) (c)

Figure 3: Example recovered images using the Gaussian ensemble. (a) 700 constraints. (b) 1100 constraints. (c) 1250 constraints. The total number of pixels is 46 × 81 = 3726. Note that the error is plotted on a logarithmic scale.



(a)

(b)

Figure 4: For each (n, p, r) triple, we repeated the following procedure ten 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 set $Y_0 = Y_L Y_R'$. We select a matrix $\mathbf{A}$ from the Gaussian ensemble with p rows and n^2 columns. Then we solve the nuclear norm minimization subject to $\mathbf{A}\,\mathrm{vec}(X) = \mathbf{A}\,\mathrm{vec}(Y_0)$ We declare Y_0 to be recovered if $\|X - Y_0\|_F / \|Y_0\|_F < 10^{-3}$. The results are shown for (a) $n = 30$ and (b) $n = 40$. The color of each cell reflects the empirical recovery rate (scaled between 0 and 1). White denotes perfect recovery in all experiments, and black denotes failure for all experiments.

fectly recovered by nuclear norm minimization when $p = n^2$ as the matrix is uniquely determined. Moreover, as p is reduced slightly away from this value, low-rank matrices are still recovered 100 percent of the time for most values of r. Finally, we note that despite the asymptotic nature of our analysis, our experiments demonstrate excellent performance with low-rank matrices of size 30×30 and 40×40 matrices, showing that the heuristic is practical even in low-dimensional settings.

Intriguingly, Figure 4 also demonstrates a "phase transition" between perfect recovery and failure. As observed in several recent papers by Donoho and collaborators (see e.g. [26, 27]), the random sparsity recovery problem has two distinct connected regions of parameter space: one where the sparsity pattern is perfectly recovered, and one where no sparse solution is found. Not surprisingly, Figure 4 illustrates an analogous phenomenon in rank recovery. Computing explicit formulas for the transition between perfect recovery and failure is left for future work.

7 Discussion and future developments

Having illustrated the natural connections between affine rank minimization and affine cardinality minimization, we were able to draw on these parallels to determine scenarios where the nuclear norm heuristic exactly solves the rank minimization problem. These scenarios directly generalized conditions for which the ℓ_1 heuristic succeeded and ensembles of linear maps for which these conditions hold. Furthermore, our experimental results display similar recovery properties to those demonstrated in the empirical studies of ℓ_1 minimization. Inspired by the success of this program, we close this paper by briefly discussing several exciting directions that are natural continuations of this work building on more analogies from the compressed sensing literature. We also describe possible extensions to more general notions of parsimony.

Incoherent ensembles and partially observed transforms Taking our lead from the compressed sensing literature, it would be of great interest to extend the results of [13] to low-rank recovery. In this work, the authors show that partially observed unitary transformations of sparse vectors can be used to recover the sparse vector using ℓ_1 minimization. There are many practical applications where low-rank processes are partially observed or measured. For instance, the matrix completion problem can be thought of as partial observations under the identity transformation. As another example, there are many cases in two-dimensional Fourier spectroscopy where only partial information can be observed or measured due to experimental constraints.

Noisy measurements and low-rank approximation Our results in this paper address only the case of exact (noiseless) measurements or observations. It is of natural interest to understand the behavior of the nuclear norm heuristic in the case of noisy data. Based on the existing results for the sparse case (e.g., [14]), it would be natural to expect similar stability properties of the recovered solution, for instance in terms of the ℓ_2 norm of the computed solution. Such an analysis could also be used to study the nuclear norm heuristic as an approximation technique where a matrix has rapidly decaying singular values and a low-rank approximation is desired.

Factored measurements and alternative ensembles All of the measurement ensembles considered require the storage of $O(mnp)$ numbers. For large problems this is wholly impractical. There are many promising alternative measurement ensembles that seem to obey the same scaling laws as those presented in Section 4. For example, "factored" measurements, of the form $A_i : X \mapsto u_i^T X v_i$, where u_i, v_i are Gaussian random vectors empirically yield the same performance as the Gaussian ensemble. This factored ensemble only requires storage of $O((m+n)p)$ numbers, which is a rather significant savings for very large problems. The proof in Section 4 does not seem to extend to this ensemble, thus new machinery must be developed to guarantee properties about such low-rank measurements.

Alternative numerical methods Besides the techniques described in Section 5, there are a number of interesting additional possibilities to solve the nuclear norm minimization problem. An appealing suggestion is to combine the strength of second-order methods (as in the standard interior point approach) with the known geometry of the nuclear norm (as in the subgradient approach), and develop a customized interior point method, possibly yielding faster convergence rates, while still being relatively memory-efficient.

It is also of much interest to investigate the possible adaptation of some of the successful path-following approaches in traditional ℓ_1/cardinality minimization, such as the Homotopy [59] or LARS (least angle regression) [29]. This may be not be completely straightforward, since the efficiency of many of these methods often relies explicitly on the polyhedral structure of the feasible set of the ℓ_1 norm problem.

Searching for lower rank solutions via iterative methods Often times, in the absence of an RIP condition, the nuclear norm heuristic does not return a sufficiently low rank solution. A variety of algorithms have been proposed to attempt to further reduce the rank beyond the value returned by nuclear norm minimization. Fazel *et al.* proposed a heuristic that employs the logarithm of the determinant as a smooth approximation for rank and locally optimizes this function to obtain a sequence of semi-definite programming problems [37, 39]. The initial iteration is equivalent to (2.8), and it often provides sparser solutions than the nuclear norm heuristic in practice. This algorithm has been adapted to cardinality minimization, resulting in iterative weighted ℓ_1 norm minimization, and has been successful in this setting as well [49, 17]. Another heuristic involves ℓ_p norm minimization (locally) with $p < 1$ [18]. These algorithms do not currently have any theoretical guarantees, and it bears investigation if the probabilistic analysis developed in this paper can be applied to determine when these iterative algorithms can find low-rank solutions beyond the guarantees derived for nuclear norm minimization.

Geometric interpretations For the case of cardinality/ℓ_1 minimization, a beautiful geometric interpretation has been set forth by Donoho and Tanner [26, 27]. Key to their results is the notion of *central k-neighborliness* of a centrosymmetric polytope, namely the property that every subset of $k+1$ vertices not including an antipodal pair spans a k-face. In particular, they show that the ℓ_1 heuristic always succeeds whenever the image of the ℓ_1 unit ball (the cross-polytope) under the linear mapping $\mathcal{A}$ is a centrally k-neighborly polytope.

In the case of rank minimization, the direct application of these concepts fails, since the unit ball of the nuclear norm is not a polyhedral set. Nevertheless, it seems likely that a similar explanation could be developed, where the key feature would be the preservation under a linear map of the extremality of the components of the boundary of the nuclear norm unit ball defined by low-rank conditions.

Jordan algebras As we have seen, our results for the rank minimization problem closely parallel the earlier developments in cardinality minimization. A convenient mathematical framework that allows the simultaneous consideration of these cases as well as a few new ones, is that of *Jordan algebras* and the related symmetric cones [33]. In the Jordan-algebraic setting, there is an intrinsic notion of rank that agrees with the cardinality of the support in the case of the nonnegative orthant or the rank of a matrix in the case of the positive semidefinite cone. Besides mathematical elegance, a direct Jordan-algebraic approach would transparently yield similar results for the case of second-order (or Lorentz) cone constraints.

As specific examples of the power and elegance of this approach, we mention the work of Faybusovich [35] and Schmieta and Alizadeh [66] that provide a unified development of interior point methods for symmetric cones, as well as Faybusovich's work on convexity theorems for quadratic mappings [36].

Parsimonious models and optimization Sparsity and low-rank are two specific classes of parsimonious (or low-complexity) descriptions. Are there other kinds of easy-to-describe parametric models that are amenable to exact solutions via convex optimizations techniques? Given the intimate connections between linear and semidefinite programming and the Jordan algebraic approaches described earlier, it is likely that this will require alternative tractable convex optimization formulations.

8 Acknowledgements

We thank Stephen Boyd, Emmanuel Candès, José Costa, John Doyle, Ali Jadbabaie, Ali Rahimi, and Michael Wakin for their useful comments and suggestions. We also thank the IMA in Minneapolis for hosting us during the initial stages of our collaboration.

References

[1] P.-A. Absil, A. Edelman, and P. Koev. On the largest principal angle between random subspaces. *Linear Algebra Appl.*, 414(1):288–294, 2006.

[2] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and Systems Science*, 66(4):671–687, 2003. Special issue of invited papers from PODS'01.

[3] H. C. Andrews and C. L. Patterson, III. Singular value decomposition (SVD) image coding. *IEEE Transactions on Communications*, 24(4):425–432, 1976.

[4] 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`.

[5] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[6] C. Beck and R. D'Andrea. Computational study and comparisons of LFT reducibility methods. In *Proceedings of the American Control Conference*, 1998.

[7] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Athena Scientific, Belmont, Massachusetts, 1996.

[8] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 2nd edition, 1999.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2003.

[10] S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming (Series B)*, 95:329–357, 2003.

[11] S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[12] E. J. Candès. Compressive sampling. In *International Congress of Mathematicians. Vol. III*, pages 1433–1452. Eur. Math. Soc., Zürich, 2006.

[13] E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

[14] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications of Pure and Applied Mathematics*, 59:1207–1223, 2005.

[15] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[16] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[17] E. J. Candés, M. P. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. Technical report, 2008. Preprint available at `http://www.eecs.umich.edu/~wakin/publications.html`.

[18] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *Signal Process. Lett.*, 14(10):707–710, 2007.

[19] S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to nonlinear system identification. *International Journal of Control*, 50(5):1873–1896, 1989.

[20] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computing*, 20(1):33–61, 1998.

[21] J. F. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38(5):826–844, 1973.

[22] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.

[23] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook on the Geometry of Banach spaces*, pages 317–366. Elsevier Scientific, 2001.

[24] G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *SPIE Journal of Optical Engineering*, 33(7):2183–2191, 1994.

[25] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[26] D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA*, 102(27):9452–9457, 2005.

[27] D. L. Donoho and J. Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA*, 102(27):9446–9451, 2005.

[28] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[29] B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[30] 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.

[31] L. El Ghaoui, F. Oustry, and M. A. Rami. A cone complementarity linearization algorithm for static output-feedback and related problems. *IEEE Transactions on Automatic Control*, 42(8):1171–1176, 1997.

[32] N. El Karoui. *New results about random covariance matrices and statistical applications.* PhD thesis, Stanford University, 2004.

[33] J. Faraut and A. Korányi. *Analysis on symmetric cones.* Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, 1994.

[34] B. Fares, P. Apkarian, and D. Noll. An augmented Lagrangian method for a class of LMI-constrained problems in robust control theory. *International Journal of Control*, 74(4):384–360, 2001.

[35] L. Faybusovich. Euclidean Jordan algebras and interior-point algorithms. *Positivity*, 1(4):331–357, 1997.

[36] L. Faybusovich. Jordan-algebraic approach to convexity theorems for quadratic mappings. *SIAM Journal on Optimization*, 17(2):558–576, 2006.

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

[38] 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.

[39] M. Fazel, H. Hindi, and S. Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *Proceedings of the American Control Conference*, 2003.

[40] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[41] W. Gander. Algorithms for the polar decomposition. *SIAM J. Sci. Statist. Comput.*, 11(6):1102–1115, 1990.

[42] K. M. Grigoriadis and E. B. Beran. Alternating projection algorithms for linear matrix inequalities problems with rank constraints. In L. El Ghaoui and S. Niculescu, editors, *Advances in Linear Matrix Inequality Methods in Control*, chapter 13, pages 251–267. SIAM, 2000.

[43] A. Hassibi, J. How, and S. Boyd. Low-authority controller design via convex optimization. *AIAA Journal of Guidance, Control, and Dynamics*, 22(6):862–872, 1999.

[44] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods.* Springer-Verlag, New York, 1993.

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

[46] A. S. Lewis. The mathematics of eigenvalue optimization. *Mathematical Programming*, 97(1–2):155–176, 2003.

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

[48] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. `www.ee.ucla.edu/~vandenbe/nucnrm.html`, 2008.

[49] M. Lobo, M. Fazel, and S. P. Boyd. Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1):376–394, 2007.

[50] G. G. Lorentz, M. von Golitschek, and Y. Makovoz. *Constructive Approximation: Advanced problems*, volume 304 of *Grundlehren der Mathematischen Wissenschaften.* Springer, 1996.

[51] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[52] I. Markovsky. Structured low-rank approximation and its applications. *Automatica*, 44:891–909, 2007.

[53] G. Marsaglia and G. P. H. Styan. When does $\text{rank}(A+B) = \text{rank}(A) + \text{rank}(B)$? *Canad. Math. Bull.*, 15:451–452, 1972.

[54] M. Mesbahi and G. P. Papavassilopoulos. On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Transactions on Automatic Control*, 42(2):239–243, 1997.

[55] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quart. J. Math. Oxford Ser. (2)*, 11:50–59, 1960.

[56] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal of Computing*, 24(2):227–234, 1995.

[57] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization.* Wiley, New York, 1983.

[58] R. Orsi, U. Helmke, and J. B. Moore. A Newton-like method for solving rank constrained linear matrix inequalities. *Automatica*, 42(11):1875–1882, 2006.

[59] M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–403, 2000.

[60] P. A. Parrilo and S. Khatri. On cone-invariant linear matrix inequalities. *IEEE Trans. Automat. Control*, 45(8):1558–1563, 2000.

[61] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems and Computers*, 1993.

[62] S. Qian and D. Chen. Signal representation using adaptive normalized gaussian functions. *Signal Processing*, 36:329–355, 1994.

[63] 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.

[64] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[65] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[66] S. H. Schmieta and F. Alizadeh. Associative and Jordan algebras, and polynomial time interior-point algorithms for symmetric cones. *Math. Oper. Res.*, 26(3):543–564, 2001.

[67] I. J. Schoenberg. Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". *Annals of Mathematics*, 36(3):724–732, July 1935.

[68] R. E. Skelton, T. Iwasaki, and K. Grigoriadis. *A Unified Algebraic Approach to Linear Control Design.* Taylor and Francis, 1998.

[69] E. Sontag. *Mathematical Control Theory.* Springer-Verlag, New York, 1998.

[70] N. Srebro. *Learning with Matrix Factorizations.* PhD thesis, Massachusetts Institute of Technology, 2004.

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

[72] S. J. Szarek. The finite dimensional basis problem with an appendix on nets of the Grassmann manifold. *Acta Mathematica*, 151:153–179, 1983.

[73] S. J. Szarek. Metric entropy of homogeneous spaces. In *Quantum probability (Gdańsk, 1997)*, volume 43 of *Banach Center Publ.*, pages 395–410. Polish Acad. Sci., Warsaw, 1998. Preprint available at `arXiv:math/9701213v1`.

[74] H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39–52, 1979.

[75] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[76] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, 58(1):267–288, 1996.

[77] K. C. Toh, M. Todd, and R. H. Tütüncü. *SDPT3 - a MATLAB software package for semidefinite-quadratic-linear programming.* Available from `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`.

[78] M. W. Trosset. Distance matrix completion by numerical optimization. *Computational Optimization and Applications*, 17(1):11–22, October 2000.

[79] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

[80] L. Vandenberghe, S. Boyd, and A. El-Gamal. Optimal wire and transistor sizing for circuits with non-tree topology. In *Proc. of IEEE/ACM International Conference on Computer Aided Design*, pages 252–259, 1997.

[81] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk. An architecture for compressive imaging. In *Proc. International Conference on Image Processing – ICIP 2006*, oct 2006.

[82] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Applications*, 170:1039–1053, 1992.

[83] H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors. *Handbook of Semidefinite Programming.* Kluwer Academic Publishers, Boston, 2000.

[84] Y. Q. Yin, Z. D. Bai, and P. R. Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78:509–512, 1988.