Forty-Sixth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 23-26, 2008

Exact Low-rank Matrix Completion via Convex Optimization

Emmanuel J. Candès† and Benjamin Recht♯

† Applied and Computational Mathematics, Caltech, Pasadena, CA 91125

♯ Center for the Mathematics of Information, Caltech, Pasadena, CA 91125

Abstract— **Suppose that one observes an incomplete subset of entries selected uniformly at random from a low-rank matrix. When is it possible to complete the matrix and recover the entries that have not been seen? We show that in very general settings, one can perfectly recover all of the missing entries from a sufficiently large random subset by solving a convex programming problem. This program finds the matrix with the minimum nuclear norm agreeing with the observed entries. The techniques used in this analysis draw upon parallels in the field of compressed sensing, demonstrating that objects other than signals and images can be perfectly reconstructed from very limited information.**

I. Introduction

In many practical problems of interest, one would like to recover a matrix from a sampling of its entries. As a motivating example, consider the task of inferring answers in a partially filled out survey. That is, suppose that questions are being asked to a collection of individuals. Then we can form a matrix where the rows index each individual and the columns index the questions. We collect data to fill out this table but unfortunately, many questions are left unanswered. Is it possible to make an educated guess about what the missing answers should be? How can one make such a guess? Formally, we may view this problem as follows. We are interested in recovering a data matrix $\mathbf{M}$ with n_1 rows and n_2 columns but only get to observe a number m of its entries which is comparably much smaller than $n_1 n_2$, the total number of entries. Can one recover the matrix $\mathbf{M}$ from m of its entries? In general, everyone would agree that this is impossible without some additional information.

In many instances, however, the matrix we wish to recover is known to be structured in the sense that it is low-rank or approximately low-rank. (We recall for completeness that a matrix with n_1 rows and n_2 columns has rank r if its rows or columns span an r-dimensional space.) Below are two examples of practical scenarios where one would like to be able to recover a low-rank matrix from a sampling of its entries.

- *The Netflix problem.* In the area of recommender systems, users submit ratings on a subset of entries in a database, and the vendor provides recommendations based on the user's preferences [19], [21]. Because users only rate a few items, one would like to infer their preference for unrated items.
  A special instance of this problem is the now famous Netflix problem [2]. Users (rows of the data matrix) are given the opportunity to rate movies (columns of the data matrix) but users typically rate only very few movies so that there are very few scattered observed entries of this data matrix. Yet one would like to complete this matrix so that the vendor (here Netflix) might recommend titles that any particular user is likely to be willing to order. In this case, the data matrix of all user-ratings may be approximately low-rank because it is commonly believed that only a few factors contribute to an individual's tastes or preferences.
- *Triangulation from incomplete data.* Suppose we are given partial information about the distances between objects and would like to reconstruct the low-dimensional geometry describing their locations. For example, we may have a network of low-power wirelessly networked sensors scattered randomly across a region. Suppose each sensor only has the ability to construct distance estimates based on signal strength readings from its nearest fellow sensors. From these noisy distance estimates, we can form a partially observed distance matrix. We can then estimate the true distance matrix whose rank will be equal to two if the sensors are located in a plane or three if they are located in three dimensional space [16], [20]. In this case, we only need to observe a few distances per node to have enough information to reconstruct the positions of the objects.

These examples are of course far from exhaustive and there are many other problems which fall in this general category. For instance, we may have some very limited information about a covariance matrix of interest. Yet, this covariance matrix may be low-rank or approximately low-rank because the variables only depend upon a comparably smaller number of factors.

II. Impediments and solutions

Suppose for simplicity that we wish to recover a square $n \times n$ matrix $\mathbf{M}$ of rank r.[1] Such a matrix $\mathbf{M}$ can be represented by n^2 numbers, but it only has $(2n-r)r$ degrees of freedom. This fact can be revealed by counting parameters in the singular value decomposition (the number of degrees of freedom associated with the description of the singular values and of the left and right singular vectors). When

[1] We emphasize that there is nothing special about $\mathbf{M}$ being square and all of our discussion would apply to arbitrary rectangular matrices as well. The advantage of focusing on square matrices is a simplified exposition and reduction in the number of parameters of which we need to keep track.

978-1-4244-2926-4/08/$25.00 ©2008 IEEE

the rank is small, this is considerably smaller than n^2. For instance, when $\mathbf{M}$ encodes a 10-dimensional phenomenon, then the number of degrees of freedom is about $20n$ offering a reduction in dimensionality by a factor about equal to $n/20$. When n is large (e.g. in the thousands or millions), the data matrix carries much less information than its ambient dimension suggests. The problem is now whether it is possible to recover this matrix from a sampling of its entries without having to probe all the n^2 entries, or more generally collect n^2 or more measurements about $\mathbf{M}$.

A. Which matrices?

In general, one cannot hope to be able to recover a low-rank matrix from a sample of its entries. Consider the rank-1 matrix $\mathbf{M}$ equal to

$$\mathbf{M} = \mathbf{e}_1\mathbf{e}_n^* = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \qquad \text{(II.1)}$$

where here and throughout, $\mathbf{e}_i$ is the ith canonical basis vector in Euclidean space (the vector with all entries equal to 0 but the ith equal to 1). This matrix has a 1 in the top-right corner and all the other entries are 0. Clearly this matrix cannot be recovered from a sampling of its entries unless we pretty much see all the entries. The reason is that for most sampling sets, we would only get to see zeros so that we would have no way of guessing that the matrix is not zero. For instance, if we were to see 90% of the entries selected at random, then 10% of the time we would only get to see zeroes.

It is therefore impossible to recover *all* low-rank matrices from a set of sampled entries but can one recover *most* of them? To investigate this issue, we introduce a simple model of low-rank matrices. Consider the singular value decomposition (SVD) of a matrix $\mathbf{M}$

$$\mathbf{M} = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^*, \qquad \text{(II.2)}$$

where the $\mathbf{u}_k$'s and $\mathbf{v}_k$'s are the left and right singular vectors, and the σ_k's are the singular values (the roots of the eigenvalues of $\mathbf{M}^*\mathbf{M}$). Then we could think of a *generic* low-rank matrix as follows: the family $\{\mathbf{u}_k\}_{1\le k\le r}$ is selected uniformly at random among all families of r orthonormal vectors, and similarly for the the family $\{\mathbf{v}_k\}_{1\le k\le r}$. The two families may or may not be independent of each other. We make no assumptions about the singular values σ_k. In the sequel, we will refer to this model as the *random orthogonal model*. This model is convenient in the sense that it is both very concrete and simple, and useful in the sense that it will help us fix the main ideas. In the sequel, however, we will consider far more general models. The question for now is whether or not one can recover such a generic matrix from a sampling of its entries.

B. Which sampling sets?

Clearly, one cannot hope to reconstruct any low-rank matrix $\mathbf{M}$—even of rank 1—if the sampling set avoids any column or row of $\mathbf{M}$. Suppose that $\mathbf{M}$ is of rank 1 and of the form $\mathbf{x}\mathbf{y}^*$, $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ so that the (i,j)th entry is given by

$$M_{ij} = x_i y_j.$$

Then if we do not have samples from the first row for example, one could never guess the value of the first component x_1, by any method whatsoever; no information about x_1 is observed. There is of course nothing special about the first row and this argument extends to any row or column. To have any hope of recovering an unknown matrix, one needs at least one observation per row and one observation per column.

We have just seen that if the sampling is adversarial, e.g. one observes all of the entries of $\mathbf{M}$ but those in the first row, then one would not even be able to recover matrices of rank 1. But what happens for most sampling sets? Can one recover a low-rank matrix from almost all sampling sets of cardinality m? Formally, suppose that the set Ω of locations corresponding to the observed entries ($(i,j) \in \Omega$ if M_{ij} is observed) is a set of cardinality m sampled uniformly at random. Then can one recover a generic low-rank matrix M, perhaps with very large probability, from the knowledge of the value of its entries in the set Ω?

C. Which algorithm?

If the number of measurements is sufficiently large, and if the entries are sufficiently uniformly distributed as above, one might hope that there is only *one* low-rank matrix with these entries. If this were true, one would want to recover the data matrix by solving the optimization problem

$$\begin{array}{ll} \text{minimize} & \text{rank}(\mathbf{X}) \\ \text{subject to} & X_{ij} = M_{ij} \quad (i,j) \in \Omega, \end{array} \qquad \text{(II.3)}$$

where $\mathbf{X}$ is the decision variable and $\text{rank}(\mathbf{X})$ is equal to the rank of the matrix $\mathbf{X}$. The program (II.3) is a common sense approach which simply seeks the simplest explanation fitting the observed data. If there were only one low-rank object fitting the data, this would recover $\mathbf{M}$. This is unfortunately of little practical use because this optimization problem is not only NP-hard, but all known algorithms which provide exact solutions require time doubly exponential in the dimension n of the matrix in both theory and practice [11].

If a matrix has rank r, then it has exactly r nonzero singular values so that the rank function in (II.3) is simply the number of nonvanishing singular values. In this paper, we consider an alternative which minimizes the sum of the singular values over the constraint set. This sum is called the *nuclear norm*,

$$\|\mathbf{X}\|_* = \sum_{k=1}^{n} \sigma_k(\mathbf{X}) \qquad \text{(II.4)}$$

where, here and below, $\sigma_k(\mathbf{X})$ denotes the kth largest singular value of $\mathbf{X}$. The heuristic optimization is then given by

$$\begin{array}{ll} \text{minimize} & \|\mathbf{X}\|_* \\ \text{subject to} & X_{ij} = M_{ij} \quad (i,j) \in \Omega. \end{array} \qquad \text{(II.5)}$$

Whereas the rank function counts the number of nonvanishing singular values, the nuclear norm sums their amplitude

and in some sense, is to the rank functional what the convex ℓ_1 norm is to the counting ℓ_0 norm in the area of sparse signal recovery. The main point here is that the nuclear norm is a convex function and, as we will discuss in Section V can be optimized efficiently via semidefinite programming.

D. A first typical result

Our first result shows that, perhaps unexpectedly, this heuristic optimization recovers a generic $\mathbf{M}$ when the number of randomly sampled entries is large enough. We will prove the following:

Theorem 2.1: Let $\mathbf{M}$ be an $n_1 \times n_2$ matrix of rank r sampled from the random orthogonal model, and put $n = \max(n_1, n_2)$. Suppose we observe m entries of $\mathbf{M}$ with locations sampled uniformly at random. Then there are numerical constants C and c such that if

$$m \geq C n^{5/4} r \log n, \tag{II.6}$$

the minimizer to the problem (II.5) is unique and equal to $\mathbf{M}$ with probability at least $1 - cn^{-3}$; that is to say, the semidefinite program (II.5) recovers all the entries of $\mathbf{M}$ with no error. In addition, if $r \leq n^{1/5}$, then the recovery is exact with probability at least $1 - cn^{-3}$ provided that

$$m \geq C n^{6/5} r \log n. \tag{II.7}$$

The theorem states that a surprisingly small number of entries are sufficient to complete a generic low-rank matrix. For small values of the rank, e.g. when $r = O(1)$ or $r = O(\log n)$, one only needs to see on the order of $n^{6/5}$ entries (ignoring logarithmic factors) which is considerably smaller than n^2—the total number of entries of a squared matrix. The real feat, however, is that the recovery algorithm is tractable and very concrete. Hence the contribution is twofold:

- Under the hypotheses of Theorem 2.1, there is a unique low-rank matrix which is consistent with the observed entries.
- Further, this matrix can be recovered by the convex optimization (II.5). In other words, for most problems, the nuclear norm relaxation is *formally equivalent* to the combinatorially hard rank minimization problem (II.3).

Theorem 2.1 is in fact a special instance of a far more general theorem that covers a much larger set of matrices $\mathbf{M}$. We describe this general class of matrices and precise recovery conditions in the next section.

III. Main results

As seen in our first example (II.1), it is impossible to recover a matrix which is equal to zero in nearly all of its entries unless we see all the entries of the matrix. To recover a low-rank matrix, this matrix cannot be in the null space of the sampling operator giving the values of a subset of the entries. Now it is easy to see that if the singular vectors of a matrix $\mathbf{M}$ are highly concentrated, then $\mathbf{M}$ could very well be in the null-space of the sampling operator. For instance consider the rank-2 symmetric matrix $\mathbf{M}$ given by

$$\mathbf{M} = \sum_{k=1}^{2} \sigma_k \mathbf{u}_k \mathbf{u}_k^*, \qquad \begin{aligned} \mathbf{u}_1 &= (\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}, \\ \mathbf{u}_2 &= (\mathbf{e}_1 - \mathbf{e}_2)/\sqrt{2}, \end{aligned}$$

where the singular values are arbitrary. Then this matrix vanishes everywhere except in the top-left 2×2 corner and one would basically need to see all the entries of $\mathbf{M}$ to be able to recover this matrix exactly by any method whatsoever. There is an endless list of examples of this sort. Hence, we arrive at the notion that, somehow, the singular vectors need to be sufficiently spread—that is, uncorrelated with the standard basis—in order to minimize the number of observations needed to recover a low-rank matrix.[2] This motivates the following definition.

Definition 3.1: Let U be a subspace of $\mathbb{R}^n$ of dimension r and $\mathbf{P}_U$ be the orthogonal projection onto U. Then the *coherence* of U (vis-à-vis the standard basis $(\mathbf{e}_i)$) is defined to be

$$\mu(U) \equiv \frac{n}{r} \max_{1 \leq i \leq n} \|\mathbf{P}_U \mathbf{e}_i\|^2. \tag{III.1}$$

Note that for any subspace, the smallest $\mu(U)$ can be is 1, achieved, for example, if U is spanned by vectors whose entries all have magnitude $1/\sqrt{n}$. The largest possible value for $\mu(U)$ is n/r which would correspond to any subspace that contains a standard basis element. We shall be primarily interested in subspace with low coherence as matrices whose column and row spaces have low coherence cannot really be in the null space of the sampling operator. For instance, we will see that the random subspaces discussed above have nearly minimal coherence.

To state our main result, we introduce two assumptions about an $n_1 \times n_2$ matrix $\mathbf{M}$ whose SVD is given by $\mathbf{M} = \sum_{1 \leq k \leq r} \sigma_k \mathbf{u}_k \mathbf{v}_k^*$ and with column and row spaces denoted by U and V respectively.

- A0 The coherences obey $\max(\mu(U), \mu(V)) \leq \mu_0$ for some positive μ_0.
- A1 The $n_1 \times n_2$ matrix $\sum_{1 \leq k \leq r} \mathbf{u}_k \mathbf{v}_k^*$ has a maximum entry bounded by $\mu_1 \sqrt{r/(n_1 n_2)}$ in absolute value for some positive μ_1.

The μ's above may depend on r and n_1, n_2. Moreover, note that **A1** always holds with $\mu_1 = \mu_0 \sqrt{r}$ since the (i,j)th entry of the matrix $\sum_{1 \leq k \leq r} \mathbf{u}_k \mathbf{v}_k^*$ is given by $\sum_{1 \leq k \leq r} u_{ik} v_{jk}$ and by the Cauchy-Schwarz inequality,

$$\left| \sum_{1 \leq k \leq r} u_{ik} v_{jk} \right| \leq \sqrt{\sum_{1 \leq k \leq r} |u_{ik}|^2} \sqrt{\sum_{1 \leq k \leq r} |v_{jk}|^2} \leq \frac{\mu_0 r}{\sqrt{n_1 n_2}}.$$

Hence, for sufficiently small ranks, μ_1 is comparable to μ_0. As we show in the full version of this paper [10], for larger ranks, both subspaces selected from the uniform distribution and spaces constructed as the span of singular vectors with bounded entries are not only incoherent with the standard basis, but also obey **A1** with high probability for values of μ_1 at most logarithmic in n_1 and/or n_2. We will assume that μ_1 is greater than or equal to 1.

We are in the position to state our main result: if a matrix has row and column spaces that are incoherent with the standard basis, then nuclear norm minimization can recover

[2] Both the left and right singular vectors need to be uncorrelated with the standard basis. Indeed, the matrix $\mathbf{e}_1 \mathbf{v}^*$ has its first row equal to $\mathbf{v}$ and all the others equal to zero. Clearly, this rank-1 matrix cannot be recovered unless we basically see all of its entries.

this matrix from a random sampling of a small number of entries.

Theorem 3.2: Let $\mathbf{M}$ be an $n_1 \times n_2$ matrix of rank r obeying **A0** and **A1** and put $n = \max(n_1, n_2)$. Suppose we observe m entries of $\mathbf{M}$ with locations sampled uniformly at random. Then there exist constants C, c such that if

$$m \geq C \max(\mu_1^2, \mu_0^{1/2}\mu_1, \mu_0 n^{1/4}) nr(\beta \log n) \tag{III.2}$$

for some $\beta > 2$, then the minimizer to the problem (II.5) is unique and equal to $\mathbf{M}$ with probability at least $1 - cn^{-\beta}$. For $r \leq \mu_0^{-1} n^{1/5}$ this estimate can be improved to

$$m \geq C \mu_0 n^{6/5} r(\beta \log n) \tag{III.3}$$

with the same probability of success.

Theorem 3.2 asserts that if the coherence is low, few samples are required to recover $\mathbf{M}$. For example, if $\mu_0 = O(1)$ and the rank is not too large, then the recovery is exact with large probability provided that

$$m \geq C n^{6/5} r \log n. \tag{III.4}$$

We give two illustrative examples of matrices with incoherent column and row spaces. This list is by no means exhaustive.

1) The first example is the random orthogonal model. For values of the rank r greater than $\log n$, $\mu(U)$ and $\mu(V)$ are $O(1)$, $\mu_1 = O(\log n)$ both with very large probability. Hence, the recovery is exact provided that m obeys (II.6) or (II.7). Specializing Theorem 3.2 to these values of the parameters gives Theorem 2.1. Hence, Theorem 2.1 is a special case of our general recovery result.
2) The second example is more general and, in a nutshell, simply requires that the components of the singular vectors of $\mathbf{M}$ are small. Assume that the $\mathbf{u}_j$ and $\mathbf{v}_j$'s obey

$$\max_{ij} |\langle \mathbf{e}_i, \mathbf{u}_j \rangle|^2 \leq \mu_B / n, \quad \max_{ij} |\langle \mathbf{e}_i, \mathbf{v}_j \rangle|^2 \leq \mu_B / n, \tag{III.5}$$

for some value of $\mu_B = O(1)$. Then the maximum coherence is at most μ_B since $\mu(U) \leq \mu_B$ and $\mu(V) \leq \mu_B$. Further, we show in [10] that **A1** holds most of the time with $\mu_1 = O(\sqrt{\log n})$. Thus, for matrices with singular vectors obeying (III.5), the recovery is exact provided that m obeys (III.4) for values of the rank not exceeding $\mu_B^{-1} n^{1/5}$.

The proof of Theorem 3.2 can be found in the full version of this paper [10]. There we establish sufficient conditions which guarantee that the true low-rank matrix $\mathbf{M}$ is the unique solution to (II.5). One of these conditions is the existence of a dual vector obeying two crucial optimality conditions. We construct such a dual vector and then demonstrate that it obeys the desired properties provided that the number of measurements is sufficiently large.

IV. Extensions

Our main result (Theorem 3.2) extends to a variety of other low-rank matrix completion problems beyond the sampling of entries. Indeed, suppose we have two orthonormal bases $\mathbf{f}_1, \ldots, \mathbf{f}_n$ and $\mathbf{g}_1, \ldots, \mathbf{g}_n$ of $\mathbb{R}^n$, and that we are interested in solving the rank minimization problem

$$\begin{array}{ll} \text{minimize} & \text{rank}(\mathbf{X}) \\ \text{subject to} & \mathbf{f}_i^* \mathbf{X} \mathbf{g}_j = \mathbf{f}_i^* \mathbf{M} \mathbf{g}_j, \quad (i,j) \in \Omega, \end{array} \tag{IV.1}$$

This comes up in a number of applications. As a motivating example, there has been a great deal of interest in the machine learning community in developing specialized algorithms for the *multiclass* and *multitask* learning problems (see, e.g., [1], [4], [3]). In multiclass learning, the goal is to build multiple classifiers with the same training data to distinguish between more than two categories. For example, in face recognition, one might want to classify whether an image patch corresponds to an eye, nose, or mouth. In multitask learning, we have a large set of data, but have a variety of different classification tasks, and, for each task, only partial subsets of the data are relevant. For instance, in activity recognition, we may have acquired sets of observations of multiple subjects and want to determine if each observed person is walking or running. However, a different classifier is to be learned for each individual, and it is not clear how having access to the full collection of observations can improve classification performance. Multitask learning aims precisely to take advantage of the access to the full database to improve performance on the individual tasks.

In the abstract formulation of this problem for linear classifiers, we have K classes to distinguish and are given training examples $\mathbf{f}_1, \ldots, \mathbf{f}_n$. For each example, we are given partial labeling information about which classes it belongs or does not belong to. That is, for each example $\mathbf{f}_j$ and class k, we may either be told that $\mathbf{f}_j$ belongs to class k, be told $\mathbf{f}_j$ does not belong to class k, or provided no information about the membership of $\mathbf{f}_j$ to class k. For each class $1 \leq k \leq K$, we would like to produce a linear function $\mathbf{w}_k$ such that $\mathbf{w}_k^* \mathbf{f}_i > 0$ if $\mathbf{f}_i$ belongs to class k and $\mathbf{w}_k^* \mathbf{f}_i < 0$ otherwise. Formally, we can search for the vector $\mathbf{w}_k$ that satisfies the equality constraints $\mathbf{w}_k^* \mathbf{f}_i = y_{ik}$ where $y_{ik} = 1$ if we are told that $\mathbf{f}_i$ belongs to class k, $y_{ik} = -1$ if we are told that $\mathbf{f}_i$ does not belong to class k, and y_{ik} unconstrained if we are not provided information. A common hypothesis in the multitask setting is that the $\mathbf{w}_k$ corresponding to each of the classes together span a very low dimensional subspace with dimension significantly smaller than K [1], [4], [3]. That is, the basic assumption is that

$$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$$

is low-rank. Hence, the multiclass learning problem can be cast as (IV.1) with observations of the form $\mathbf{f}_i^* \mathbf{W} \mathbf{e}_j$.

To see that our theorem provides conditions under which (IV.1) can be solved via nuclear norm minimization, note that there exist unitary transformations $\mathbf{F}$ and $\mathbf{G}$ such that $\mathbf{e}_j = \mathbf{F}\mathbf{f}_j$ and $\mathbf{e}_j = \mathbf{G}\mathbf{g}_j$ for each $j = 1, \ldots, n$. Hence,

$$\mathbf{f}_i^* \mathbf{X} \mathbf{g}_j = \mathbf{e}_i^* (\mathbf{F} \mathbf{X} \mathbf{G}^*) \mathbf{e}_j.$$

Then if the conditions of Theorem 3.2 hold for the matrix $\mathbf{F}\mathbf{X}\mathbf{G}^*$, it is immediate that nuclear norm minimization finds

the unique optimal solution of (IV.1) when we are provided a large enough random collection of the inner products $\mathbf{f}_i^* \mathbf{M} \mathbf{g}_j$. In other words, all that is needed is that the column and row spaces of $\mathbf{M}$ be respectively incoherent with the basis $(\mathbf{f}_i)$ and $(\mathbf{g}_i)$.

From this perspective, we additionally remark that our results likely extend to the case where one observes a small number of arbitrary linear functionals of a hidden matrix $\mathbf{M}$. Set $N = n^2$ and $\mathbf{A}_1, \ldots, \mathbf{A}_N$ be an orthonormal basis for the linear space of $n \times n$ matrices with the usual inner product $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{trace}(\mathbf{X}^* \mathbf{Y})$. Then we expect our results should also apply to the rank minimization problem

$$\begin{array}{ll} \text{minimize} & \mathrm{rank}(\mathbf{X}) \\ \text{subject to} & \langle \mathbf{A}_k, \mathbf{X} \rangle = \langle \mathbf{A}_k, \mathbf{M} \rangle \quad k \in \Omega, \end{array} \tag{IV.2}$$

where $\Omega \subset \{1, \ldots, N\}$ is selected uniformly at random. In fact, (IV.2) is (II.3) when the orthobasis is the canonical basis $(\mathbf{e}_i \mathbf{e}_j^*)_{1 \le i,j \le n}$. Here, those low-rank matrices which have small inner product with all the basis elements $\mathbf{A}_k$ may be recoverable by nuclear norm minimization. To avoid unnecessary confusion and notational clutter, we leave this general low-rank recovery problem for future work.

V. Connections, alternatives and prior art

Nuclear norm minimization is a recent heuristic introduced by Fazel in [15], and is an extension of the trace heuristic often used by the control community, see e.g. [5], [17]. Indeed, when the matrix variable is symmetric and positive semidefinite, the nuclear norm of $\mathbf{X}$ is the sum of the (nonnegative) eigenvalues and thus equal to the trace of $\mathbf{X}$. Hence, for positive semidefinite unknowns, (II.5) would simply minimize the trace over the constraint set:

$$\begin{array}{ll} \text{minimize} & \mathrm{trace}(\mathbf{X}) \\ \text{subject to} & X_{ij} = M_{ij} \quad (i,j) \in \Omega \ . \\ & \mathbf{X} \succeq 0 \end{array}$$

This is a semidefinite program. Even for the general matrix $\mathbf{M}$ which may not be positive definite or even symmetric, the nuclear norm heuristic can be formulated in terms of semidefinite programming as, for instance, the program (II.5) is equivalent to

$$\begin{array}{ll} \text{minimize} & \mathrm{trace}(\mathbf{W}_1) + \mathrm{trace}(\mathbf{W}_2) \\ \text{subject to} & X_{ij} = M_{ij} \quad (i,j) \in \Omega \\ & \begin{bmatrix} \mathbf{W}_1 & \mathbf{X} \\ \mathbf{X}^* & \mathbf{W}_2 \end{bmatrix} \succeq 0 \end{array}$$

with optimization variables $\mathbf{X}$, $\mathbf{W}_1$ and $\mathbf{W}_2$, (see, e.g., [15], [22]). There are many efficient algorithms and high-quality software available for solving these types of problems.

Our work is inspired by results in the emerging field of *compressive sampling* or *compressed sensing*, a new paradigm for acquiring information about objects of interest from what appears to be a highly incomplete set of measurements [7], [9], [14]. In practice, this means for example that high-resolution imaging is possible with fewer sensors, or that one can speed up signal acquisition time in biomedical applications by orders of magnitude, simply by taking far fewer specially coded samples. Mathematically speaking, we wish to reconstruct a signal $\mathbf{x} \in \mathbb{R}^n$ from a small number measurements $\mathbf{y} = \Phi \mathbf{x}$, $\mathbf{y} \in \mathbb{R}^m$, and m is much smaller than n; i.e. we have far fewer equations than unknowns. In general, one cannot hope to reconstruct $\mathbf{x}$ but assume now that the object we wish to recover is known to be structured in the sense that it is sparse (or approximately sparse). This means that the unknown object depends upon a smaller number of unknown parameters. Then it has been shown that ℓ_1 minimization allows recovery of sparse signals from remarkably few measurements: supposing Φ is chosen randomly from a suitable distribution, then with very high probability, all sparse signals with about k nonzero entries can be recovered from on the order of $k \log n$ measurements. For instance, if $\mathbf{x}$ is k-sparse in the Fourier domain, i.e. $\mathbf{x}$ is a superposition of k sinusoids, then it can be perfectly recovered with high probability—by ℓ_1 minimization—from the knowledge of about $k \log n$ of its entries sampled uniformly at random [7].

From this viewpoint, the results in this paper greatly extend the theory of compressed sensing by showing that other types of interesting objects or structures, beyond sparse signals and images, can be recovered from a limited set of measurements. Moreover, the techniques for proving our main results build upon ideas from the compressed sensing literature together with probabilistic tools such as the powerful techniques of Bourgain and of Rudelson for bounding norms of operators between Banach spaces.

Our notion of incoherence generalizes the concept of the same name in compressive sampling. Notably, in [6], the authors introduce the notion of the incoherence of a unitary transformation. Letting $\mathbf{U}$ be an $n \times n$ unitary matrix, the *coherence* of $\mathbf{U}$ is given by

$$\mu(\mathbf{U}) = n \max_{j,k} |U_{jk}|^2.$$

This quantity ranges in values from 1 for a unitary transformation whose entries all have the same magnitude to n for the identity matrix. Using this notion, [6] showed that with high probability, a k-sparse signal could be recovered via linear programming from the observation of the inner product of the signal with $m = \Omega(\mu(\mathbf{U}) k \log n)$ randomly selected columns of the matrix $\mathbf{U}$. This result provided a generalization of the celebrated results about partial Fourier observations described in [7], a special case where $\mu(\mathbf{U}) = 1$. This paper generalizes the notion of incoherence to problems beyond the setting of sparse signal recovery.

In [18], the authors studied the nuclear norm heuristic applied to a related problem where partial information about a matrix $\mathbf{M}$ is available from m equations of the form

$$\langle \mathbf{A}^{(k)}, M \rangle = \sum_{ij} A_{ij}^{(k)} M_{ij} = b_k, \qquad k = 1, \ldots, m, \tag{V.1}$$

where for each k, $\{A_{ij}^{(k)}\}_{ij}$ is an i.i.d. sequence of Gaussian or Bernoulli random variables and the sequences $\{\mathbf{A}^{(k)}\}$ are also independent from each other (the sequences $\{\mathbf{A}^{(k)}\}$ and $\{b_k\}$ are available to the analyst). Building on the concept of *restricted isometry* introduced in [8] in the context of

sparse signal recovery, [18] establishes the first sufficient conditions for which the nuclear norm heuristic returns the minimum rank element in the constraint set. They prove that the heuristic succeeds with large probability whenever the number m of available measurements is greater than a constant times $2nr\log n$ for $n \times n$ matrices. Although this is an interesting result, a serious impediment to this approach is that one needs to essentially measure random projections of the unknown data matrix—a situation which unfortunately does not commonly arise in practice. Further, the measurements in (V.1) give some information about *all* the entries of $\mathbf{M}$ whereas in our problem, information about most of the entries is simply not available. In particular, the results and techniques introduced in [18] do not begin to address the matrix completion problem of interest to us in this paper. As a consequence, our methods are completely different; for example, they do not rely on any notions of restricted isometry. Instead, as we discussed above, we prove the existence of a Lagrange multiplier for the optimization (II.5) that certifies the unique optimal solution is precisely the matrix that we wish to recover.

Finally, we would like to briefly discuss the possibility of other recovery algorithms when the sampling happens to be chosen in a very special fashion. For example, suppose that $\mathbf{M}$ is generic and that we precisely observe every entry in the first r rows and columns of the matrix. Write $\mathbf{M}$ in block form as

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}$$

with $\mathbf{M}_{11}$ an $r \times r$ matrix. In the special case that $\mathbf{M}_{11}$ is invertible and $\mathbf{M}$ has rank r, then it is easy to verify that $\mathbf{M}_{22} = \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12}$. One can prove this identity by forming the SVD of $\mathbf{M}$, for example. That is, if $\mathbf{M}$ is generic, and the upper $r \times r$ block is invertible, and we observe *every* entry in the first r rows and columns, we can recover $\mathbf{M}$. This result immediately generalizes to the case where one observes precisely r rows and r columns and the $r \times r$ matrix at the intersection of the observed rows and columns is invertible. However, this scheme has many practical drawbacks that stand in the way of a generalization to a completion algorithm from a general set of entries. First, if we miss *any* entry in these rows or columns, we cannot recover $\mathbf{M}$, nor can we leverage any information provided by entries of $\mathbf{M}_{22}$. Second, if the matrix has rank less than r, and we observe r rows and columns, a combinatorial search to find the collection that has an invertible square sub-block is required. Moreover, because of the matrix inversion, the algorithm is rather fragile to noise in the entries.

VI. Discussion

A. Improvements

The results discussed here show that under suitable conditions, one can reconstruct an $n \times n$ matrix of rank r from a small number of its sampled entries provided that this number is on the order of $n^{1.2}r\log n$, at least for moderate values of the rank. One would like to know whether better results hold in the sense that exact matrix recovery would be guaranteed with a reduced number of measurements. In particular, recall that an $n \times n$ matrix of rank r depends on $(2n-r)r$ degrees of freedom; is it true then that it is possible to recover most low-rank matrices from on the order of nr—up to logarithmic multiplicative factors—randomly selected entries? Can the sample size be merely proportional to the true complexity of the low-rank object we wish to recover?

In this direction, we would like to emphasize that there is nothing in our approach that apparently prevents us from getting stronger results. Our proof architecture requires bounding an infinite matrix series in the operator norm. We develop a bound on the spectral norm of each of the first four terms of this series and a general argument to bound the remainder of the series in [10]. Presumably, one could bound higher order terms by the same techniques. Getting an appropriate bound on the fifth term would lower the exponent of n from $6/5$ to $7/6$. The appropriate bound on the sixth term would further lower the exponent to $8/7$, and so on. To obtain an optimal result, one would need to reach bound $O(\log n)$ terms. We refer the interested reader to [10] for an discussion of how such an extension might be achieved.

B. Further directions

It would be of interest to extend our results to the case where the unknown matrix is approximately low-rank. Suppose we write the SVD of a matrix $\mathbf{M}$ as

$$\mathbf{M} = \sum_{1 \le k \le n} \sigma_k \mathbf{u}_k \mathbf{v}_k^*,$$

where $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_n \ge 0$ and assume for simplicity that none of the σ_k's vanish. In general, it is impossible to complete such a matrix exactly from a partial subset of its entries. However, one might hope to be able to recover a good approximation if, for example, most of the singular values are small or negligible. For instance, consider the truncated SVD of the matrix $\mathbf{M}$,

$$\mathbf{M}_r = \sum_{1 \le k \le r} \sigma_k \mathbf{u}_k \mathbf{v}_k^*,$$

where the sum extends over the r largest singular values and let $\mathbf{M}_\star$ be the solution to (II.5). Then one would not expect to have $\mathbf{M}_\star = \mathbf{M}$ but it would be of great interest to determine whether the size of $\mathbf{M}_\star - \mathbf{M}$ is comparable to that of $\mathbf{M} - \mathbf{M}_r$ provided that the number of sampled entries is sufficiently large. For example, one would like to know whether it is reasonable to expect that $\|\mathbf{M}_\star - \mathbf{M}\|_*$ is on the same order as $\|\mathbf{M} - \mathbf{M}_r\|_*$ (one could ask for a similar comparison with a different norm). If the answer is positive, then this would say that approximately low-rank matrices can be accurately recovered from a small set of sampled entries.

Another important direction is to determine whether the reconstruction is robust to noise as in some applications, one would presumably observe

$$Y_{ij} = M_{ij} + z_{ij}, \quad (i,j) \in \Omega,$$

where z is a deterministic or stochastic perturbation. In this setup, one would perhaps want to minimize the nuclear norm

subject to $\sum_{(i,j)\in\Omega}(X_{ij} - M_{ij})^2 \leq \varepsilon^2$ where ε is an upper bound on the mean noise level. Can one expect that this algorithm or a variation thereof provides accurate answers? That is, can one expect that the error between the recovered and the true data matrix be proportional to the noise level?

Acknowledgments

E. C. was partially supported by a National Science Foundation grant CCF-515362, by the 2006 Waterman Award (NSF) and by an ONR grant. The authors would like to thank Ali Jadbabaie, Pablo Parrilo, Ali Rahimi, Terence Tao, and Joel Tropp for fruitful discussions about parts of this paper. E. C. would like to thank Arnaud Durand for his careful proof-reading and comments.

REFERENCES

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert, "Low-rank matrix factorization with attributes," Ecole des Mines de Paris, Tech. Rep. N24/06/MM, 2006.

[2] *Proceedings of KDD Cup and Workshop*. ACM SIGKDD and Netflix, 2007, proceedings available online at http://www.cs.uic.edu/~liub/KDD-cup-2007/proceedings.html.

[3] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *Proceedings of the Twenty-fourth International Conference on Machine Learning*, 2007.

[4] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Neural Information Processing Systems*, 2007.

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

[6] E. J. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007. [Online]. Available: http://stacks.iop.org/0266-5611/23/969

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

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

[9] E. J. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[10] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," 2008, submitted for publication. Preprint available at http://lanl.arxiv.org/abs/0805.4471.

[11] A. L. Chistov and D. Y. Grigoriev, "Complexity of quantifier elimination in the theory of algebraically closed fields," in *Proceedings of the 11th Symposium on Mathematical Foundations of Computer Science*, ser. Lecture Notes in Computer Science, vol. 176. Springer Verlag, 1984, pp. 17–31.

[12] V. H. de la Peña, "Decoupling and Khintchine's inequalities for U-statistics," *Ann. Probab.*, vol. 20, no. 4, pp. 1877–1892, 1992.

[13] V. H. de la Peña and S. J. Montgomery-Smith, "Decoupling inequalities for the tail probabilities of multivariate U-statistics," *Ann. Probab.*, vol. 23, no. 2, pp. 806–816, 1995.

[14] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

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

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

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

[18] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization," 2007, submitted to *SIAM Review*.

[19] 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.

[20] A. M.-C. So and Y. Ye, "Theory of semidefinite programming for sensor network localization," *Mathematical Programming, Series B*, vol. 109, 2007.

[21] N. Srebro, "Learning with matrix factorizations," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[22] L. Vandenberghe and S. P. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.