Dense Error Correction for Low-Rank Matrices via Principal Component Pursuit

Arvind Ganesh[†], John Wright[*], Xiaodong Li[‡], Emmanuel J. Candès[‡,§] and Yi Ma[*,†]
[*]Microsoft Research Asia, Beijing, P.R.C
[†]Dept. of Electrical and Computer Engineering, UIUC, Urbana, IL 61801
[‡]Dept. of Mathematics, Stanford University, Stanford, CA 94305
[§]Dept. of Statistics, Stanford University, Stanford, CA 94305

***Abstract*—We consider the problem of recovering a low-rank matrix when some of its entries, whose locations are not known a priori, are corrupted by errors of arbitrarily large magnitude. It has recently been shown that this problem can be solved efficiently and effectively by a convex program named Principal Component Pursuit (PCP), provided that the fraction of corrupted entries and the rank of the matrix are both sufficiently small. In this paper, we extend that result to show that the same convex program, with a slightly improved weighting parameter, exactly recovers the low-rank matrix even if "almost all" of its entries are arbitrarily corrupted, provided the signs of the errors are random. We corroborate our result with simulations on randomly generated matrices and errors.**

I. Introduction

Low-rank matrix recovery and approximation have been extensively studied lately for their great importance in theory and practice. Low-rank matrices arise in many real data analysis problems when the high-dimensional data of interest lie on a low-dimensional linear subspace. This model has been extensively and successfully used in many diverse areas, including face recognition [1], system identification [2], and information retrieval [3], just to name a few.

Principal Component Analysis (PCA) [4] is arguably the most popular algorithm to compute low-rank approximations to a high-dimensional data matrix. Essentially, PCA solves the following optimization problem:

$$\min_{L} \|D - L\| \quad \text{s.t.} \quad \operatorname{rank}(L) \leq r, \tag{1}$$

where $D \in \mathbb{R}^{m \times n}$ is the given data matrix, and $\|\cdot\|$ denotes the matrix spectral norm. The optimal solution to the above problem is the best rank-r approximation (in an ℓ^2 sense) to D [5]. Furthermore, PCA offers the optimal solution when the matrix D is corrupted by i.i.d. Gaussian noise. In addition to theoretical guarantees, the PCA can be computed stably and efficiently via the Singular Value Decomposition (SVD).

The major drawback of PCA is its brittleness to errors of large magnitude, even if such errors affect only a few entries of the matrix D. In fact, a single corrupted entry can throw the low-rank matrix $\hat{L}$ estimated by PCA arbitrarily far from the true solution. Unfortunately, these kinds of *non-Gaussian*, gross errors and corruptions are prevalent in modern data. For example, shadows in a face image corrupt only a small part of the image, but the corrupted pixels can be arbitrarily far from their true values in magnitude.

Thus, the problem at hand is to recover a low-rank matrix L_0 (the principal components) from a corrupted data matrix $D \doteq L_0 + S_0$, where the entries of S_0 can have arbitrary magnitude. Although this problem is intractable (NP-hard) to solve under general conditions, recent studies have discovered that certain convex program can effectively solve this problem under surprisingly broad conditions. The work of [6], [7] has proposed a convex program to recover low-rank matrices when a fraction of their entries have been corrupted by errors of arbitrary magnitude *i.e.,* when the matrix S_0 is sufficiently sparse. This approach, dubbed Principal Component Pursuit (PCP) by [6], suggests solving the following convex optimization problem:

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 \quad \text{s.t.} \quad D = L + S, \tag{2}$$

where $\|\cdot\|_*$ and $\|\cdot\|_1$ denote the matrix nuclear norm (sum of singular values) and 1-norm (sum of absolute values of matrix entries), respectively, and $\lambda > 0$ is a weighting parameter. For square matrices of size $n \times n$, the main result of [6] can be summarized as follows:

> *If the singular vectors of L_0 are not too coherent with the standard basis, and the support of S_0 is random, then solving the convex program* (2) *with $\lambda = n^{-1/2}$ exactly recovers L_0 of rank $O(n/\log^2 n)$ from errors S_0 affecting ρn^2 of the entries, where $\rho > 0$ is a sufficiently small positive constant.*

In this work, we extend the above result to show that under the same assumptions, (2) recovers low-rank matrices even if the fraction of corrupted entries ρ is arbitrarily close to one, provided the signs of the errors are *random*. Equivalently speaking, *almost all* of the matrix entries can be badly corrupted by random errors. The analysis in this paper is a nontrivial modification to the arguments of [6] and leads to a better estimate of the weighting parameter λ that enables this *dense error-correction* performance. We verify our result with simulations on randomly generated matrices.

II. Assumptions and Main Result

For convenience of notation, we consider square matrices of size $n \times n$. The results stated here easily extend to non-square matrices.

Corresponding author: Arvind Ganesh (abalasu2@illinois.edu).
This work was partially supported by the grants NSF IIS 08-49292, NSF ECCS 07-01676, and ONR N00014-09-1-0230.

Assumption A: Incoherence Model for L_0. It is clear that for some low-rank and sparse pairs (L_0, S_0), the problem of separating $M = L_0 + S_0$ into the components that generated it is not well-posed, e.g., if L_0 is itself a sparse matrix. In both matrix completion and matrix recovery, it has proved fruitful to restrict attention to matrices whose singular vectors are not aligned with the canonical basis. This can be formalized via the notion of *incoherence* introduced in [8]. If $L_0 = U\Sigma V^*$ denotes a reduced singular value decomposition of L_0, with $U, V \in \mathbb{R}^{n\times r}$, and $\Sigma \in \mathbb{R}^{r\times r}$, then L_0 is μ-incoherent if

$$\left\{ \begin{array}{rcl} \max_i \|U^* e_i\|^2 & \leq & \mu r/n, \\ \max_i \|V^* e_i\|^2 & \leq & \mu r/n, \\ \|UV^*\|_\infty & \leq & \sqrt{\mu r/n^2}, \end{array} \right. \quad (3)$$

where the e_i's are the canonical basis vectors in $\mathbb{R}^n$. Here, $\|\cdot\|_\infty$ denotes the matrix ∞-norm (maximum absolute value of matrix entries).

Assumption B: Random Signs and Support for S_0. Similarly, it is clear that for some very sparse patterns of corruption, exact recovery is not possible, e.g., if S_0 affects an entire row or column of the observation. In [6], such ambiguities are avoided by placing a random model on $\Omega \doteq \mathrm{supp}(S_0)$, which we also adopt. In this model, each entry (i, j) is included in Ω independently with probability ρ. We say $\Omega \sim \mathrm{Ber}(\rho)$ whenever Ω is sampled from the above distribution. We further introduce a random model for the signs of S_0: we assume that for $(i, j) \in \Omega$, $\mathrm{sgn}((S_0)_{ij})$ is an independent random variable taking values ± 1 with probability $1/2$. Equivalently, under this model, if $E = \mathrm{sgn}(S_0)$, then

$$E_{ij} = \left\{ \begin{array}{rcl} 1, & \text{w.p.} & \rho/2, \\ 0, & \text{w.p.} & 1-\rho, \\ -1, & \text{w.p.} & \rho/2. \end{array} \right. \quad (4)$$

This error model differs from the one assumed in [6], in which the error signs come from any fixed (even adversarial) $n \times n$ sign pattern. The stronger assumption that the signs are random is necessary for dense error correction.[1]

Our main result states that under the above assumptions and models, PCP corrects large fractions of errors. In fact, provided the dimension is high enough and the matrix L_0 is sufficiently low-rank, ρ can be any constant less than one:

Theorem 1 (Dense Error Correction via PCP). Fix any $\rho < 1$. Suppose that L_0 is an $n \times n$ matrix of rank r obeying (3) with incoherence parameter μ, and the entries of $\mathrm{sign}(S_0)$ are sampled i.i.d. according to (4). Then as n becomes large[2], Principal Component Pursuit (2) exactly recovers (L_0, S_0) with high probability, provided

$$\lambda = C_1 \left(4\sqrt{1-\rho} + \frac{9}{4}\right)^{-1} \sqrt{\frac{1-\rho}{\rho n}}, \quad r < \frac{C_2 n}{\mu \log^2 n}, \quad (5)$$

where $0 < C_1 \leq 4/5$ and $C_2 > 0$ are certain constants.[3]

[1]The symmetric distribution for the signs is for notational convenience. Any distribution with non-zero probabilities for ± 1 would suffice.

[2]For ρ closer to one, the dimension n must be larger; formally, $n > n_0(\rho)$. By "high probability", we mean with probability at least $1 - cn^{\beta}$ for some fixed $\beta > 0$.

[3]C_2 is not a numerical constant, and possibly depends on the choice of ρ.

In other words, provided the rank of a matrix is of the order of $n/\mu \log^2 n$, PCP can recover the matrix exactly even when an arbitrarily large fraction of its entries are corrupted by errors of arbitrary magnitude and the locations of the uncorrupted entries are unknown. Furthermore, Theorem 1 holds true for some positive numerical constant C_2 if the rank r is restricted to be at most $C_2 n/\mu \log^3 n$ and $\lambda = (n \log n)^{-1/2}$.

Relations to Existing Results. While [6] has proved that PCP succeeds, with high probability, in recovering L_0 and S_0 exactly with $\lambda = n^{-1/2}$, the analysis required that the fraction of corrupted entries ρ is small. The new result shows that, with random error signs, PCP succeeds with ρ arbitrarily close to one. This result also suggests using a slightly modified weighting parameter λ. Although the new λ is of the same order as $n^{-1/2}$, we identify a dependence on ρ that is crucial for correctly recovering L_0 when ρ is large.

This dense error correction result is not an isolated phenomenon when dealing with high-dimensional highly correlated signals. In a sense, this work is inspired by a conceptually similar result for recovering sparse signal via ℓ_1 minimization [9]. To summarize, to recover a sparse signal x from corrupted linear measurements: $y = Ax + e$, one can solve the convex program $\min \|x\|_1 + \|e\|_1$, s.t. $y = Ax + e$. It has been shown in [9] that if A is sufficiently coherent and x sufficiently sparse, the convex program can exactly recover x even if the fraction of nonzero entries in e approaches one.

The result is also similar in spirit to results on matrix completion [8], [10], [11], which show that under similar incoherence assumptions, low-rank matrices can be recovered from vanishing fractions of their entries.

III. Main Ideas of the Proof

The proof of Theorem 1 follows a similar line of arguments presented in [6], and is based on the idea of constructing a dual certificate W whose existence certifies the optimality of (L_0, S_0). As in [6], the dual certificate is constructed in two parts via a combination of the "golfing scheme" of David Gross [11], and the method of least squares. However, several details of the construction must be modified to accommodate a large ρ.

Before continuing, we fix some notation. Given the compact SVD of $L_0 = U\Sigma V^*$, we let $T \subset \mathbb{R}^{n\times n}$ denote the linear subspace $\{UX^* + YV^* \,|\, X, Y \in \mathbb{R}^{n\times r}\}$. By a slight abuse of notation, we also denote by Ω the linear subspace of matrices whose support is a subset of Ω. We let $\mathcal{P}_T$ and $\mathcal{P}_\Omega$ denote the projection operators T and Ω, respectively.

The following lemma introduces a dual vector that in turn, ensures that (L_0, S_0) is the unique optimal solution to (2).

Lemma 1. (Dual Certificate) Assume $\lambda < 1 - \alpha$ and $\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq 1 - \epsilon$ for some $\alpha, \epsilon \in (0, 1)$. Then, (L_0, S_0) is the unique solution to (2) if there is a pair (W, F) obeying

$$UV^* + W = \lambda\left(\mathrm{sgn}(S_0) + F + \mathcal{P}_\Omega D\right)$$

with $\mathcal{P}_T W = 0$ and $\|W\| \leq \alpha$, $\mathcal{P}_\Omega F = 0$ and $\|F\|_\infty \leq \frac{1}{2}$, and $\|\mathcal{P}_\Omega D\|_F \leq \epsilon^2$.

We prove this lemma in the appendix. Lemma 1 generalizes Lemma 2.5 of [6] as follows:

1) [6] assumes that $\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq 1/2$, whereas we only require that $\|\mathcal{P}_\Omega \mathcal{P}_T\|$ is bounded away from one. By Lemma 2, the former assumption is justified only for small values of ρ (or for small amounts of corruption).
2) While [6] requires that $\|W\| \leq 1/2$, we impose a more general bound on $\|W\|$. We find that a value of α closer to 1 gives a better estimate of λ.

For example, by setting $\alpha = 9/10$, to prove that (L_0, S_0) is the unique optimal solution to (2), it is sufficient to find a dual vector W satisfying

$$\begin{cases} \mathcal{P}_T W = 0, \\ \|W\| < \frac{9}{10}, \\ \|\mathcal{P}_\Omega(UV^* + W - \lambda \mathrm{sgn}(S_0))\|_F \leq \lambda\epsilon^2, \\ \|\mathcal{P}_{\Omega^\perp}(UV^* + W)\|_\infty < \frac{\lambda}{2}, \end{cases} \tag{6}$$

assuming that $\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq 1 - \epsilon$ and $\lambda < 1/10$.

We construct a dual certificate in two parts, $W = W^L + W^S$ using a variation of the golfing scheme [11] presented in [6].

1) *Construction of W^L using the golfing scheme.* The golfing scheme writes $\Omega^c = \cup_{j=1}^{j_0} \Omega_j$, where the $\Omega_j \subseteq [n] \times [n]$ are independent Ber(q), with q chosen so that $(1-q)^{j_0} = \rho$.[4] The choice of q ensures that indeed $\Omega \sim \mathrm{Ber}(\rho)$, while the independence of the Ω_j's allows a simple analysis of the following iterative construction:

   Starting with $Y_0 = 0$, we iteratively define
   $$Y_j = Y_{j-1} + q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_T(UV^* - Y_{j-1}),$$
   and set
   $$W^L = \mathcal{P}_{T^\perp} Y_{j_0}. \tag{7}$$
2) *Construction of W^S using least squares.* We set
   $$W^S = \arg\min \|Q\|_F \quad \text{s.t.} \quad \mathcal{P}_\Omega Q = \lambda\, \mathrm{sgn}(S_0), \quad \mathcal{P}_T Q = 0.$$
   Since $\|\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega\| = \|\mathcal{P}_\Omega \mathcal{P}_T\|^2 < 1$, it is not difficult to show that the solution is given by the Neumann series
   $$W^S = \lambda \mathcal{P}_{T^\perp} \sum_{k \geq 0} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k \mathrm{sgn}(S_0). \tag{8}$$

In the remainder of this section, we present three lemmas that establish the desired main result Theorem 1. The first lemma validates the principal assumption of Lemma 1 that $\|\mathcal{P}_\Omega \mathcal{P}_T\|$ is bounded away from one. The other two lemmas collectively prove that the dual certificate $W = W^L + W^S$ generated by the procedure outlined above satisfies (6) with high probability, and thereby, prove Theorem 1 by virtue of Lemma 1.

Lemma 2. (Corollary 2.7 in [6]) Suppose that $\Omega \sim \mathrm{Ber}(\rho)$ and L_0 obeys the incoherence model (3). Then, with high probability, $\|\mathcal{P}_\Omega \mathcal{P}_T\|^2 \leq \rho + \delta$, provided that $1 - \rho \geq C_0 \delta^{-2} \frac{\mu r \log n}{n}$ for some numerical constant $C_0 > 0$.

This result plays a key role in establishing the following two bounds on W^L and W^S, respectively.

[4]The value of j_0 is specified in Lemma 3.

Lemma 3. Assume that $\Omega \sim \mathrm{Ber}(\rho)$, and $\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq \sigma \doteq \sqrt{\rho} + \delta < 1$. Set $j_0 = 2\lceil \log n \rceil$. Then, under the assumptions of Theorem 1, the matrix W^L obeys, with high probability,

(a) $\|W^L\| < 1/10$,
(b) $\|\mathcal{P}_\Omega(UV^* + W^L)\|_F < \lambda(1-\sigma)^2$,
(c) $\|\mathcal{P}_{\Omega^\perp}(UV^* + W^L)\|_\infty < \frac{\lambda}{4}$.

The proof of this lemma follows that of Lemma 2.8 of [6] exactly – the only difference is that here we need to use tighter constants that hold for larger n. The main tools needed are bounds on the operator norm of $\mathcal{P}_{\Omega_j}\mathcal{P}_T$ (which follow from Lemma 2), as well as bounds on

$$\|Q - q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_T Q\|_\infty / \|Q\|_\infty, \quad \|Q - q^{-1}\mathcal{P}_{\Omega_j} Q\| / \|Q\|_\infty,$$

for any fixed nonzero Q (which are given by Lemmas 3.1 and 3.2 of [6]). These bounds can be invoked thanks to the independence between the Ω_j's in the golfing scheme. We omit the details here due to limited space and invite the interested reader to consult [6].

Lemma 4. Assume that $\Omega \sim \mathrm{Ber}(\rho)$, and that the signs of S_0 are i.i.d. symmetric (and independent of Ω). Then, under the assumptions of Theorem 1, the matrix W^S obeys, with high probability,

(a) $\|W^S\| < 8/10$,
(b) $\|\mathcal{P}_{\Omega^\perp} W^S\|_\infty < \frac{\lambda}{4}$.

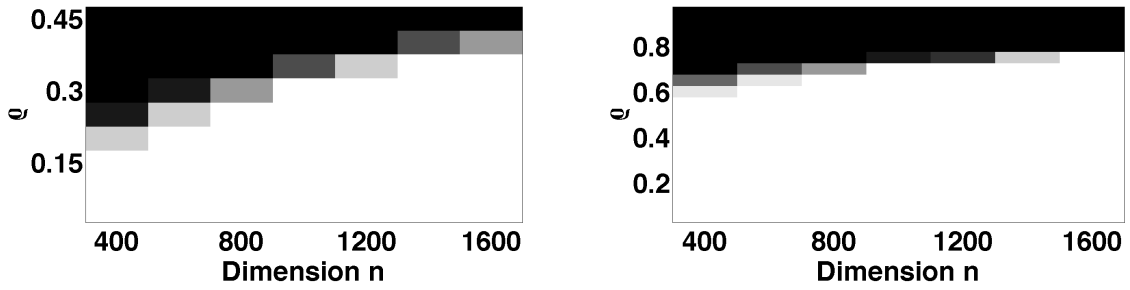
See the appendix for the proof details. The proof of this lemma makes heavy use of the randomness in $\mathrm{sgn}(S_0)$, and the fact that these signs are independent of Ω. The idea is to first bound the norm of the linear operator $\mathcal{R} = \mathcal{P}_{T^\perp} \sum_{k \geq 1} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k$, and then, conditioning on Ω, we use Hoeffding's inequality to obtain a tail bound for $x^* \mathcal{R}(\mathrm{sgn}(S_0)) y$ for any fixed x, y. This extends to a bound on $\|W^S\| = \sup_{\|x\| \leq 1, \|y\| \leq 1} x^* \mathcal{R}(\mathrm{sgn}(S_0)) y$ via a union bound across an appropriately chosen net. We state this argument formally in the appendix.

Although the line of argument here is similar to the proof of Lemma 2.9 in [6], there are some important differences since that work assumed that ρ (and hence, $\|\mathcal{P}_\Omega \mathcal{P}_T\|$) is small. Our analysis gives a tighter probabilistic bound for $\|\mathcal{P}_{T^\perp} \sum_{k \geq 1} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k E\|$, which in turn yields a better estimate of the weighting parameter λ as a function of ρ.

IV. Simulations

In this section, we provide simulation results on randomly generated matrices to support our main result, and suggest potential improvements to the value of λ predicted by our analysis in this paper. For a given dimension n, rank r, and sparsity parameter ρ, we generate L_0 and S_0 as follows:

1) $L_0 = R_1 R_2^*$, where $R_1, R_2 \in \mathbb{R}^{n \times r}$ are random matrices whose entries are i.i.d. distributed according to a normal distribution with mean zero and variance $100/n$.
2) S_0 is a sparse matrix with exactly ρn^2 non-zero entries, whose support is chosen uniformly at random from all



(a) $r = 1, C_1 = 0.8$ (b) $r = 1, C_1 = 4$

Fig. 1. **Dense error correction for varying dimension.** Given n, r, and ρ, we generate $L_0 = R_1 R_2^*$ as the product of two independent $n \times r$ i.i.d. $\mathcal{N}(0, 100/n)$ matrices, and S_0 is a sparse matrix with ρn^2 non-zero entries taking values ± 1 with probability $1/2$. For each pair (n, ρ), the plots show the fraction of successful recoveries over a total of 10 independent trials. Here, white denotes reliable recovery in all trials, and black denotes failure in all trials, with a linear scale for intermediate fractions.

possible supports of size ρn^2.[5] The non-zero entries of S_0 take value ± 1 with probability $1/2$.

We use the augmented Lagrange multiplier method (ALM) [12] to solve (2). This algorithm exhibits good convergence behavior, and since its iterations each have the same complexity as an SVD, it is scalable to reasonably large matrices. Let $(\hat{L}, \hat{S})$ be the optimal solution to (2). The recovery is considered successful if $\frac{\|L_0 - \hat{L}\|_F}{\|L_0\|_F} < 0.01$, *i.e.,* the relative error in the recovered low-rank matrix is less than 1%.

For our first experiment, we fix $\mathrm{rank}(L_0) = 1$. This case demonstrates the best possible error correction behavior for any given dimension n. We vary n from 400 upto 1600, and for each n consider varying $\rho \in (0, 1)$. For each (n, ρ) pair, we choose

$$\lambda = C_1 \cdot \left(4\sqrt{1-\rho} + \frac{9}{4}\right)^{-1} \sqrt{\frac{1-\rho}{n\rho}} \tag{9}$$

with $C_1 = 0.8$ as suggested by Theorem 1. Figure 1(a) plots the fraction of successes across 10 independent trials. Notice that the amount of corruption that PCP can handle increases monotonically with dimension n.

We have found that the λ given by our analysis is actually somewhat pessimistic for moderate n – better error correction behavior in relatively low dimensions can be observed by choosing λ according to (9), but with a larger constant $C_1 = 4$. Figure 1(b) verifies this by repeating the same experiment as in Figure 1(a), but with the modified λ. Indeed, we see larger fractions of error successfully corrected. For instance, we observe that for $n = 1600$, choosing $C_1 = 0.8$ enables reliable recovery when upto 35% of the matrix entries are corrupted, whereas with $C_1 = 4$, PCP can handle upto 75% of corrupted entries. As discussed below, this suggests there is still room for improving our bounds, either by tighter analysis of the current construction or by constructing dual certificates W^S of smaller norm.

V. Discussion

This work showed that PCP in fact corrects large fractions of random errors, provided the matrix to be recovered satisfies the incoherence condition and the corruptions are random in both sign and support. The fact that a higher value of the constant C_1 offers better error-correction performance in moderate dimensions suggests that the analysis in this work can be further strengthened. In our analysis, the value of λ is essentially determined by the spectral norm of W^S; it is reasonable to believe that dual certificates of smaller spectral norm can be constructed by methods other than least squares. Finally, while we have stated our results for the case of square matrices, similar results can be obtained for non-square matrices with minimal modification to the proof.

Appendix: Proof of Lemma 1 and Lemma 4

Proof of Lemma 1.

Proof: Let $UV^* + W_0$ be a subgradient of the nuclear norm at L_0, and $\mathrm{sgn}(S_0) + F_0$ be a subgradient of the ℓ_1-norm at S_0. For any feasible solution $(L_0 + H, S_0 - H)$ to (2),

$$\begin{aligned}\|L_0 + H\|_* + \lambda \|S_0 - H\|_1 \;\geq\; \\ \|L_0\|_* + \lambda \|S_0\|_1 + \langle UV^* + W_0, H\rangle - \lambda\langle \mathrm{sgn}(S_0) + F_0, H\rangle\end{aligned}$$

Choosing W_0 such that $\langle W_0, H\rangle = \|\mathcal{P}_{T^\perp} H\|_*$ and F_0 such that $\langle F_0, H\rangle = -\|\mathcal{P}_{\Omega^\perp} H\|_1$[6] gives

$$\begin{aligned}&\|L_0 + H\|_* + \lambda\|S_0 - H\|_1 \\ &\geq \|L_0\|_* + \lambda\|S_0\|_1 + \|\mathcal{P}_{T^\perp} H\|_* + \lambda\|\mathcal{P}_{\Omega^\perp} H\|_1 \\ &\quad + \langle UV^* - \lambda\mathrm{sgn}(S_0), H\rangle.\end{aligned}$$

By assumption, $UV^* - \lambda\mathrm{sgn}(S_0) = \lambda F - W + \lambda\mathcal{P}_\Omega D$. Since $\|W\| \leq \alpha$, and $\|F\|_\infty \leq \frac{1}{2}$, we have

$$\begin{aligned}&|\langle UV^* - \lambda\mathrm{sgn}(S_0), H\rangle| \\ &\leq \alpha\|\mathcal{P}_{T^\perp} H\|_* + \tfrac{\lambda}{2}\|\mathcal{P}_{\Omega^\perp} H\|_1 + \lambda|\langle \mathcal{P}_\Omega D, H\rangle|.\end{aligned}$$

Substituting the above relation, we get

$$\begin{aligned}&\|L_0 + H\|_* + \lambda\,\|S_0 - H\|_1 \\ &\geq \|L_0\|_* + \lambda\|S_0\|_1 + (1-\alpha)\|\mathcal{P}_{T^\perp} H\|_* + \tfrac{\lambda}{2}\|\mathcal{P}_{\Omega^\perp} H\|_1 \\ &\quad -\lambda|\langle \mathcal{P}_\Omega D, H\rangle| \\ &\geq \|L_0\|_* + \lambda\|S_0\|_1 + (1-\alpha)\|\mathcal{P}_{T^\perp} H\|_* + \tfrac{\lambda}{2}\|\mathcal{P}_{\Omega^\perp} H\|_1 \\ &\quad -\lambda\epsilon^2\|P_\Omega H\|_F\end{aligned}$$

[5] As argued in Appendix 7.1 of [6], from the perspective of success of the algorithm, this uniform model is essentially equivalent to the Bernoulli model.

[6] For instance, $F_0 = -\mathrm{sgn}(\mathcal{P}_{\Omega^\perp} H)$ and $W_0 = \mathcal{P}_{T^\perp} W$, where $\|W\| = 1$ and $\langle W, \mathcal{P}_{T^\perp} H\rangle = \|\mathcal{P}_{T^\perp} H\|_*$. Such a W exists due to the duality between $\|\cdot\|$ and $\|\cdot\|_*$.

We note that

$$\begin{aligned}\|\mathcal{P}_\Omega H\|_F &\leq \|\mathcal{P}_\Omega \mathcal{P}_T H\|_F + \|\mathcal{P}_\Omega \mathcal{P}_{T^\perp} H\|_F \\ &\leq (1-\epsilon)\left(\|\mathcal{P}_\Omega H\|_F + \|\mathcal{P}_{\Omega^\perp} H\|_F\right) + \|\mathcal{P}_{T^\perp} H\|_F\end{aligned}$$

and, therefore,

$$\begin{aligned}\|\mathcal{P}_\Omega H\|_F &\leq \tfrac{1-\epsilon}{\epsilon}\|\mathcal{P}_{\Omega^\perp} H\|_F + \tfrac{1}{\epsilon}\|\mathcal{P}_{T^\perp} H\|_F \\ &\leq \tfrac{1-\epsilon}{\epsilon}\|\mathcal{P}_{\Omega^\perp} H\|_1 + \tfrac{1}{\epsilon}\|\mathcal{P}_{T^\perp} H\|_*.\end{aligned}$$

In conclusion, we have

$$\begin{aligned}&\|L_0 + H\|_* + \lambda \|S_0 - H\|_1 \\ &\geq \|L_0\|_* + \lambda\|S_0\|_1 + \left((1-\alpha) - \lambda\epsilon\right)\|\mathcal{P}_{T^\perp} H\|_* \\ &\quad + \lambda\left(\tfrac{1}{2} - (1-\epsilon)\epsilon\right)\|\mathcal{P}_{\Omega^\perp} H\|_1.\end{aligned}$$

Because $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1$, the intersection of $\Omega \cap T = \{0\}$, and hence, for any nonzero H, at least one of the above terms involving H is strictly positive. ∎

Proof of Lemma 4.

Proof:

Proof of (a). Let $E = \operatorname{sgn}(S_0)$. By assumption, the distribution of each entry of E is given by (4). Using (8) we can express W^S as:

$$\begin{aligned}W^S &= \lambda \mathcal{P}_{T^\perp} E + \lambda \mathcal{P}_{T^\perp} \sum_{k\geq 1} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k E \\ &:= \mathcal{P}_{T^\perp} W_0^S + \mathcal{P}_{T^\perp} W_1^S.\end{aligned}$$

For the first term, we have $\|\mathcal{P}_{T^\perp} W_0^S\| \leq \lambda\|E\|$. Using standard arguments on the norm of a matrix with i.i.d. entries, we have $\|E\| \leq 4\sqrt{n\rho}$ with overwhelming probability [13].

For the second term, we set $\mathcal{R} = \mathcal{P}_{T^\perp} \sum_{k\geq 1} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k$, so $W_1^S = \lambda \mathcal{R}(E)$. Notice that whenever $\|\mathcal{P}_\Omega \mathcal{P}_T\| < 1$,

$$\begin{aligned}\|\mathcal{R}\| &= \|\mathcal{P}_{T^\perp} \sum_{k\geq 1} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k\| \\ &\leq \|\mathcal{P}_{T^\perp} \mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega\| \cdot \|\sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^k\| \\ &\leq \|\mathcal{P}_{T^\perp} \mathcal{P}_\Omega \mathcal{P}_T\| \cdot \|\mathcal{P}_T \mathcal{P}_\Omega\| \cdot \sum_{k\geq 0} \|\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega\|^k \\ &= \|\mathcal{P}_{T^\perp} \mathcal{P}_{\Omega^\perp} \mathcal{P}_T\| \cdot \|\mathcal{P}_T \mathcal{P}_\Omega\| \cdot \sum_{k\geq 0} \|\mathcal{P}_\Omega \mathcal{P}_T\|^{2k} \\ &\leq \frac{\|\mathcal{P}_{\Omega^\perp} \mathcal{P}_T\| \cdot \|\mathcal{P}_\Omega \mathcal{P}_T\|}{1 - \|\mathcal{P}_T \mathcal{P}_\Omega\|^2}. \qquad (10)\end{aligned}$$

Consider the two events:

$$\begin{aligned}\mathcal{E}_1 &:= \{\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq \sqrt{\rho} + \delta\}, \\ \mathcal{E}_2 &:= \{\|\mathcal{P}_{\Omega^\perp} \mathcal{P}_T\| \leq \sqrt{1-\rho} + \delta\}.\end{aligned}$$

For any fixed $\eta > 0$, we can choose $\delta(\eta, \rho) > 0$, such that on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\|\mathcal{R}\| \leq (1+\eta)\sqrt{\frac{\rho}{1-\rho}}. \qquad (11)$$

Since $\Omega \sim \text{Ber}(\rho)$ and $\Omega^c \sim \text{Ber}(1-\rho)$, by Lemma 2, $\mathcal{E}_1 \cap \mathcal{E}_2$ occurs with high probability provided

$$r \leq \delta(\eta, \rho)^2 \min(\rho, 1-\rho) n / \mu \log n. \qquad (12)$$

Since by assumption $r \leq Cn/\mu \log^2 n$, (12) holds for n sufficiently large.

For any $\tau \in (0,1)$, let N_τ denote an τ-net for $\mathbb{S}^{n-1}$ of size at most $(3/\tau)^n$ (see [14] Lemma 3.18). Then, it can be shown that

$$\|\mathcal{R}(E)\| = \sup_{x,y \in \mathbb{S}^{n-1}} \langle y, \mathcal{R}(E) x\rangle \leq (1-\tau)^{-2} \sup_{x,y \in N_\tau} \langle y, \mathcal{R}(E) x\rangle$$

For a fixed pair $(x, y) \in N_\tau \times N_\tau$, we define $X(x,y) \doteq \langle y, \mathcal{R}(E) x\rangle = \langle \mathcal{R}(yx^*), E\rangle$. Conditional on $\Omega = \operatorname{supp}(E)$, the signs of E are i.i.d. symmetric and by Hoeffding's inequality, we have

$$\mathbb{P}(|X(x,y)| > t | \Omega) \leq 2\exp\left(-\frac{2t^2}{\|\mathcal{R}(xy^*)\|_F^2}\right).$$

Since $\|xy^*\|_F = 1$, we have $\|\mathcal{R}(xy^*)\|_F \leq \|\mathcal{R}\|$, so

$$\mathbb{P}\left(\sup_{x,y \in N_\tau} |X(x,y)| > t | \Omega\right) \leq 2|N_\tau|^2 \exp\left(-\frac{2t^2}{\|\mathcal{R}\|^2}\right),$$

and for any fixed $\Omega \in \mathcal{E}_1 \cap \mathcal{E}_2$

$$\mathbb{P}\left(\|\mathcal{R}(E)\| > t | \Omega\right) \leq 2\left(\frac{3}{\tau}\right)^{2n} \exp\left(-\frac{2(1-\tau)^4(1-\rho)t^2}{(1+\eta)^2 \rho}\right).$$

In particular, for any $C > (1+\eta)(1-\tau)^{-2}\sqrt{\log\left(\frac{3}{\tau}\right)}$, $\Omega \in \mathcal{E}_1 \cap \mathcal{E}_2$,

$$\mathbb{P}\left(\|\mathcal{R}(E)\| > C\sqrt{\tfrac{\rho n}{1-\rho}} \mid \Omega\right) < \exp(-C'n),$$

where $C'(C) > 0$. Since $\inf_{0<\tau<1} (1-\tau)^{-2}\sqrt{\log(3/\tau)} < 9/4$, by an appropriate choice of τ and $\eta > 0$, we have

$$\mathbb{P}\left(\|\mathcal{R}(E)\| > \frac{9}{4}\sqrt{\frac{\rho n}{1-\rho}}\right) < \exp(-C'n) + \mathbb{P}((\mathcal{E}_1 \cap \mathcal{E}_2)^c).$$

Thus,

$$\|W^S\| < \lambda\left(4\sqrt{\rho} + \frac{9}{4}\sqrt{\frac{\rho}{1-\rho}}\right)\sqrt{n} \leq 8/10$$

with high probability, provided n is sufficiently large.

Proof of (b) follows the proof of Lemma 2.9 (b) of [6]. ∎

REFERENCES

[1] J. Wright, A. Yang, A. Ganesh, Y. Ma, and S. Sastry, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, Feb 2009.

[2] M. Fazel, H. Hindi, and S. Boyd, "Rank minimization and applications in system theory," in *American Control Conference*, June 2004.

[3] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, no. 2, Oct 2000.

[4] I. Jolliffe, "Principal component analysis," 1986.

[5] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, pp. 211–218, 1936.

[6] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" preprint *submitted to Journal of the ACM*, 2009.

[7] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Sparse and low-rank matrix decompositions," in *IFAC Symposium on System Identification*, 2009.

[8] E. Candès and B. Recht, "Exact matrix completion via convex optimzation," *Found. of Comput. Math.*, 2008.

[9] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," to appear in *IEEE Transactions on Information Theory*, 2008.

[10] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," to appear in *IEEE Transactions on Information Theory*, 2009.

[11] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," preprint, 2009.

[12] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *UIUC Technical Report UILU-ENG-09-2215*, Oct 2009.

[13] R. Vershynin, "Math 280 lecture notes," 2007, available at `http://www-stat.stanford.edu/~dneedell/280.html`.

[14] M. Ledoux, *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.