Compressive Principal Component Pursuit

John Wright[1], Arvind Ganesh[2], Kerui Min[3], and Yi Ma[3,4]

[1] Electrical Engineering Department, Columbia University, New York

[2] IBM Research India

[3] Electrical and Computer Engineering Department, UIUC, Urbana

[4] Microsoft Research Asia, Beijing, China

March 19, 2013

Abstract

We consider the problem of recovering a target matrix that is a superposition of low-rank and sparse components, from a small set of linear measurements. This problem arises in compressed sensing of structured high-dimensional signals such as videos and hyperspectral images, as well as in the analysis of transformation invariant low-rank structure recovery. We analyze the performance of the natural convex heuristic for solving this problem, under the assumption that measurements are chosen uniformly at random. We prove that this heuristic exactly recovers low-rank and sparse terms, provided the number of observations exceeds the number of intrinsic degrees of freedom of the component signals by a polylogarithmic factor. Our analysis introduces several ideas that may be of independent interest for the more general problem of compressed sensing and decomposing superpositions of multiple structured signals.

1 Introduction

In recent years, there has been tremendous interest in recovering low-dimensional structure in high-dimensional signal or data spaces. This interest has been fueled by the striking discovery that efficient techniques based on convex programming can accurately recover low-complexity signals such as sparse vectors or low-rank matrices from severely compressive, incomplete, or even corrupted observations.

One representative example arises in *Robust Principal Component Analysis (RPCA)*. There, the goal is to recover a low-rank matrix $\boldsymbol{L}_0$ from grossly corrupted observations. For example, suppose we observe $\boldsymbol{M} = \boldsymbol{L}_0 + \boldsymbol{S}_0$, where $\boldsymbol{S}_0$ is a sparse error. Under mild conditions, the following convex program, called *Principal Component Pursuit* (PCP) [CLMW11, CSPW11]:

$$\text{minimize} \quad \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \boldsymbol{L} + \boldsymbol{S} = \boldsymbol{M}, \tag{1.1}$$

precisely recovers $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$. In (1.1), $\|\cdot\|_*$ is the matrix *nuclear norm* (sum of singular values) and $\|\cdot\|_1$ is the ℓ^1 norm (sum of magnitudes). For data analysis applications, this suggests that a low-rank matrix $\boldsymbol{L}_0$ can be recovered from the observation $\boldsymbol{M}$ despite large-magnitude sparse errors.

This result has been extended and generalized in a number of directions: to include additional small dense noise $\boldsymbol{M} = \boldsymbol{L}_0 + \boldsymbol{S}_0 + \boldsymbol{N}$ [ZLW+10], large fractions of random errors $\boldsymbol{S}_0$ [GLW+10], and even column-sparse or row-sparse errors [XSC11, MT11].

The conditions under which recovery is known to occur are broad: provided the low-rank term satisfies a technical *incoherence* condition, correct recovery can occur even when $\operatorname{rank}(\boldsymbol{L}_0)$ almost proportional to dimension of the matrix $\boldsymbol{M}$, and the number of nonzero entries in $\boldsymbol{S}_0$ is proportional to the number of entries in $\boldsymbol{M}$ [CLMW11]. On the other hand, in many applications of interest, the rank may actually be significantly smaller than dimension (say 3 [WGS+10], or 9 [BJ03]). Moreover, cardinality of the sparse term may also be quite small. In such a situation our number mn of observations could be extravagantly large compared to the number degrees of freedom in the unknowns $\boldsymbol{L}_0, \boldsymbol{S}_0$. Is it possible to recover $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$ from smaller sets of linear measurements?

1.1 Compressive RPCA

The low-rank and sparse model described above captures properties of many signals of interest, including foreground and background in video surveillance [CLMW11], videos [PGW+12, SA11], structured textures [ZGLM12], hyperspectral datacubes [WSB11, GV11] and more. The ability to recover low-rank and sparse models from small sets of linear measurements could be very useful for developing new sensing architectures for such signals [Don06, WSB11]. Mathematically, our observations have the form

$$\boldsymbol{D} \;\doteq\; \mathcal{P}_Q[\boldsymbol{M}] \;=\; \mathcal{P}_Q[\boldsymbol{L}_0 + \boldsymbol{S}_0], \tag{1.2}$$

where $Q \subseteq \mathbb{R}^{m \times n}$ is a linear subspace, and $\mathcal{P}_Q$ denotes the projection operator onto that subspace. Can we simultaneously recover the low-rank and sparse components correctly from highly compressive measurements via the natural convex program

$$\text{minimize} \quad \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \mathcal{P}_Q[\boldsymbol{L} + \boldsymbol{S}] = \boldsymbol{D}\,? \tag{1.3}$$

While this question is largely open, there is good reason to believe the answer may be positive. For example, [CLMW11, Li11, CJSC13] have studied the "robust matrix completion" problem, with $\mathcal{P}_Q = \mathcal{P}_\Omega$, where Ω is a small subset of the entries of the matrix. When $\mathcal{P}_Q = \mathcal{P}_\Omega$, it is impossible to exactly recover $\boldsymbol{S}_0$ (many of the entries are simply not observed!), but the low-rank term $\boldsymbol{L}_0$ *can* be recovered from near-minimal sets of samples [Li11]. Moreover, Chen et. al. have shown that in this setting a small number of deterministic errors can simultaneously be corrected [CJSC13]. However, in many applications the sparse term $\boldsymbol{S}_0$ is actually the quantity of interest: for example, in visual surveillance, $\boldsymbol{S}_0$ might capture moving foreground objects. To recover both $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$, we must require measurements Q that are incoherent with *both* the low-rank and the sparse term.

In this paper, we investigate the performance of (1.3) when Q is a randomly chosen subspace, (incoherent with $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$ with high probability). As the simulation results in Figure 2 suggest, as long as the rank and sparsity are low enough, we can expect the convex program to correctly recover both the low-rank and sparse components from a reduced set of random linear measurements. A similar recovery problem was recently considered by [WSB11], again, with the goal of designing sensing strategies capable of recovering both $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$. We will discuss the results of [WSB11] and other related works in more detail in Section 3, after we have stated our main result.

(a) The input image with initial region of interest (red box), and the recovered transformation that aligns it to a canonical frame (green box)

(b) The region in the canonical frame (green box) which can be viewed as a low-rank matrix.

Figure 1: **Transform Invariant Low-rank Texture** [ZGLM12] utilizes the Transformed RPCA formulation.

1.2 Transformed RPCA

Aside from the perspective of compressive sensing, there are many other practical scenarios that require recovering a low-rank matrix from partial, incomplete, or corrupted measurements. One example is when the given data is a transformed version of the low-rank and sparse matrices:

$$\boldsymbol{M} \circ \tau = \boldsymbol{L}_0 + \boldsymbol{S}_0, \tag{1.4}$$

where τ is an unknown nonlinear transformation from some continuous group $\mathcal{G}$. The goal is to simultaneously recover $\boldsymbol{L}_0, \boldsymbol{S}_0$ and τ from $\boldsymbol{M}$. One can view this as a "transformed RPCA" problem. The constraint (1.4) is often highly nonlinear. One popular approach is to linearize the measurements against parameters of the transformation:

$$\boldsymbol{M} \circ \tau + \mathcal{J}[\Delta\tau] \approx \boldsymbol{L}_0 + \boldsymbol{S}_0,$$

where $\mathcal{J}$ is the Jacobian of $\boldsymbol{M} \circ \tau$ against of τ. We can then solve for an increment $\Delta\tau = \tau_{k+1} - \tau_k$ in the transformation parameters via the convex program:

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S},\Delta\tau} \ \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \boldsymbol{M} \circ \tau_k + \mathcal{J}[\Delta\tau] = \boldsymbol{L} + \boldsymbol{S}. \tag{1.5}$$

Mathematically, this program is equivalent to (1.3). To see this, let Q be the orthogonal complement to the range of $\mathcal{J}$, so that $\mathcal{P}_Q\mathcal{J} = 0$. Let

$$\boldsymbol{D} \ \doteq \ \mathcal{P}_Q[\boldsymbol{M} \circ \tau + \mathcal{J}[\Delta\tau]] \ = \ \mathcal{P}_Q[\boldsymbol{M} \circ \tau] \ \approx \ \mathcal{P}_Q[\boldsymbol{L}_0 + \boldsymbol{S}_0].$$

After $\Delta\tau$ is eliminated in this way, the problem now becomes recovering the low-rank and sparse components from $\boldsymbol{D}$:

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \ \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \boldsymbol{D} = \mathcal{P}_Q[\boldsymbol{L} + \boldsymbol{S}], \tag{1.6}$$

which has the same form as above.

Empirically, this iterative linearization scheme performs well in applications such as aligning multiple images [PGW+12] and rectifying low-rank textures [ZGLM12], see Figure 1 for an illustration. Hence, it is important to understand under what conditions we should expect the associated convex program to perform correctly. However, there are some important differences from the compressive sensing scenario:

1. In the transformed RPCA case, we often are dealing with a finite dimensional deformation group $\mathcal{G}$ whose dimension, say p, is either fixed (as in [ZGLM12]) or grows very slowly compared to the number of entries in the matrix (as in [PGW+12]).

2. Unlike compressive sensing where the measurement operator $\mathcal{P}_Q(\cdot)$ can be arbitrarily chosen, here it is determined by the given data and the associated transformation group. We can no longer model it as a random projection. Hence, we hope to have deterministic conditions which can be directly verified with the given data.

1.3 Compressive Sensing of Decomposable Components

From both the compressive and transformed RPCA problems, we see the need to understand under what conditions we should expect to correctly recover the low-rank and sparse components from compressive or partial measurements: $\boldsymbol{D} = \mathcal{P}_Q[\boldsymbol{L}_0 + \boldsymbol{S}_0]$. In particular, we are interested in when the convex program:

$$\text{minimize } \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \boldsymbol{D} = \mathcal{P}_Q[\boldsymbol{L} + \boldsymbol{S}], \tag{1.7}$$

finds the correct solution $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$. Following the terminology of [CLMW11], in this paper we refer to this convex program as *Compressive Principal Component Pursuit* (CPCP).

One fundamental question is how many measurements q are needed for the above program (1.7) to correctly recover $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$. Ideally, this number should be related to the number of intrinsic degrees of freedom in $(\boldsymbol{L}_0, \boldsymbol{S}_0)$. If the rank r of $\boldsymbol{L}_0$ is known, we can fully specify $\boldsymbol{L}_0$ using $(m+n-r)r$ real numbers. So, we can consider $\boldsymbol{L}_0$ to have $(m+n-r)r$ degrees of freedom. $\boldsymbol{S}_0$ is slightly more complicated: to specify it, we need to specify both its support Ω, and the $\|\boldsymbol{S}_0\|_0$ real values of $\boldsymbol{S}_0$ on its support.[1] Our bounds will only explicitly depend on the second quantity, $\|\boldsymbol{S}_0\|_0$, which we can consider to be the number of "real degrees of freedom" in $\boldsymbol{S}_0$. We can *define* a quantity which counts these real degrees of freedom:[2]

$$\#\texttt{degrees of freedom}(\boldsymbol{L}_0, \boldsymbol{S}_0) \doteq (m+n-r)r + \|\boldsymbol{S}_0\|_0 . \tag{1.8}$$

Intuitively, the best we can possibly hope for is a number of measurements q on this order. We will show that when the measurements are random (say Gaussian), the desired $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ can indeed be exactly recovered from a number of measurements that is close to this lower bound. Provided

$$\#\texttt{measurements} \geq O(\log^2 m) \times \#\texttt{degrees of freedom}(\boldsymbol{L}_0, \boldsymbol{S}_0),$$

[1] Recall that $\|\boldsymbol{M}\|_0$ is the number of nonzero entries in a matrix $\boldsymbol{M}$.

[2] To be more precise, this quantity is the dimension of the set $\mathcal{M}$ of all pairs $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ consisting of a rank r matrix $\boldsymbol{L}_0$ and a $\|\boldsymbol{S}_0\|_0$-sparse matrix $\boldsymbol{S}_0$, which is a union of manifolds. As our goal in this section is to describe our results qualitatively and give the reader an intuitive feeling for why they are nearly optimal, we simply *define* the quantity number #degrees of freedom as in (1.8), and delay a more general and rigorous treatment to Section 4.

the compressive principal component pursuit program (1.7) correctly recovers this pair with very high probability. This bound is nearly optimal, differing from this hard lower bound by only a polylogarithmic factor.

Our analysis actually pertains to a more general class of problems, in which we seek to decompose a given observation into multiple incoherent components:

$$\text{minimize} \quad \sum_i \lambda_i \left\| \boldsymbol{X}_i \right\|_{(i)} \quad \text{subject to} \quad \sum_i \boldsymbol{X}_i = \boldsymbol{M}. \tag{1.9}$$

Here, $\| \cdot \|_{(i)}$ are norms that encourage various types of low-complexity structure. Principal Component Pursuit [CLMW11, CSPW11], Outlier Pursuit [XSC11, MT11] and Morphological Component Analysis [BSE07] are all special cases of this general problem. Roughly speaking, our analysis will suggest that, if the above program succeeds in recovering all the components $\{\boldsymbol{X}_i\}$ from $\boldsymbol{M}$, one should also expect to recover them from the highly compressive measurements $\mathcal{P}_Q[\boldsymbol{M}]$. The number of measurements required is again governed by the intrinsic degrees of freedom in the components $\{\boldsymbol{X}_i\}$, times at most a polylog(m) oversampling factor. These notions will be made precise in Section 4 below. Because of this additional generality, the results in this paper are potentially applicable to a broad class of source separation or signal decomposition problems that may arise in signal processing, communications, and pattern recognition.

The remainder of this paper is organized as follows. In Section 2, we first introduce the precise mathematical model and present the main result of the paper, as well as extensions to nonvanishing error fractions and deterministic observation operators. In Section 3, we discuss the implications of our results and their relationships with existing work in the literature. Section 4 discusses the more general setting of (1.9) and lays out the framework of our analysis. The remaining sections complete the proof of our main result.

2 Models and Main Results

Our main technical contribution is a procedure for producing a certificate of optimality for $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ for the Compressive Principal Component Pursuit problem, given that the pair is optimal for Principal Component Pursuit. In this sense, our mathematical approach is modular – it partially decouples the analysis of the the compressive measurements from the analysis of the core low-rank and sparse recovery problem. Combining with existing models and analyses of PCP, we can prove that the pair $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ is indeed recoverable by the convex optimization.

We first recall conditions under which $\boldsymbol{M} = \boldsymbol{L}_0 + \boldsymbol{S}_0$ can be exactly separated into its constituents, by PCP. Intuitively, we should not expect to recover all possible low-rank pairs and sparse pairs $(\boldsymbol{L}_0, \boldsymbol{S}_0)$. Indeed, imagine the case when $\boldsymbol{M}$ is rank-one and one-sparse (i.e., $\boldsymbol{M} = \boldsymbol{e}_i \boldsymbol{e}_j^*$ for some i, j). In this situation the answers $(\boldsymbol{L} = \boldsymbol{e}_i \boldsymbol{e}_j^*, \boldsymbol{S} = \boldsymbol{0})$ and $(\boldsymbol{L} = \boldsymbol{0}, \boldsymbol{S} = \boldsymbol{e}_i \boldsymbol{e}_j^*)$ both seem reasonable – the decomposition problem is ambiguous!

To make the problem meaningful, we need conditions that ensure that (i) the low-rank term $\boldsymbol{L}_0$ does not "look sparse" and (ii) the sparse term $\boldsymbol{S}_0$ does not "look low-rank." One popular way formalizing the first intuition of doing this is via the notion of *incoherence* introduced by [CR08]. If the low-rank matrix $\boldsymbol{L}_0$ has rank-reduced singular value decomposition $\boldsymbol{L}_0 = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^*$, then we say that $\boldsymbol{L}_0$ is μ-incoherent if

$$\forall\, i \ \left\| \boldsymbol{U}^* \boldsymbol{e}_i \right\|_2^2 \ \leq \ \frac{\mu r}{m}, \quad \forall\, j \ \left\| \boldsymbol{V}^* \boldsymbol{e}_j \right\|_2^2 \ \leq \ \frac{\mu r}{n}, \quad \text{and} \quad \left\| \boldsymbol{U} \boldsymbol{V}^* \right\|_\infty \ \leq \ \sqrt{\frac{\mu r}{mn}}. \tag{2.1}$$

Intuitively, these conditions ensure that the singular vectors of $\boldsymbol{L}_0$ are not too concentrated on only a few coordinates – the singular vectors do not "look sparse." For further discussion of the implications of this condition, we refer the reader to [CR08].

At the same time, we need to ensure that the sparse term does not "look low-rank." One appealing way of doing this is via a random model: we assume that each (i, j) is an element of $\operatorname{supp}(\boldsymbol{S}_0)$ independently with probability ρ bounded by some small constant. We assume that the signs of the nonzero entries are independent symmetric ± 1 random variables (i.e., Rademacher random variables). In stating our theorems, we call such a distribution an "iid Bernoulli-Rademacher model."

Thus far, we have discussed only the low-rank and sparse terms, but not the properties of the measurements Q. We state a result for Q chosen uniformly at random from the set of all q-dimensional subspaces of $\mathbb{R}^{m \times n}$. More precisely, Q is distributed according to the Haar measure on the Grassmannian $\mathbb{G}(\mathbb{R}^{m \times n}, q)$. On a more intuitive level, this means that Q is equal in distribution to the linear span of a collection of q independent iid $\mathcal{N}(0, 1)$ matrices. In notation more familiar from compressed sensing, we may let $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_q$ denote such a set of matrices, and define an operator $\mathcal{Q} : \mathbb{R}^{m \times n} \to \mathbb{R}^q$ via

$$\mathcal{Q}[\boldsymbol{M}] \;=\; \left(\langle \boldsymbol{Q}_1, \boldsymbol{M} \rangle, \ldots, \langle \boldsymbol{Q}_q, \boldsymbol{M} \rangle\right)^* \;\in\; \mathbb{R}^q. \tag{2.2}$$

Our analysis also pertains to the equivalent convex program:

$$\text{minimize} \quad \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \mathcal{Q}[\boldsymbol{L} + \boldsymbol{S}] = \mathcal{Q}[\boldsymbol{L}_0 + \boldsymbol{S}_0]. \tag{2.3}$$

Since $\mathcal{Q}$ has full rank q almost surely, (2.3) and (1.7) are completely equivalent.

With these assumptions, the following theorem gives a tight bound on the number of (random) measurements required to correctly recover the pair $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ from $\mathcal{P}_Q[\boldsymbol{M}]$ via CPCP:

Theorem 2.1 (Compressive PCP Recovery). *Let $\boldsymbol{L}_0, \boldsymbol{S}_0 \in \mathbb{R}^{m \times n}$, with $m \geq n$, and suppose that $\boldsymbol{L}_0 \neq \boldsymbol{0}$ is a rank-r, μ-incoherent matrix with*

$$r \;\leq\; \frac{c_{\text{rank}}\, n}{\mu \log^2 m}, \tag{2.4}$$

and $\operatorname{sign}(\boldsymbol{S}_0)$ is iid Bernoulli-Rademacher with nonzero probability $\rho < c_{\text{sparse}}$. Let $Q \subset \mathbb{R}^{m \times n}$ be a random subspace of dimension

$$\dim(Q) \;\geq\; C_{\text{meas}} \cdot (\rho m n + m r) \cdot \log^2 m \tag{2.5}$$

distributed according to the Haar measure, probabilistically independent of $\operatorname{sign}(\boldsymbol{S}_0)$. Then with probability at least $1 - C m^{-9}$ in $(\operatorname{sign}(\boldsymbol{S}_0), Q)$, the solution to

$$\textit{minimize} \quad \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \textit{subject to} \quad \mathcal{P}_Q[\boldsymbol{L} + \boldsymbol{S}] = \mathcal{P}_Q[\boldsymbol{L}_0 + \boldsymbol{S}_0] \tag{2.6}$$

with $\lambda = 1/\sqrt{m}$ is unique, and equal to $(\boldsymbol{L}_0, \boldsymbol{S}_0)$. Above, c_{rank}, c_{sparse}, C_{meas}, and C are positive numerical constants.

Here, the magnitudes of the nonzeros in $\boldsymbol{S}_0$ are arbitrary, and no randomness is assumed in $\boldsymbol{L}_0$. The randomness in our this result occurs in the sign and support pattern of $\boldsymbol{S}_0$ and in the measurements Q. We note in passing that the randomness in the signs of $\boldsymbol{S}_0$ can be removed using

the techniques of [CLMW11] Sections 2.1-2.2. For interested readers, Appendix D sketches this argument.

The bounds on r and ρ essentially match those of [CLMW11] for the fully observed case, possibly with different constants. So, again, r and $\|\boldsymbol{S}_0\|_0$ can be rather large. On the other hand, when these quantities are small, the bound on $\dim(Q)$ ensures that the number of measurements needed for accurate recovery is also commensurately small. As we will describe in more detail in Section 4, this result is obtained via general arguments that can also be applied to other decomposition problems.

Two Extensions: Many Errors, Deterministic Q. When the number of errors is very large – say, $\rho = \Omega(1)$ and so a constant fraction of the entries are grossly corrupted – the bound (2.5) in Theorem 2.1 becomes trivial. Indeed, (2.5) is only meaningful when $\rho = O(1/\log^2 m)$. Using a very different argument, which is specific to the Compressive PCP problem, one can also show that it is possible to correct constant fractions of errors ($\rho = \Omega(1)$), with reduced sets of random measurements Q [GMWM12]. The price is that the reduction in measurements may be rather small – that work show successful recovery when the codimension $p = \dim(Q^\perp)$ satisfies $p < Cn$ for an appropriate constant C. From the compressed sensing perspective, this result does not imply strong savings: the number of measurements required is still on the order of mn. However, compared to the argument for Theorem 2.1, the proof of this result is much easier to derandomize.

Derandomizing the arguements and assumptions is important for the transformed low-rank recovery problem, since for that problem, the subspace Q is not random and actually depends on $\boldsymbol{L}$ and $\boldsymbol{S}$. Using arguments of [GMWM12], it is possible to show that if the rank r is small (say $r = o(n^{1/3})$), and $Q^\perp$ is not too coherent with low-rank matrices (i.e., it has a basis whose elements all have small operator norm), then correct recovery is possible. In the application to transformed low-rank recovery, $Q^\perp = \text{span}(\mathcal{J})$ is the linear span of the Jacobian operator with respect to the transformation parameters. To first order, this is the subspace along which we can move, by transforming the given image or images. The operator incoherence condition on $Q^\perp$ can be viewed as asserting that this subspace does not contain any approximately low-rank matrices. We view this result as a step towards a rigorous characterization of when transformed low-rank recovery is possible, at least locally. Readers interested in the details of this these results may refer to the technical report [GMWM12].

Notation. Bold uppercase letters $\boldsymbol{A}, \boldsymbol{B}, \ldots$ denote matrices. Bold lowercase letters $\boldsymbol{x}, \boldsymbol{y}$ denote vectors. Script uppercase letters $\mathcal{A}, \mathcal{B}, \ldots$ denote operators on matrices. In particular, if $S \subset \mathbb{R}^{m \times n}$ is a linear subspace, we will let $\mathcal{P}_S$ denote the orthogonal projection onto S. The notations C, c will always refer to numerical constants. When used in different sections they may not refer to the same constant. All logarithms are base-e. When applied to subsets of a vector space, "+" will denote Minkowski summation, i.e., $A + B = \{\boldsymbol{a} + \boldsymbol{b} \mid \boldsymbol{a} \in A,\ \boldsymbol{b} \in B\}$.

Definition 2.2. *We will say that subspaces $S_1, \ldots, S_k$ are **independent** if*

$$\dim(S_1 + \cdots + S_k) \ = \ \dim(S_1) + \cdots + \dim(S_k).$$

For a matrix $\boldsymbol{M}$, we will let $\|\boldsymbol{M}\|_{2,2}$ denote the ℓ^2 operator norm. For a linear operator $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m' \times n'}$, we will let $\|\mathcal{A}\|_{F,F} = \sup_{\|\boldsymbol{X}\|_F \leq 1} \|\mathcal{A}\boldsymbol{X}\|_F$ denote the operator norm induced by the Frobenius norm.

3 Relationship to the Literature

As mentioned above, in recent years there has been a large amount of work on matrix recovery and decomposition, for example see [CLMW11, CSPW11, ZLW+10, GLW+10, XSC11, MT11, ANW11, HKZ11] and references therein. The aforementioned works mostly pertain to the case when the matrix $\boldsymbol{M}$ is fully observed, and hence are not directly comparable to our result. In Section 4, we will see that our analysis gives a tool for transforming a certificate of optimality for the fully observed problem into a certificate of optimality for the compressive problem. Because this technique is modular, it may be possible to apply it in conjunction with the aforementioned works to prove correct recovery under different assumptions, and even with different regularizers.

Compared to the fully observed problem, there is much less dedicated work on low-rank and sparse recovery from compressive measurements. Recently, motivated by applications in compressive foreground and background separation and compressive hyperspectral image acquisition, [WSB11] introduced a greedy algorithm for this problem, which aims at the objective function

$$\text{minimize}_{\boldsymbol{L},\boldsymbol{S}} \ \|\boldsymbol{D} - \mathcal{P}_Q[\boldsymbol{L}+\boldsymbol{S}]\|_F \quad \text{subject to} \quad \text{rank}\,(\boldsymbol{L}) \leq r,\ \|\boldsymbol{S}\|_0 \leq k. \tag{3.1}$$

Their algorithm is similar in spirit to the CoSaMP algorithm of [NT08] for recovering sparse signals, and performs well on numerical examples. Analyzing its behavior theoretically and proving performance guarantees is currently an open problem.

As mentioned in the introduction, there *is* dedicated theoretical work analyzing matrix recovery problems with both sparse errors and missing entries [CLMW11, Li11, CJSC13]. These works use an entry-wise sampling model that is different from the one considered in [WSB11] and here. That model is very sensible for problems in collaborative filtering, but may not be as useful for compressive sensing, because it does not allow the recovery of the sparse error term. Because the undersampling model assumed in [CLMW11, Li11, CJSC13] is different from the one considered here, these results are not strictly comparable to ours. It is worth noting that both [Li11] and [CJSC13] prove that the low-rank component $\boldsymbol{L}_0$ and *the observed part* of the sparse component $\boldsymbol{S}_0$ can be recovered from a near minimal number of observed entries. This finding is similar in spirit to our observation that $\boldsymbol{L}_0$ and the entire sparse component $\boldsymbol{S}_0$ can be recovered from a near minimal number of random projections, although the details differ.

As the body of results on specific problems such as matrix recovery grows, there has been an increasing interest in unifying or generalizing the basic insights obtained from studying special cases. A number of groups have produced results that pertain to general structured regularizers. For example, Negahban et. al. [NRWY10] have introduced a general geometric framework for analyzing low-complexity signal recovery, highlighting the role of the regularizer in overcoming a lack of strong convexity in the loss. Agarwal et. al. [ANW11] use this framework to analyze sparse and low-rank decomposition, and have obtained tight results for estimation in noise, stronger than previously known results by [ZLW+10]. Their analysis proceeds under different (weaker) assumptions, which preclude exact recovery.

In a similar vein, Chandrasekaran et. al. [CRPW10] have recently produced a very general analysis of structured signal recovery with Gaussian measurements. That work exploits the geometry of the atomic norm ball – in particular, relating the required number of measurements to the Gaussian width of the tangent cone at the desired solution. Based on this, they give tight bounds on the number of measurements needed to recover a low-rank matrix or sparse vector. However, once the atomic set contains both low-rank and sparse matrices, it is less clear how to analyze the Gaussian width of the tangent cone. Indeed, the non-trivial analysis in [CLMW11, CSPW11] can be viewed

as simply showing that the desired solution lies on the boundary of the norm ball. Estimating the width of the tangent cone at that point seems to entail additional analytical difficulty.

For Gaussian measurements, the recent work of Candès and Recht [CR11] also gives simple bounds for exact recovery, under the assumption that the regularizer (or norm) is *decomposable*. If we wished to apply similar analysis to our problem, we would need to work with the quotient norm on $\boldsymbol{M}$:

$$\|\boldsymbol{M}\|_\diamond \doteq \inf_{\boldsymbol{L}+\boldsymbol{S}=\boldsymbol{M}} \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1. \tag{3.2}$$

This is the *infimal convolution* of two decomposable terms. Its subdifferential has a number of nice properties which we will exploit in our analysis, but decomposability (in the sense of [CR11]) does not appear to be one of them. Nevertheless, the results in this paper show that under suitable conditions, we should expect the same type of compressive sensing results for this class of generalized norms for superpositions of low-complexity components.

In this paper, we generalize the analysis of decomposable regularizers to their sums (or strictly speaking infimal convolutions) and obtain nearly optimal bounds on the required number of measurements for exact recovery and decomposition of low-complexity components via convex optimization. In particular, our results provide strong theoretical justification for conducting robust principal component analysis with highly compressive measurements.

4 General Certificate Upgrades

In this section, we present the technical result used to obtain Theorem 2.1 above. As promised, this result will have implications for compressive variants of a large number of conceivable signal decomposition problems. In full generality, we can imagine that the fully observed data $\boldsymbol{M}$ are given as a sum of structured terms:

$$\boldsymbol{M} \;=\; \boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_\tau, \tag{4.1}$$

where each $\boldsymbol{X}_i$ satisfies a low-complexity model such as sparsity or rank-deficiency, possibly also including more exotic types of structured sparsity [Bac10]. For each type of structure, we have a corresponding regularizer $\|\cdot\|_{(i)}$. The natural convex heuristic for decomposing $\boldsymbol{M}$ into its components would solve

$$\text{minimize} \quad \sum_i \lambda_i \|\boldsymbol{X}_i\|_{(i)} \quad \text{subject to} \quad \sum_i \boldsymbol{X}_i = \boldsymbol{M}, \tag{4.2}$$

where the $\lambda_i > 0$ are scalar weight factors. Many authors have studied special cases of this problem, and given conditions under which correct decomposition occurs. A prime example is Principal Component Pursuit; others include Outlier Pursuit [XSC11, MT11] and Morphological Component Analysis [BSE07].

The goal of this paper is not to study (4.2) per se, but rather to understand what happens to it when we only observe compressive measurements of $\boldsymbol{M}$ (or when $\boldsymbol{M}$ itself is subject to some transformation):

$$\text{minimize} \quad \sum_i \lambda_i \|\boldsymbol{X}_i\|_{(i)} \quad \text{subject to} \quad \mathcal{P}_Q\left[\sum_i \boldsymbol{X}_i\right] = \mathcal{P}_Q\boldsymbol{M}. \tag{4.3}$$

Suppose we know that (4.2) correctly decomposes $\boldsymbol{M}$ into $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau$. Does this imply that (4.3) can also recover $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau$? At a slightly more technical level, we can ask whether a certificate

of optimality for the decomposition problem (4.2) can be refined to also certify optimality for the compressive decomposition problem (4.3). Theorem 4.7 below will imply that this is true under broad circumstances. Provided we have proved optimality for (4.2), we can move to optimality for (4.3), as long as the number of measurements $\dim(Q)$ is sufficiently large. In this sense, our analysis is modular: any technique can be used to perform the analysis of the original decomposition problem, provided it constructs an (approximate) dual certificate.

Duality and Optimality. Our result pertains to a class of *decomposable* norms $\|\cdot\|_{(i)}$ [NRWY10, CR11]. This notion includes many sparsity inducing norms, such as the ℓ^1 norm and nuclear norm (as above), as well as sums of block ℓ^p norms.

Definition 4.1. *We say that a norm $\|\cdot\|$ is **locally decomposable** at $\boldsymbol{X}$ if there exists a subspace T and a matrix $\boldsymbol{S}$ such that*

$$\partial\|\cdot\|(\boldsymbol{X}) \;=\; \{\,\boldsymbol{\Lambda} \mid \mathcal{P}_T\boldsymbol{\Lambda} \;=\; \boldsymbol{S},\; \|\mathcal{P}_{T^\perp}\boldsymbol{\Lambda}\|^* \;\leq\; 1\}\,, \tag{4.4}$$

where $\|\cdot\|^*$ denotes the dual norm of $\|\cdot\|$, and $\mathcal{P}_{T^\perp}$ is nonexpansive with respect to $\|\cdot\|^*$.

For example, **the ℓ^1 norm** satisfies this definition with

$$\begin{aligned} T &= \operatorname{supp}(\boldsymbol{X})\,, \\ \boldsymbol{S} &= \operatorname{sign}(\boldsymbol{X})\,, \\ T^\perp &= (\operatorname{supp}(\boldsymbol{X}))^c. \end{aligned}$$

For $\boldsymbol{X}$ with compact singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, **the nuclear norm** satisfies Definition 4.1 with

$$\begin{aligned} T &= \left\{\boldsymbol{U}\boldsymbol{P} + \boldsymbol{Q}\boldsymbol{V}^T \mid \boldsymbol{P} \in \mathbb{R}^{r\times n}, \boldsymbol{Q} \in \mathbb{R}^{m\times r}\right\}, \\ \boldsymbol{S} &= \boldsymbol{U}\boldsymbol{V}^T, \\ T^\perp &= \left\{\boldsymbol{M} \mid \boldsymbol{U}^T\boldsymbol{M} = \boldsymbol{0}, \boldsymbol{M}\boldsymbol{V} = \boldsymbol{0}\right\}. \end{aligned}$$

Finally, if $\Omega_1 \ldots \Omega_k$ are a disjoint partition of $\{1\ldots m\} \times \{1\ldots n\}$, the **group sparse norm**

$$\|\boldsymbol{X}\| = \sum_{i=1}^{k} \|\mathcal{P}_{\Omega_i}\boldsymbol{X}\|_F$$

satisfies Definition 4.1. Indeed, let $I = \{i \mid \mathcal{P}_{\Omega_i}\boldsymbol{X} \neq \boldsymbol{0}\}$. Then in Definition 4.1, we can set

$$\begin{aligned} T &= \left\{\boldsymbol{X} \mid \mathcal{P}_{\Omega_j}[\boldsymbol{X}] = \boldsymbol{0}\;\forall\, j \in I^c\right\}, \\ \boldsymbol{S} &= \sum_{i\in I} \frac{\mathcal{P}_{\Omega_i}[\boldsymbol{X}]}{\left\|\mathcal{P}_{\Omega_j}[\boldsymbol{X}]\right\|_F}, \\ T^\perp &= \{\boldsymbol{X} \mid P_{\Omega_i}[\boldsymbol{X}] = \boldsymbol{0}\;\forall i \in I\}\,. \end{aligned}$$

Definition 4.1 is completely equivalent to that of [CR11] (and there is termed "decomposability"). We have added the modifier "local" to distinguish it from an earlier definition of [NRWY10], which

also encompasses all of the aforementioned structure-inducing norms, but is not strictly equivalent. The modifier "local" is appropriate, because Definition 4.1 only refers to properties of the subdifferential at a particular point $\boldsymbol{X}$. Appendix A discusses in more detail the relationship between these two definitions.

Here, we assume that each $\|\cdot\|_{(i)}$ is locally decomposable at the target solution $\boldsymbol{X}_{i,\star}$, so per the above definition we have a sequence of subspaces T_i and matrices $\boldsymbol{S}_i$ that define the subdifferentials of each of the regularizers $\|\cdot\|_{(i)}$. With this notation in mind, we can state a simple sufficient optimality condition for (4.3):

Lemma 4.2. *Consider a feasible solution $\boldsymbol{x}_\star = (\boldsymbol{X}_{1,\star}, \ldots, \boldsymbol{X}_{\tau,\star})$ to (4.3). Suppose that each of the norms $\|\cdot\|_{(i)}$ is locally decomposable at $\boldsymbol{X}_{i,\star}$. If $T_1, \ldots, T_\tau, Q^\perp$ are independent subspaces and there exists $\boldsymbol{\Lambda}$ satisfying $\mathcal{P}_{T_i}\boldsymbol{\Lambda} = \lambda_i \boldsymbol{S}_i$ and $\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^* < \lambda_i$ for each i, and $\mathcal{P}_{Q^\perp}\boldsymbol{\Lambda} = \boldsymbol{0}$, then $\boldsymbol{x}_\star$ is the unique optimal solution to (4.3).*

Notice that this condition implies that $\boldsymbol{\Lambda}$ lies in the subdifferential of $\lambda_i\|\cdot\|_{(i)}$ for each i. The proof of Lemma 4.2 follows a familiar form[3], and is given in Appendix B. Notice that if we take $Q = \mathbb{R}^{m\times n}$ in Lemma 4.2, we obtain a sufficient optimality condition for the original decomposition problem (4.2). The condition given by Lemma 4.2 is not so convenient to directly work with, because it demands that $\boldsymbol{\Lambda}$ exactly satisfies a set of equality constraints $\mathcal{P}_{T_i}\boldsymbol{\Lambda} = \lambda_i \boldsymbol{S}_i$. One very useful device, due to Gross [Gro11], is to trade off between the equality constraints and the dual norm inequality constraints $\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^* < \lambda_i$, tightening the latter while loosening the former. The following definition gives this idea a name:

Definition 4.3. *We call $\boldsymbol{\Lambda}$ an (α, β)-**inexact certificate** for a putative solution $(\boldsymbol{X}_{1,\star}, \ldots, \boldsymbol{X}_{\tau,\star})$ to (4.2) with parameters $(\lambda_1, \ldots, \lambda_\tau)$ if for each i, $\|\mathcal{P}_{T_i}\boldsymbol{\Lambda} - \lambda_i \boldsymbol{S}_i\|_F \leq \alpha$, and $\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^* < \lambda_i\beta$.*

Comparing to the optimality condition in Lemma 4.2, we can see that this definition is most meaningful when α is small, and $\beta \leq 1$. Indeed, a number of simple and powerful analyses of problems such as matrix completion and robust low-rank matrix recovery proceed by constructing an inexact certificate for which α is polynomial in m^{-1}, and β is a moderate constant, say, $1/2$ [CLMW11, Gro11, Rec11, Li11].

Definition 4.3 pertains to the decomposition problem (4.2), and does not involve the measurement operator Q in any way. Adding one additional constraint, $\mathcal{P}_{Q^\perp}\boldsymbol{\Lambda} = \boldsymbol{0}$, we obtain an inexact certificate for the compressive decomposition problem (4.3):

Definition 4.4. *We call $\boldsymbol{\Lambda}$ an (α, β)-**inexact certificate** for a putative solution $(\boldsymbol{X}_{1,\star}, \ldots, \boldsymbol{X}_{\tau,\star})$ to (4.3) with parameters $(\lambda_1, \ldots, \lambda_\tau)$ if*

(i) $\boldsymbol{\Lambda}$ is an (α, β) inexact certificate for (4.2), and

(ii) $\mathcal{P}_{Q^\perp}\boldsymbol{\Lambda} = \boldsymbol{0}$.

As we will see, an inexact certificate is easier to produce than the "exact" $\boldsymbol{\Lambda}$ demanded in the optimality condition Lemma 4.2. Is it still sufficient to certify optimality? The following lemma shows the answer is *yes*, provided α and β are small enough:

[3] See, e.g., the proof of Lemma 2.4 in [CLMW11], Proposition 2 in [CSPW11], or arguments in [CT05].

Lemma 4.5. *Consider a feasible solution $\boldsymbol{x}_\star = (\boldsymbol{X}_{1,\star}, \ldots, \boldsymbol{X}_{\tau,\star})$ to the optimization problem (4.3). Suppose that each of the norms $\|\cdot\|_{(i)}$ is locally decomposable at $\boldsymbol{X}_{i,\star}$, and that each of the $\|\cdot\|_{(i)}$ majorizes the Frobenius norm. Then if $T_1, \ldots, T_\tau, Q^\perp$ are independent subspaces with*

$$\|\mathcal{P}_{T_i}\mathcal{P}_{T_j}\|_{F,F} \;<\; \frac{1}{\tau - 1} \quad \forall\, i \neq j, \tag{4.5}$$

and there exists an (α, β)-inexact certificate $\hat{\boldsymbol{\Lambda}}$, with

$$\beta + \frac{\sqrt{\tau}}{(1 - \|\mathcal{P}_{Q^\perp}\mathcal{P}_{T_1+\cdots+T_\tau}\|_{F,F}^2)\sqrt{1 - (\tau-1)\max_{ij}\|\mathcal{P}_{T_i}\mathcal{P}_{T_j}\|_{F,F}}} \times \frac{\alpha}{\min_l \lambda_l} < 1, \tag{4.6}$$

then $\boldsymbol{x}_\star$ is the unique optimal solution.

We prove this lemma in Section 6, using a least squares perturbation argument. The additional technical condition that $\|\cdot\|_{(i)}$ majorizes the Frobenius norm (i.e., for all $\boldsymbol{X}$, $\|\boldsymbol{X}\|_{(i)} \geq \|\boldsymbol{X}\|_F$) is immediately satisfied by structure inducing norms such as the nuclear and ℓ^1 norms. In any case, it can always be ensured by rescaling.

Remark 4.6. *The denominator in the condition of Lemma 4.5 depends on our knowledge of the relative orientation of the subspaces $T_1, \ldots, T_\tau$ and Q. We have stated the lemma in a way that assumes bounds on the angles of each pair (T_i, T_j) and between $T_1 + \cdots + T_\tau$ and Q, but demands no additional knowledge. A tighter accounting is possible if more is known about the configuration of $(T_1, \ldots, T_\tau, Q)$.*

It may seem counterintuitive that the denominator in (4.6) depends on the scalings λ_i. In fact, the numerator α also depends on λ_i, since α bounds the norm of the error $\mathcal{P}_{T_i}\hat{\boldsymbol{\Lambda}} - \lambda_i \boldsymbol{S}_i$. In particular, merely scaling the (λ_i) to be very large does not ensure that (4.6) will be satisfied, since any certificate $\hat{\boldsymbol{\Lambda}}$ for this new problem will have a proportionally larger relaxation parameter α.

Thus, to show that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau$ solve the compressive decomposition problem (4.3), we just have produce an inexact certificate $\boldsymbol{\Lambda}$ following the specification of Definition 4.4 with (α, β) sufficiently small. This is fortuitous, since many existing analyses of the original decomposition problem (4.2) already give certificates for that problem. For example, for Principal Component Pursuit, we can leverage existing constructions in [CLMW11]. To prove that the desired solution remains optimal even when we only see a few measurements Q, we will show that a certificate for (4.2) can be "upgraded" to a certificate for (4.3), with very high probability in the choice of random Q, and only a small loss in the parameters (α, β).

Of course, intuitively speaking, this should only be possible if the number of measurements is sufficient: if the number of measurements in Q is smaller than the number of degrees of freedom in $\boldsymbol{x}_\star$, then reconstruction from the compressive measurements $\mathcal{P}_Q \boldsymbol{M}$ should not be possible. Interestingly, however, we will see that the number of measurements does not need to be too much larger than the number of degrees of freedom in $\boldsymbol{x}_\star$: oversampling by $O(\log^2 m)$ will suffice. We have been a bit vague about what we mean by the number of degrees of freedom in the signal. To be precise, our theorem will refer to the quantity $\dim(T_1 + \cdots + T_\tau)$. As mentioned above, for the ℓ^1 norm, $\dim(T_i)$ is the number of nonzero entries in the solution $\boldsymbol{X}_i$. For the nuclear norm, one can check that $\dim(T_i)$ is the number of degrees of freedom in specifying a matrix whose rank is equal to that of $\boldsymbol{X}_i$.

Our main theorem states that with very high probability it is possible to "upgrade" a certificate for the decomposition problem (4.2) to one for the compressive decomposition problem (4.3), with only small loss in parameters (α, β). As it turns out, the loss in the dual norm $\|\cdot\|_{(i)}^*$ will be bounded by the expected dual norm of a standard Gaussian matrix.[4] We will let ν_i denote this quantity:

$$\nu_i \doteq \mathbb{E}\left[\|\boldsymbol{G}\|_{(i)}^*\right], \qquad \boldsymbol{G} \sim_{iid} \mathcal{N}(0,1). \tag{4.7}$$

We have the following theorem:

Theorem 4.7 (Certificate Upgrade). *Consider the general decomposition problem* (4.2)*, and suppose that each of the norms $\|\cdot\|_{(i)}$ majorizes the Frobenius norm. Let $\boldsymbol{x}_\star = (\boldsymbol{X}_{1,\star}, \ldots, \boldsymbol{X}_{\tau,\star})$ be feasible for* (4.2)*, and suppose there exists an (α, β)-inexact certificate $\hat{\boldsymbol{\Lambda}}$ for $\boldsymbol{x}_\star$ for the decomposition problem* (4.2) *with parameters (λ_i).*

Then if $Q \subset \mathbb{R}^{m\times n}$ is a random subspace distributed according to the Haar measure, with

$$\dim(Q) \;\geq\; C_{\text{subspace}} \cdot \dim(T_1 + \cdots + T_\tau) \cdot \log\, m, \tag{4.8}$$

there exists an (α', β')-inexact certificate for $\boldsymbol{x}_\star$ for the compressive *decomposition problem* (4.3) *with*

$$\alpha' \;\leq\; \alpha + m^{-3}\|\hat{\boldsymbol{\Lambda}}\|_F, \tag{4.9}$$

$$\beta' \;\leq\; \beta + C_1 \max_i \frac{\nu_i + \sqrt{\log m}}{\lambda_i}\left(\frac{\|\hat{\boldsymbol{\Lambda}}\|_F^2 \log m}{\dim(Q)}\right)^{1/2}, \tag{4.10}$$

with probability at least $1 - C_2 \cdot \tau \cdot m^{-9}$ in Q. Above, C_{subspace}, C_1 and C_2 are positive numerical constants.

Remark 4.8. *As will become clear in the proof, the degrees m^{-3} and m^{-9} above are arbitrary, and can be set to be any constants by appropriate choice of $C_{\text{subspace}}, C_1, C_2$.*

Remark 4.9 (Scaling in (4.10)**).** *A casual glance at* (4.10) *may suggest that we can make β' arbitrarily close to β, by setting (λ_i) large. This is actually not the case: the initial certificate $\hat{\boldsymbol{\Lambda}}$ must satisfy $\mathcal{P}_{T_i}\hat{\boldsymbol{\Lambda}} \approx \lambda_i \boldsymbol{S}_i$. Scaling all of the λ_i by the same amount will also scale $\hat{\boldsymbol{\Lambda}}$, causing no effective change to the right hand side of* (4.10)*.*

On the other hand, Theorem 4.7 suggests an interesting practical role for the expected norms ν_i in choosing the relative *values of λ_i. Namely, it suggests setting $\lambda_i \propto \nu_i$. This is consistent (within logarithmic factors) with suggestions in [CLMW11], and could suggest a principled way of combining many such structure-inducing norms as in* (4.3)*.*

Our proof uses a variant of the "golfing scheme" of Gross [Gro11]. This general approach divides the observations into (probabilistically) independent subsets, and then constructs a dual certificate iteratively, one subset at a time. While this approach may lose logarithmic factors in the number of measurements, the independence between the current block and the error makes subsequent analysis much more convenient.

[4] This quantity equal, up to a scale of $C\sqrt{mn}$ to the *mean width*, or *Gaussian width*, of the norm ball for $\|\cdot\|_{(i)}$. See, e.g., [Ver09] for more details and calculations of mean widths for various norm balls of interest.

Proof of Theorem 4.7. Let

$$S = T_1 + \cdots + T_\tau + \mathrm{span}(\hat{\mathbf{\Lambda}}), \tag{4.11}$$

where + denotes Minkowski summation. Then S is a linear subspace of dimension at most $\dim(T_1 + \cdots + T_\tau) + 1$ containing $\hat{\mathbf{\Lambda}}$. Our goal is to generate a certificate $\mathbf{\Lambda}_\star$ that is close to $\hat{\mathbf{\Lambda}}$ on S, and also satisfies

$$\mathcal{P}_{Q^\perp} \mathbf{\Lambda}_\star = \mathbf{0}. \tag{4.12}$$

Such a $\mathbf{\Lambda}_\star$ would inherit the good properties of $\hat{\mathbf{\Lambda}}$ on $T_1 + \cdots + T_\tau \subseteq S$, and also satisfy the additional equality constraint (4.12) – in effect, certifying that the measurements are sufficient.

To this end, we will set

$$\mathbf{\Lambda}_0 \;=\; \mathbf{0}, \tag{4.13}$$

We will generate inductively a sequence $(\mathbf{\Lambda}_j)_{j=1,\ldots,k}$ for appropriate k, such that with high probability $\mathbf{\Lambda}_\star = \mathbf{\Lambda}_k$ is the desired certificate. The initial guess $\mathbf{\Lambda}_0$ obviously satisfies (4.12), but could be very far from $\hat{\mathbf{\Lambda}}$ on S. Define the error at step j to be

$$\boldsymbol{E}_j \;=\; \mathcal{P}_S[\mathbf{\Lambda}_j] - \hat{\mathbf{\Lambda}} \;\in\; S. \tag{4.14}$$

We will generate a sequence of corrections, each of which lies in Q, that drive $\boldsymbol{E}_j$ toward zero.

By orthogonal invariance, Q is equal in distribution to the linear span of

$$\boldsymbol{H}_1, \ldots, \boldsymbol{H}_{\dim(Q)},$$

where $\boldsymbol{H}_j$ are independent iid $\mathcal{N}(0, 1/mn)$ random matrices. Choose from $\{1, \ldots, \dim(Q)\}$,

$$k = \lceil 3 \log_2 m \rceil \tag{4.15}$$

disjoint subsets $I_1, \ldots, I_k$ of size

$$\gamma \;=\; \left\lfloor \frac{\dim(Q)}{k} \right\rfloor. \tag{4.16}$$

Our choice of constant ensures that $2^{-k} \leq m^{-3}$. We will require that

$$\gamma \;\geq\; C_3 \cdot \dim(S), \tag{4.17}$$

where C_3 is a numerical constant to be specified later. Since by assumption

$$\dim(Q) \;\geq\; C_{\text{subspace}} \cdot \dim(T_1 + \cdots + T_\tau) \cdot \log(m),$$

once C_3 is chosen, we can ensure that C_{subspace} is large enough that (4.17) holds.

Let $\mathcal{A}_j : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ denote the semidefinite operator that acts via

$$\mathcal{A}_j[\cdot] = \sum_{i \in I_j} \boldsymbol{H}_i \langle \boldsymbol{H}_i, \cdot \rangle. \tag{4.18}$$

Notice that $\mathbb{E}[\mathcal{A}_j] = \frac{\gamma}{mn} \mathcal{I}$. For $j = 1, \ldots, k$, let

$$\begin{aligned} \mathbf{\Lambda}_j &= \mathbf{\Lambda}_{j-1} - \frac{mn}{\gamma} \mathcal{A}_j \boldsymbol{E}_{j-1} \\ &= -\sum_{i=1}^{j} \frac{mn}{\gamma} \mathcal{A}_i \boldsymbol{E}_{i-1}. \end{aligned} \tag{4.19}$$

Then we have

$$\begin{aligned} \boldsymbol{E}_j &= \mathcal{P}_S[\boldsymbol{\Lambda}_j] - \hat{\boldsymbol{\Lambda}} \\ &= \mathcal{P}_S[\boldsymbol{\Lambda}_{j-1}] - \hat{\boldsymbol{\Lambda}} - \mathcal{P}_S \frac{mn}{\gamma} \mathcal{A}_j \boldsymbol{E}_{j-1} \\ &= \mathcal{P}_S \left(\mathcal{I} - \frac{mn}{\gamma} \mathcal{A}_j \right) \mathcal{P}_S \boldsymbol{E}_{j-1}. \end{aligned} \tag{4.20}$$

In paragraph (i) below, we will use this expression to control the Frobenius norm of $\boldsymbol{E}_k$. We may further write

$$\begin{aligned} \boldsymbol{\Lambda}_k &= \mathcal{P}_S[\boldsymbol{\Lambda}_k] + \mathcal{P}_{S^\perp}[\boldsymbol{\Lambda}_k], \\ &= \hat{\boldsymbol{\Lambda}} + \boldsymbol{E}_k - \sum_{j=1}^{k} \mathcal{P}_{S^\perp} \frac{mn}{\gamma} \mathcal{A}_j \mathcal{P}_S \boldsymbol{E}_{j-1}, \end{aligned} \tag{4.21}$$

where in (4.21) we have used that $\boldsymbol{E}_j \in S$ for all j. In paragraphs (ii)-(iii) below, we will use this final expression to control the dual norms of $\boldsymbol{\Lambda}_k$.

(i) Driving $\boldsymbol{E}$ to zero. Ensuring that C_3 in (4.17) is sufficiently large that the hypotheses of Lemma 5.1 are verified, we have that with probability at least $1 - C_4 \exp(-c_1\gamma)$,

$$\left\| \mathcal{P}_S \left(\mathcal{I} - \frac{mn}{\gamma} \mathcal{A}_j \right) \mathcal{P}_S \right\|_{F,F} = \left\| \mathcal{P}_S \frac{mn}{\gamma} \mathcal{A}_j \mathcal{P}_S - \mathcal{P}_S \right\|_{F,F} \leq \frac{1}{2}. \tag{4.22}$$

Hence, by (4.20), we have $\|\boldsymbol{E}_j\|_F \leq \frac{1}{2} \|\boldsymbol{E}_{j-1}\|_F$ for each j, on the complement of a bad event $\mathcal{E}_{\text{err}}$ of probability at most $C_4 k \exp(-c_1\gamma)$. On $\mathcal{E}_{\text{err}}^c$, $\|\boldsymbol{E}_j\|_F \leq 2^{-j} \|\boldsymbol{E}_0\|_F$ for each j, giving

$$\|\boldsymbol{E}_k\|_F \leq 2^{-k} \|\hat{\boldsymbol{\Lambda}}\|_F \quad \text{and} \quad \sum_{j=0}^{k} \|\boldsymbol{E}_j\|_F \leq 2 \|\hat{\boldsymbol{\Lambda}}\|_F. \tag{4.23}$$

(ii) Analysis of α'. From the definition, we may set $\alpha' = \max_i \|\mathcal{P}_{T_i} \boldsymbol{\Lambda}_k - \lambda_i \boldsymbol{S}_i\|_F$. On $\mathcal{E}_{\text{err}}^c$,

$$\begin{aligned} \|\mathcal{P}_{T_i} \boldsymbol{\Lambda}_k - \lambda_i \boldsymbol{S}_i\|_F &= \|\mathcal{P}_{T_i}[\hat{\boldsymbol{\Lambda}} + \boldsymbol{E}_k] - \lambda_i \boldsymbol{S}_i\|_F, \\ &\leq \|\mathcal{P}_{T_i} \hat{\boldsymbol{\Lambda}} - \lambda_i \boldsymbol{S}_i\|_F + \|\boldsymbol{E}_k\|_F, \\ &\leq \|\mathcal{P}_{T_i} \hat{\boldsymbol{\Lambda}} - \lambda_i \boldsymbol{S}_i\|_F + 2^{-k} \|\hat{\boldsymbol{\Lambda}}\|_F, \\ &\leq \|\mathcal{P}_{T_i} \hat{\boldsymbol{\Lambda}} - \lambda_i \boldsymbol{S}_i\|_F + m^{-3} \|\hat{\boldsymbol{\Lambda}}\|_F. \end{aligned} \tag{4.24}$$

Since the first term is bounded by α, the claim in (4.9) is established.

(iii) Analysis of β'. Similarly, for β', we can take

$$\beta' = \max_{i=1,\ldots,\tau} \lambda_i^{-1} \|\boldsymbol{\Lambda}_k\|_{(i)}^* \tag{4.25}$$

From (4.21) and the triangle inequality, we have

$$\|\boldsymbol{\Lambda}_k\|_{(i)}^* \;\leq\; \|\hat{\boldsymbol{\Lambda}}\|_{(i)}^* + \|\boldsymbol{E}_k\|_F + \sum_{j=1}^{k} \left\| \mathcal{P}_{S^\perp} \frac{mn}{\gamma} \mathcal{A}_j \mathcal{P}_S \boldsymbol{E}_{j-1} \right\|_{(i)}^*, \tag{4.26}$$

where we used the fact that whenever the primal norm majorizes the Frobenius norm, its dual minorizes the Frobenius norm. Applying Lemma 5.2, this is bounded by

$$\begin{aligned} \|\boldsymbol{\Lambda}_k\|_{(i)}^* &\leq \lambda_i \beta + 2^{-k} \|\boldsymbol{E}_0\|_F + 10 \frac{\nu_i + \sqrt{\log m}}{\sqrt{\gamma}} \sum_{j=1}^{k} \|\boldsymbol{E}_{j-1}\|_F, \\ &\leq \lambda_i \beta + \left(2^{-k} + 20 \frac{\nu_i + \sqrt{\log m}}{\sqrt{\gamma}} \right) \|\boldsymbol{E}_0\|_F, \\ &\leq \lambda_i \beta + 21 \frac{\nu_i + \sqrt{\log m}}{\sqrt{\gamma}} \|\hat{\boldsymbol{\Lambda}}\|_F \end{aligned} \tag{4.27}$$

on the complement of an event $\mathcal{E}_\infty$ of probability at most $2km^{-10} + k \exp\left(-\frac{\gamma}{2}\right) + \mathbb{P}[\,\mathcal{E}_{\text{err}}\,]$. In the final line, we have used that $2^{-k} \leq m^{-3}$ and $\gamma \leq m^2$. Since $\gamma \;\geq\; \frac{c \cdot \dim(Q)}{\log m}$, for some numerical constant C_5,

$$\lambda_i^{-1} \|\boldsymbol{\Lambda}_k\|_{(i)}^* \;\leq\; \beta + C_5 \frac{\nu_i + \sqrt{\log m}}{\lambda_i} \left(\frac{\|\hat{\boldsymbol{\Lambda}}\|_F^2 \log m}{\dim(Q)} \right)^{1/2}. \tag{4.28}$$

Taking a union bound over $i = 1, \ldots, \tau$ demonstrates the desired bound on β', completing the proof. □

5 Key Probabilistic Lemmas

This section introduces two probabilistic lemmas used in the analysis of the golfing scheme. The first lemma shows that $\mathcal{P}_S \frac{mn}{\gamma} \mathcal{A}_j \mathcal{P}_S \approx \mathcal{P}_S$. Its (routine) proof is delayed to the appendix. The second lemma is crucial for controlling the dual norms of $\mathcal{P}_{S^\perp} \frac{mn}{\gamma} \mathcal{A}_j \mathcal{P}_S$, and is proved in this section.

Lemma 5.1. *There exist numerical constants $C_1, C_2, c > 0$ such that the following holds. Let $S \subseteq \mathbb{R}^{m \times n}$ be a fixed linear subspace, and let $\mathcal{A} = \sum_{j=1}^{\gamma} \boldsymbol{H}_j \langle \boldsymbol{H}_j, \cdot \rangle$, where $(\boldsymbol{H}_j)$ is a sequence of independent iid $\mathcal{N}(0, 1/mn)$ random matrices, and let $R = \text{range}(\mathcal{A}) \subseteq \mathbb{R}^{m \times n}$. Then if*

$$\gamma \;\geq\; C_1 \cdot \dim(S), \tag{5.1}$$

with probability at least $1 - C_2 \exp(-c\gamma)$,

$$\left\| \mathcal{P}_S \frac{mn}{\gamma} \mathcal{A} \mathcal{P}_S - \mathcal{P}_S \right\|_{F,F} \;\leq\; \frac{1}{2}, \tag{5.2}$$

and

$$\left\| \mathcal{P}_S \mathcal{P}_R \mathcal{P}_S - \frac{\gamma}{mn} \mathcal{P}_S \right\|_{F,F} \;\leq\; \frac{1}{16} \frac{\gamma}{mn}. \tag{5.3}$$

The proof of this result follows a familiar covering argument (see, e.g., [Ver11]). For completeness, we give this proof in Appendix C. We next state and prove the key probabilistic lemma for analyzing the "upgrade" procedure introduced in the proof of Theorem 4.7. This lemma allows us to control the dual norm of the constructed certificate. The result is as follows:

Lemma 5.2. *Let S be any fixed subspace of $\mathbb{R}^{m\times n}$ ($m \geq n$), $\boldsymbol{M}$ any fixed matrix. Let $\mathcal{A} = \sum_{l=1}^{\gamma} \boldsymbol{H}_l \langle \boldsymbol{H}_l, \cdot \rangle$ be a random semidefinite operator constructed from a sequence of independent iid $\mathcal{N}(0, 1/mn)$ matrices $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_\gamma$. Let $\|\cdot\|$ be any norm that majorizes the Frobenius norm, and let $\|\cdot\|^*$ be its dual norm. Set $\nu = \mathbb{E}\,[\,\|\boldsymbol{G}\|^*]$, with $\boldsymbol{G}$ iid $\mathcal{N}(0,1)$. Then we have*

$$\left\| \mathcal{P}_{S^\perp} \frac{mn}{\gamma} \mathcal{A} \mathcal{P}_S \boldsymbol{M} \right\|^* \;\leq\; 10 \, \|\mathcal{P}_S \boldsymbol{M}\|_F \, \frac{\nu + \sqrt{\log m}}{\sqrt{\gamma}}, \tag{5.4}$$

with probability at least $1 - m^{-10} - \exp\left(-\frac{\gamma}{2}\right)$.

Proof. Let $\Gamma = \{\boldsymbol{N} \mid \|\boldsymbol{N}\| \leq 1\} \subset \{\boldsymbol{N} \mid \|\boldsymbol{N}\|_F \leq 1\}$ denote the unit ball for $\|\cdot\|$. Then

$$\|\mathcal{P}_{S^\perp} \mathcal{A} \mathcal{P}_S \boldsymbol{M}\|^* = \sup_{\boldsymbol{N} \in \Gamma} \langle \boldsymbol{N}, \mathcal{P}_{S^\perp} \mathcal{A} \mathcal{P}_S \boldsymbol{M} \rangle \,.$$

Inner products of this form are particularly easy to control because they involve projections of $\mathcal{A}$ onto orthogonal subspaces. Since S and $S^\perp$ are orthogonal and $\boldsymbol{H}_l$ is iid Gaussian, $\mathcal{P}_S \boldsymbol{H}_l$ and $\mathcal{P}_{S^\perp} \boldsymbol{H}_l$ are probabilistically independent. So, letting $\boldsymbol{H}'_1, \ldots, \boldsymbol{H}'_\gamma$ denote an independent copy of $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_\gamma$, we have

$$\begin{aligned}
\mathcal{P}_{S^\perp} \mathcal{A} \mathcal{P}_S [\,\cdot\,] \quad &= \quad \mathcal{P}_{S^\perp} \sum_l \boldsymbol{H}_l \, \langle \boldsymbol{H}_l, \mathcal{P}_S [\,\cdot\,] \rangle \\
&= \quad \sum_l \mathcal{P}_{S^\perp} [\boldsymbol{H}_l] \, \langle \mathcal{P}_S \boldsymbol{H}_l, \cdot \rangle \\
&\equiv_d \quad \sum_l \mathcal{P}_{S^\perp} [\boldsymbol{H}_l] \, \langle \mathcal{P}_S \boldsymbol{H}'_l, \cdot \rangle \quad \doteq \quad \mathcal{D},
\end{aligned} \tag{5.5}$$

where $\equiv_d$ denotes equality in distribution. Hence, we have

$$\|\mathcal{P}_{S^\perp} \mathcal{A} \mathcal{P}_S \boldsymbol{M}\|^* \equiv_d \|\mathcal{D} \boldsymbol{M}\|^*.$$

We find the second term more convenient to analyze. Conditioned on $\boldsymbol{H}'_1, \ldots, \boldsymbol{H}'_\gamma$,

$$\xi_{\boldsymbol{N}} \;\doteq\; \langle \boldsymbol{N}, \mathcal{D} \boldsymbol{M} \rangle \;= \sum_l \langle \mathcal{P}_S \boldsymbol{H}'_l, \boldsymbol{M} \rangle \, \langle \mathcal{P}_{S^\perp} \boldsymbol{N}, \boldsymbol{H}_l \rangle \tag{5.6}$$

is zero-mean Gaussian. We have $\|\mathcal{D}\boldsymbol{M}\|^* = \sup_{\boldsymbol{N} \in \Gamma} \xi_{\boldsymbol{N}}$. A quick calculation shows that for any $\boldsymbol{N}$ and $\boldsymbol{N}'$,

$$\begin{aligned}
\mathbb{E}\left[(\xi_{\boldsymbol{N}} - \xi_{\boldsymbol{N}'})^2 \mid \boldsymbol{H}'_1, \ldots, \boldsymbol{H}'_\gamma \right] \quad &= \quad \frac{\left\| \mathcal{P}_{S^\perp} (\boldsymbol{N} - \boldsymbol{N}') \right\|_F^2}{mn} \sum_{l=1}^{\gamma} \langle \mathcal{P}_S \boldsymbol{H}'_l, \boldsymbol{M} \rangle^2 \\
&\leq \quad \frac{\|\boldsymbol{N} - \boldsymbol{N}'\|_F^2}{mn} \sum_{l=1}^{\gamma} \langle \mathcal{P}_S \boldsymbol{H}'_l, \boldsymbol{M} \rangle^2 \;=\; \frac{\Xi^2 \|\boldsymbol{N} - \boldsymbol{N}'\|_F^2}{mn},
\end{aligned} \tag{5.7}$$

where we have let

$$\Xi \;=\; \left(\sum_{l=1}^{\gamma} \left\langle \boldsymbol{H}'_l, \mathcal{P}_S \boldsymbol{M} \right\rangle^2\right)^{1/2} \tag{5.8}$$

Consider a second zero-mean Gaussian process $(\zeta_{\boldsymbol{N}})_{\boldsymbol{N}\in\Gamma}$, defined by letting $\boldsymbol{G}$ be an iid $\mathcal{N}(0, 1/mn)$ matrix, and setting $\zeta_{\boldsymbol{N}} \;=\; \langle \boldsymbol{N}, \boldsymbol{G}\rangle$. From the definition of ν,

$$\mathbb{E}\left[\sup_{\boldsymbol{N}\in\Gamma} \zeta_{\boldsymbol{N}}\right] = \frac{\nu}{\sqrt{mn}}. \tag{5.9}$$

Another calculation shows that

$$\mathbb{E}\left[(\zeta_{\boldsymbol{N}} - \zeta_{\boldsymbol{N}'})^2\right] = \frac{\|\boldsymbol{N} - \boldsymbol{N}'\|_F^2}{mn}. \tag{5.10}$$

By Slepian's inequality (e.g., [Ver11] Lemma 5.33, [LT91] Chapter 3), we have

$$\mathbb{E}\left[\sup_{\boldsymbol{N}} \xi_{\boldsymbol{N}} \mid \boldsymbol{H}'_1, \ldots, \boldsymbol{H}'_\gamma\right] \;\le\; \Xi \cdot \mathbb{E}\left[\sup_{\boldsymbol{N}} \zeta_{\boldsymbol{N}}\right] \;=\; \frac{\nu\,\Xi}{\sqrt{mn}}. \tag{5.11}$$

Moreover, for any fixed values of $\boldsymbol{H}'_1, \ldots, \boldsymbol{H}'_\gamma$, and any $\boldsymbol{N} \in \Gamma$, $\xi_{\boldsymbol{N}}$ is a Ξ-Lipschitz function of the iid Gaussian sequence $(\boldsymbol{H}_1, \ldots, \boldsymbol{H}_\gamma)$.[5] Hence, the supremum $\|\mathcal{D}\boldsymbol{M}\|^*$ is also Ξ-Lipschitz. By Lipschitz concentration ([Led01] Proposition 2.18),

$$\mathbb{P}\left[\, \|\mathcal{D}\boldsymbol{M}\|^* > \mathbb{E}\left[\|\mathcal{D}\boldsymbol{M}\|^* \mid (\boldsymbol{H}'_l)\right] + \frac{t\,\Xi}{\sqrt{mn}} \mid (\boldsymbol{H}'_l) \right] \le \exp\left(-\frac{t^2}{2}\right). \tag{5.12}$$

Combining with our previous estimates, we have

$$\mathbb{P}\left[\, \|\mathcal{D}\boldsymbol{M}\|^* > \frac{\Xi\,(\nu + t)}{\sqrt{mn}} \mid (\boldsymbol{H}'_l) \right] \le \exp\left(-\frac{t^2}{2}\right). \tag{5.13}$$

Since this estimate holds for any value of $(\boldsymbol{H}'_l)$, it holds unconditionally:

$$\mathbb{P}\left[\, \|\mathcal{D}\boldsymbol{M}\|^* > \frac{\Xi\,(\nu + t)}{\sqrt{mn}} \right] \le \exp\left(-\frac{t^2}{2}\right). \tag{5.14}$$

[5] Indeed, if we fix $\boldsymbol{H}'_1 \ldots \boldsymbol{H}'_\gamma$, and consider two sequences $\boldsymbol{H}_1 \ldots \boldsymbol{H}_\gamma$ and $\tilde{\boldsymbol{H}}_1 \ldots \tilde{\boldsymbol{H}}_\gamma$, we have

$$\begin{aligned}
\left|\xi_{\boldsymbol{N}}(\boldsymbol{H}_1 \ldots \boldsymbol{H}_\gamma) - \xi_{\boldsymbol{N}}(\tilde{\boldsymbol{H}}_1 \ldots \tilde{\boldsymbol{H}}_\gamma)\right| &= \left|\sum_{l=1}^{\gamma} \langle \mathcal{P}_S \boldsymbol{H}'_l, \boldsymbol{M}\rangle \left\langle \mathcal{P}_{S^\perp}\boldsymbol{N}, \boldsymbol{H}_l - \tilde{\boldsymbol{H}}_l \right\rangle\right| \\
&\le \left(\sum_l \langle \mathcal{P}_S \boldsymbol{H}'_l, \boldsymbol{M}\rangle^2\right)^{1/2} \left(\sum_l \left\langle \mathcal{P}_{S^\perp}\boldsymbol{N}, \boldsymbol{H}_l - \tilde{\boldsymbol{H}}_l \right\rangle^2\right)^{1/2} \\
&= \Xi\left(\sum_l \left\|\boldsymbol{H}_l - \tilde{\boldsymbol{H}}_l\right\|_F^2\right)^{1/2},
\end{aligned}$$

where in the final line we have used that $\|\mathcal{P}_{S^\perp}\boldsymbol{N}\|_F \le \|\boldsymbol{N}\|_F \le 1$.

Moreover, it is easy to notice that Ξ is itself a $\|\mathcal{P}_S \boldsymbol{M}\|_F$-Lipschitz function of the iid Gaussian sequence $(\boldsymbol{H}'_1, \dots, \boldsymbol{H}'_\gamma)$, with

$$\mathbb{E}[\Xi] \leq \left(\mathbb{E}\left[\Xi^2\right]\right)^{1/2} = \|\mathcal{P}_S \boldsymbol{M}\|_F \sqrt{\frac{\gamma}{mn}}. \tag{5.15}$$

From Lipschitz concentration,

$$\mathbb{P}\left[\Xi > \mathbb{E}[\Xi] + s\|\mathcal{P}_S \boldsymbol{M}\|_F\right] \leq \exp\left(-\frac{s^2 mn}{2}\right), \tag{5.16}$$

and so

$$\mathbb{P}\left[\Xi > 2\|\mathcal{P}_S \boldsymbol{M}\|_F \sqrt{\frac{\gamma}{mn}}\right] \leq \exp\left(-\frac{\gamma}{2}\right). \tag{5.17}$$

Combining this estimate with (5.14), setting $t = \sqrt{20 \log m}$, and rescaling by $\frac{mn}{\gamma}$, we obtain the result. □

6 Proof of Lemma 4.5: Upgrade to Exact Certificate

In this section, we prove Lemma 4.5, which shows how an inexact dual certificate can be upgraded to an exact certificate of optimality. We use the following lemma, whose proof mimics a standard proof of the Gershgorin disc theorem (see [HJ90]):

Lemma 6.1. *Let $T_1, \dots, T_k$ be independent subspaces of $\mathbb{R}^{m \times n}$, and $\boldsymbol{S}_1 \in T_1, \dots, \boldsymbol{S}_k \in T_k$. Then the system of equations*

$$\mathcal{P}_{T_i} \boldsymbol{X} = \boldsymbol{S}_i, \ i = 1, \dots, k, \tag{6.1}$$

has a solution $\boldsymbol{X} \in T_1 + \cdots + T_k$ satisfying

$$\|\boldsymbol{X}\|_F \leq \sqrt{\frac{\sum_i \|\boldsymbol{S}_i\|_F^2}{1 - (k-1)\max_{i \neq j} \|\mathcal{P}_{T_i} \mathcal{P}_{T_j}\|_{F,F}}}. \tag{6.2}$$

Proof. Let $\mathrm{vec} : \mathbb{R}^{m \times n} \to \mathbb{R}^{mn}$ denote the operator that vectorizes a matrix by stacking its columns. For each i, let $\boldsymbol{U}_i \in \mathbb{R}^{mn \times \dim(T_i)}$ denote a matrix whose columns form an orthonormal basis for $\mathrm{vec}[T_i]$. The system of equations is equivalent to

$$\begin{bmatrix} \boldsymbol{U}_1^* \\ \vdots \\ \boldsymbol{U}_k^* \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \boldsymbol{U}_1^* \cdot \mathrm{vec}[\boldsymbol{S}_1] \\ \vdots \\ \boldsymbol{U}_k^* \cdot \mathrm{vec}[\boldsymbol{S}_k] \end{bmatrix}, \tag{6.3}$$

where $\boldsymbol{x} = \mathrm{vec}[\boldsymbol{X}]$. Let $\boldsymbol{U}^*$ denote the matrix on the left hand side, and $\boldsymbol{s}$ denote the vector on the right hand side. Then $\|\boldsymbol{s}\|_2^2 = \sum_{i=1}^k \|\boldsymbol{S}_i\|_F^2$. The system of equations has a solution $\boldsymbol{x} \in \mathrm{range}(\boldsymbol{U}) = \mathrm{vec}[T_1 + \cdots + T_k]$ with ℓ^2 norm at most $\|\boldsymbol{s}\|_2 / \sigma_{\min}(\boldsymbol{U})$, and hence (6.1) has a solution whose Frobenius norm is bounded by the same quantity. Write

$$\boldsymbol{U}^* \boldsymbol{U} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{U}_1^* \boldsymbol{U}_2 & \dots & \boldsymbol{U}_1^* \boldsymbol{U}_k \\ \boldsymbol{U}_2^* \boldsymbol{U}_1 & \boldsymbol{I} & \dots & \boldsymbol{U}_2^* \boldsymbol{U}_k \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{U}_k^* \boldsymbol{U}_1 & \boldsymbol{U}_k^* \boldsymbol{U}_2 & \dots & \boldsymbol{I} \end{bmatrix}. \tag{6.4}$$

Let λ be any eigenvalue of $\boldsymbol{U}^*\boldsymbol{U}$, with corresponding eigenvector $\boldsymbol{x} = (\boldsymbol{x}_1^*,\ \boldsymbol{x}_2^*,\ \ldots, \boldsymbol{x}_k^*)^*$. Let $p = \arg\max_j \|\boldsymbol{x}_j\|_2$. Then, looking at just the p-th block of the equation $\lambda\boldsymbol{x} = \boldsymbol{U}\boldsymbol{U}^*\boldsymbol{x}$, we have

$$\begin{aligned} |\lambda - 1|\|\boldsymbol{x}_p\|_2 &= \Big\|\sum_{j\neq p} \boldsymbol{U}_p^*\boldsymbol{U}_j\boldsymbol{x}_j\Big\|_2 \\ &\leq \sum_{j\neq p} \|\boldsymbol{U}_p^*\boldsymbol{U}_j\|\|\boldsymbol{x}_j\|_2 \\ &\leq \|\boldsymbol{x}_p\|_2 \times (k-1)\max_{i\neq j}\|\boldsymbol{U}_i^*\boldsymbol{U}_j\|. \end{aligned} \tag{6.5}$$

Since $\|\boldsymbol{U}_i^*\boldsymbol{U}_j\| = \|\mathcal{P}_{T_i}\mathcal{P}_{T_j}\|$, we conclude that $\lambda_{\min}(\boldsymbol{U}^*\boldsymbol{U}) \geq 1-(k-1)\max_{i\neq j}\|\mathcal{P}_{T_i}\mathcal{P}_{T_j}\|$, and hence $\sigma_{\min}(\boldsymbol{U})$ is at least as large as the square root of this quantity. This establishes the result. □

Proof of Lemma 4.5. The assumption implies that $T_1, \ldots, T_\tau$ are independent subspaces, and so the system of equations

$$\mathcal{P}_{T_i}\boldsymbol{\Delta} = \lambda_i\boldsymbol{S}_i - \mathcal{P}_{T_i}\hat{\boldsymbol{\Lambda}}, \quad i = 1, \ldots, \tau \tag{6.6}$$

is feasible, and has a solution $\boldsymbol{\Delta}_0 \in T_1 + \cdots + T_\tau$ of Frobenius norm at most

$$\|\boldsymbol{\Delta}_0\|_F \ \leq\ \sqrt{\frac{\sum_i \|\lambda_i\boldsymbol{S}_i - \mathcal{P}_{T_i}\hat{\boldsymbol{\Lambda}}\|_F^2}{1-(\tau-1)\max_{i\neq j}\|\mathcal{P}_{T_i}\mathcal{P}_{T_j}\|_{F,F}}}\ \leq\ \sqrt{\frac{\alpha^2\tau}{1-(\tau-1)\max_{i\neq j}\|\mathcal{P}_{T_i}\mathcal{P}_{T_j}\|_{F,F}}}. \tag{6.7}$$

Moreover, since $T_1 + \cdots + T_\tau$ and $Q^\perp$ are independent, the system of equations

$$\mathcal{P}_{T_1+\cdots+T_\tau}\boldsymbol{\Delta} = \boldsymbol{\Delta}_0,\ \mathcal{P}_{Q^\perp}\boldsymbol{\Delta} = \boldsymbol{0} \tag{6.8}$$

is feasible (indeed, underdetermined). We consider a solution $\boldsymbol{\Delta}_\star$ of minimum Frobenius norm. Under the stated hypotheses, this solution is given by the Neumann series

$$\boldsymbol{\Delta}_\star = \mathcal{P}_Q\sum_{i=0}^{\infty}(\mathcal{P}_{T_1+\cdots+T_\tau}\mathcal{P}_{Q^\perp}\mathcal{P}_{T_1+\cdots+T_\tau})^i\boldsymbol{\Delta}_0, \tag{6.9}$$

whose norm is bounded as

$$\|\boldsymbol{\Delta}_\star\|_F \ \leq\ \frac{\|\boldsymbol{\Delta}_0\|_F}{1-\|\mathcal{P}_{T_1+\cdots+T_\tau}\mathcal{P}_{Q^\perp}\|_{F,F}^2}. \tag{6.10}$$

Set $\boldsymbol{\Lambda} = \hat{\boldsymbol{\Lambda}} + \boldsymbol{\Delta}_\star$, and observe that by construction, for each i, $\mathcal{P}_{T_i}\boldsymbol{\Lambda} = \lambda_i\boldsymbol{S}_i$. For each i, we have

$$\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^* \quad\leq\quad \|\mathcal{P}_{T_i^\perp}\hat{\boldsymbol{\Lambda}}\|_{(i)}^* \ +\ \|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_\star\|_{(i)}^*. \tag{6.11}$$

Because $\|\cdot\|_{(i)}$ majorizes the Frobenius norm, its dual minorizes the Frobenius norm, and so we have

$$\lambda_i^{-1}\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^* \quad\leq\quad \lambda_i^{-1}\|\mathcal{P}_{T_i^\perp}\hat{\boldsymbol{\Lambda}}\|_{(i)}^* \ +\ \lambda_i^{-1}\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_\star\|_F. \tag{6.12}$$

Under the stated hypotheses, this quantity is strictly smaller than one, and so $\boldsymbol{\Lambda}$ satisfies the conditions of Lemma 4.2. □

7 Proof of Theorem 2.1: Compressive PCP Recovery

In this section, we prove Theorem 2.1, using the general upgrade provided by Theorem 4.7. In the language of Section 4, we have $\|\cdot\|_{(1)} = \|\cdot\|_*$, $\|\cdot\|_{(2)} = \|\cdot\|_1$. Both of these norms majorize the Frobenius norm. For PCP, we take $\lambda_1 = 1$, $\lambda_2 = 1/\sqrt{m}$.

Let $\boldsymbol{L}_0, \boldsymbol{S}_0$ denote the target pair, and r the rank of $\boldsymbol{L}_0$. If we let $\boldsymbol{L}_0 = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^*$ denote the rank-reduced singular value decomposition of $\boldsymbol{L}_0$, and T denote the subspace

$$T \doteq \left\{\boldsymbol{U}\boldsymbol{X}^* + \boldsymbol{Y}\boldsymbol{V}^* \mid \boldsymbol{X} \in \mathbb{R}^{n\times r},\ \boldsymbol{Y} \in \mathbb{R}^{m\times r}\right\}, \tag{7.1}$$

then the subdifferential of the nuclear norm at $\boldsymbol{L}_0$ is [Wat92]

$$\partial\|\cdot\|_*(\boldsymbol{L}_0) \ = \ \{\boldsymbol{\Lambda} \mid \mathcal{P}_T\boldsymbol{\Lambda} = \boldsymbol{U}\boldsymbol{V}^*,\ \|\mathcal{P}_{T^\perp}\boldsymbol{\Lambda}\|_{2,2} \le 1\}. \tag{7.2}$$

It is easy to check that $\mathcal{P}_{T^\perp} : \boldsymbol{M} \mapsto (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^*)\boldsymbol{M}(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^*)$ is nonexpansive with respect to the operator norm $\|\cdot\|_{2,2}$, and so $\|\cdot\|_*$ indeed satisfies our criteria for local decomposability. Similarly, let $\Omega = \mathrm{supp}(\boldsymbol{S}_0)$ denote the support of the sparse term. By abuse of notation, we will also identify Ω with the subspace of matrices whose support is contained in Ω. Let $\boldsymbol{\Sigma} = \mathrm{sign}(\boldsymbol{S}_0)$, then

$$\partial\|\cdot\|_1(\boldsymbol{S}_0) \ = \ \{\boldsymbol{\Lambda} \mid \mathcal{P}_\Omega\boldsymbol{\Lambda} = \boldsymbol{\Sigma},\ \|\mathcal{P}_{\Omega^c}\boldsymbol{\Lambda}\|_\infty \le 1\}. \tag{7.3}$$

Again, $\mathcal{P}_{\Omega^c}$ does not increase the ℓ^∞ norm, and $\|\cdot\|_1$ is also decomposable. In the language of Theorem 4.7, we have $T_1 = T$, $\boldsymbol{S}_1 = \boldsymbol{U}\boldsymbol{V}^*$, $T_2 = \Omega$, $\boldsymbol{S}_2 = \boldsymbol{\Sigma}$. For the PCP problem, an (α, β)-inexact certificate is therefore a matrix $\boldsymbol{\Lambda}_{\mathrm{PCP}}$ satisfying

$$\begin{aligned}
\|\mathcal{P}_T\boldsymbol{\Lambda}_{\mathrm{PCP}} - \boldsymbol{U}\boldsymbol{V}^*\|_F &\le \alpha,\\
\|\mathcal{P}_\Omega\boldsymbol{\Lambda}_{\mathrm{PCP}} - \lambda\boldsymbol{\Sigma}\|_F &\le \alpha,\\
\|\mathcal{P}_{T^\perp}\boldsymbol{\Lambda}_{\mathrm{PCP}}\|_{2,2} &\le \beta,\\
\|\mathcal{P}_{\Omega^c}\boldsymbol{\Lambda}_{\mathrm{PCP}}\|_\infty &\le \beta\lambda.
\end{aligned}$$

Such a certificate was constructed in [CLMW11],[6] under the hypotheses of Theorem 2.1. More precisely, we have the following:

Proposition 7.1 (**Dual Certification for PCP [CLMW11]**). *Under the conditions of Theorem 2.1, on an event of probability at least $1 - Cm^{-10}$ the following hold:*

$$\textit{(i)} \qquad \|\mathcal{P}_\Omega\mathcal{P}_T\|_{F,F} \ \le \ 1/2, \tag{7.4}$$

and (ii) there exists a $(m^{-2}, 1/4)$-inexact PCP certificate $\boldsymbol{\Lambda}_{\mathrm{PCP}}$ for $(\boldsymbol{L}_0, \boldsymbol{S}_0)$, which satisfies

$$\|\boldsymbol{\Lambda}_{\mathrm{PCP}}\|_F \ \le \ C'\sqrt{\mathrm{rank}(\boldsymbol{L}_0)} + 2\lambda\sqrt{\|\boldsymbol{S}_0\|_0}. \tag{7.5}$$

Above, C and C' are numerical.

The careful reader may notice that the relaxation parameters (α, β) in Proposition 7.1 are stricter than those provided by [CLMW11], which gives $\alpha = 1/4\sqrt{m}$, $\beta = 1/2$. In fact, by modifying the constants in the construction of [CLMW11], we can achieve β smaller than any desired constant,

[6] In the notation of [CLMW11], the certificate constructed there is $\boldsymbol{\Lambda}_{\mathrm{PCP}} = \boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}^L + \boldsymbol{W}^S$.

and α smaller than any polynomial in m^{-1}, at the expense of slightly more stringent (but qualitatively equivalent) demands on $(\boldsymbol{L}_0, \boldsymbol{S}_0)$. The bound (7.5) is implied by the probabilistic lemmas in [CLMW11], but requires a bit of manipulation to obtain. Below, we will first prove Theorem 2.1, and then sketch a proof of the modifications to [CLMW11] needed to obtain the supporting result Proposition 7.1.

Proof of Theorem 2.1. From Lemma 4.5, to show that $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ is the unique optimal solution to the compressive PCP problem, it is enough to show that

(I) $\|\mathcal{P}_T \mathcal{P}_\Omega\|_{F,F} \leq 1/2$.

(II) There exists an $(\alpha', 1/2)$-inexact CPCP certificate $\boldsymbol{\Lambda}_{\text{CPCP}}$ with $\alpha' < \frac{1-\|\mathcal{P}_{Q^\perp}\mathcal{P}_{T\oplus\Omega}\|_{F,F}^2}{4\sqrt{m}}$.

We accomplish this in three parts. In paragraph (i) below, we apply Lemma 5.1 to lower bound $1 - \|\mathcal{P}_{Q^\perp}\mathcal{P}_{T\oplus\Omega}\|_{F,F}^2$. In paragraph (ii), we use Proposition 7.1 to show (I) and the existence of an inexact PCP certificate $\boldsymbol{\Lambda}_{\text{PCP}}$. In paragraph (iii) we use Theorem 4.7 to upgrade this to an inexact CPCP certificate $\boldsymbol{\Lambda}_{\text{CPCP}}$ that satisfies property (II). Paragraph (iv) completes the proof by showing that the probability of failure is appropriately small.

(i) Bounding $1-\|\mathcal{P}_{Q^\perp}\mathcal{P}_{T\oplus\Omega}\|_{F,F}^2$. We will apply Lemma 5.1 with $S = T+\Omega$. The lemma requires

$$\dim(Q) \;\geq\; C_1 \cdot \dim(T+\Omega).$$

The dimension of $T+\Omega$ is a random variable, which depends on the size of the support set Ω. Let $\mathcal{E}_\Omega$ denote the event

$$\mathcal{E}_\Omega = \{|\Omega| \;\leq\; 2\rho mn + m\}\,. \tag{7.6}$$

Notice that $|\Omega|$ is a sum of mn $\text{Ber}(\rho)$ random variables. By Bernstein's inequality,

$$\mathbb{P}\,[\,|\Omega| \geq \rho mn + t\,] \;\leq\; \exp\left(\frac{-t^2/2}{\rho mn + t/3}\right). \tag{7.7}$$

Setting $t = \rho mn + m$ and simplifying, we obtain

$$\mathbb{P}\,[\,\mathcal{E}_\Omega^c\,] \leq \exp\left(-\frac{3m}{10}\right). \tag{7.8}$$

On $\mathcal{E}_\Omega$, we have

$$\dim(\Omega + T) \;<\; 2\rho mn + m + 2mr \;\leq\; 3 \cdot (\,\rho mn + mr\,). \tag{7.9}$$

Comparing (7.9) to the condition on $\dim(Q)$ in Theorem 2.1, we can see that on $\mathcal{E}_\Omega$, the conditions of Lemma 5.1 are satisfied. Now, let $S = T + \Omega$, set $B = \{\boldsymbol{X} \in S \mid \|\boldsymbol{X}\|_F = 1\}$ and notice that

$$\begin{aligned}
1 - \|\mathcal{P}_{Q^\perp}\mathcal{P}_S\|_{F,F}^2 &= \inf_{\boldsymbol{X}\in B} \;\langle \boldsymbol{X}, \boldsymbol{X}\rangle - \langle \mathcal{P}_{Q^\perp}\mathcal{P}_S\boldsymbol{X}, \mathcal{P}_{Q^\perp}\mathcal{P}_S\boldsymbol{X}\rangle \\
&= \inf_{\boldsymbol{X}\in B} \;\left\langle \boldsymbol{X}, \left(\mathcal{P}_S - \mathcal{P}_S\mathcal{P}_{Q^\perp}\mathcal{P}_S\right)\boldsymbol{X}\right\rangle \\
&= \inf_{\boldsymbol{X}\in B} \;\left\langle \boldsymbol{X}, \left(\frac{\dim(Q)}{mn}\mathcal{P}_S + \mathcal{P}_S\mathcal{P}_Q\mathcal{P}_S - \frac{\dim(Q)}{mn}\mathcal{P}_S\right)\boldsymbol{X}\right\rangle \\
&\geq \frac{\dim(Q)}{mn} - \sup_{\boldsymbol{X}\in B} \left\langle \boldsymbol{X}, \left(\mathcal{P}_S\mathcal{P}_Q\mathcal{P}_S - \frac{\dim(Q)}{mn}\mathcal{P}_S\right)\boldsymbol{X}\right\rangle \\
&\geq \frac{\dim(Q)}{mn} - \left\|\mathcal{P}_S\mathcal{P}_Q\mathcal{P}_S - \frac{\dim(Q)}{mn}\mathcal{P}_S\right\|_{F,F}.
\end{aligned} \tag{7.10}$$

Let $\mathcal{E}_Q$ be the event $\left\{\|\mathcal{P}_S\mathcal{P}_Q\mathcal{P}_S - \frac{\dim(Q)}{mn}\mathcal{P}_S\|_{F,F} \le \frac{1}{16}\frac{\dim(Q)}{mn}\right\}$. Using Lemma 5.1 and $\dim(S) = \dim(T+\Omega) \ge m$, we have

$$\mathbb{P}\left[\,\mathcal{E}_Q \mid \mathcal{E}_\Omega\,\right] \;\ge\; 1 - C_2\exp(-c_1 m). \tag{7.11}$$

On $\mathcal{E}_Q$,

$$1 - \|\mathcal{P}_{Q^\perp}\mathcal{P}_S\|_{F,F}^2 \;\ge\; \frac{15}{16}\frac{\dim(Q)}{mn}.$$

Since by assumption $\dim(Q) \ge C_{\text{meas}} \times \log^2 m \times \dim(T+\Omega) \ge C_{\text{meas}} \times \log^2 m \times m$ and $m \ge n$, ensuring that $C_{\text{meas}} > 16/15$, we can further conclude that

$$1 - \|\mathcal{P}_{Q^\perp}\mathcal{P}_S\|_{F,F}^2 \;\ge\; \frac{1}{m}. \tag{7.12}$$

(ii) Inexact PCP Certificate. By Proposition 7.1, on an event $\mathcal{E}_{\text{PCP}}$ of probability at least $1 - C_2 m^{-10}$, we have $\|\mathcal{P}_T\mathcal{P}_\Omega\|_{F,F} < 1/2$, and there exists an $(m^{-2}, 1/4)$-inexact PCP certificate $\mathbf{\Lambda}_{\text{PCP}}$ for $(\boldsymbol{L}_0, \boldsymbol{S}_0)$, with

$$\|\mathbf{\Lambda}_{\text{PCP}}\|_F \;\le\; C_3\sqrt{\text{rank}(\boldsymbol{L}_0)} + 2\lambda|\Omega|. \tag{7.13}$$

Moreover, $\mathcal{E}_{\text{PCP}}$ is independent of Q. We rewrite the bound (7.13) a bit for later use. We have

$$m\,\|\mathbf{\Lambda}_{\text{PCP}}\|_F^2 \;\le\; m\left(\,2\,C_3\,r + 4\lambda^2|\Omega|^2\,\right),$$

which on $\mathcal{E}_\Omega$ gives

$$m\,\|\mathbf{\Lambda}_{\text{PCP}}\|_F^2 \;\le\; C_4\,(\,\rho mn + mr\,), \tag{7.14}$$

where C_4 is numerical.

(iii) Upgrade to CPCP Certificate. Now, condition on $\mathcal{E}_{\text{PCP}}$ and $\mathcal{E}_\Omega$. By our assumption on $\dim(Q)$ (and ensuring C_{meas} is sufficiently large), the conditions of Theorem 4.7 are satisfied. On an event $\mathcal{E}_{\text{upgrade}}$ of conditional probability at least $1 - C_5 m^{-9}$, the certificate $\mathbf{\Lambda}_{\text{PCP}}$ can be refined to an (α', β')-inexact CPCP certificate $\mathbf{\Lambda}_{\text{CPCP}}$, with

$$\alpha' \le m^{-2} + m^{-3}\|\mathbf{\Lambda}_{\text{PCP}}\|_F, \tag{7.15}$$

and

$$\beta' \;\le\; \frac{1}{4} + C_6\left[\frac{\|\mathbf{\Lambda}_{\text{PCP}}\|_F^2\log m}{\dim(Q)}\right]^{1/2}\max\left\{\mathbb{E}\left[\,\|\boldsymbol{G}\|_{2,2}\right] + \sqrt{\log m}\;,\; \mathbb{E}\left[\,\|\boldsymbol{G}\|_\infty\right]\sqrt{m} + \sqrt{m\log m}\right\},$$

where $\boldsymbol{G}$ is iid $\mathcal{N}(0,1)$. Furthermore, provided $mn > 1$, we have the bounds

$$\mathbb{E}\|\boldsymbol{G}\|_{2,2} \;\le\; 2\sqrt{m} \quad\text{and}\quad \mathbb{E}\|\boldsymbol{G}\|_\infty \;\le\; 3\sqrt{2\log m}. \tag{7.16}$$

The first bound follows from Theorem 2.13 of [DS01], noting that $m \ge n$. The second can be found e.g., on [Ver09] p. 44. So,

$$\beta' \;\;\le\;\; \frac{1}{4} + C_7\left[\frac{m\,\|\mathbf{\Lambda}_{\text{PCP}}\|_F^2\,\log^2 m}{\dim(Q)}\right]^{1/2} \;\;\le\;\; \frac{1}{4} + \left[\frac{C_8(\rho mn + mr)\log^2 m}{\dim(Q)}\right]^{1/2},$$

where we have used (7.14). Ensuring that the constant C_{meas} in the statement of the theorem is larger than $16C_8$, we can conclude that $\beta' \leq 1/2$.

Referring to property (II) above, all that is left to show is that $\alpha' < \frac{1-\|\mathcal{P}_{Q^\perp}\mathcal{P}_{T\oplus\Omega}\|_F^2}{4\sqrt{m}}$. Using paragraph (i), on $\mathcal{E}_\Omega \cap \mathcal{E}_Q$, it suffices to show $\alpha' < \frac{1}{4m^{3/2}}$. Using (7.14), ensuring that the constants $c_{\text{rank}}, c_{\text{sparse}}$ in the statement of Theorem 2.1 are sufficiently small (say, each smaller than $1/2C_4$), we may conclude that $\|\boldsymbol{\Lambda}_{\text{PCP}}\|_F \leq \sqrt{m}$. Hence, we have $\alpha' \leq m^{-2} + m^{-5/2}$, which is strictly smaller than $\frac{1}{4m^{3/2}}$ provided m is sufficiently large.

We have shown that on

$$\mathcal{E}_{\text{good}} \doteq \mathcal{E}_\Omega \cap \mathcal{E}_Q \cap \mathcal{E}_{\text{PCP}} \cap \mathcal{E}_{\text{upgrade}},$$

(I)-(II) hold, and hence $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ is the unique optimal solution to the CPCP problem.

(iv) Probability. We have

$$\begin{aligned}
\mathbb{P}\left[\,\mathcal{E}_{\text{good}}^c\,\right] &\leq \mathbb{P}\left[\,(\mathcal{E}_Q \cap \mathcal{E}_\Omega)^c\,\right] + \mathbb{P}\left[\,(\mathcal{E}_{\text{upgrade}} \cap \mathcal{E}_{\text{PCP}} \cap \mathcal{E}_\Omega)^c\,\right] \\
&= 1 - \mathbb{P}\left[\,\mathcal{E}_Q \mid \mathcal{E}_\Omega\,\right]\mathbb{P}\left[\,\mathcal{E}_\Omega\,\right] + 1 - \mathbb{P}\left[\,\mathcal{E}_{\text{upgrade}} \mid \mathcal{E}_{\text{PCP}} \cap \mathcal{E}_\Omega\,\right]\mathbb{P}\left[\,\mathcal{E}_{\text{PCP}} \cap \mathcal{E}_\Omega\,\right] \\
&\leq 1 - \mathbb{P}\left[\,\mathcal{E}_Q \mid \mathcal{E}_\Omega\,\right] + \mathbb{P}\left[\,\mathcal{E}_\Omega^c\,\right] + 1 - \mathbb{P}\left[\,\mathcal{E}_{\text{upgrade}} \mid \mathcal{E}_{\text{PCP}} \cap \mathcal{E}_\Omega\,\right] + \mathbb{P}\left[\,\mathcal{E}_{\text{PCP}}^c\,\right] + \mathbb{P}\left[\,\mathcal{E}_\Omega^c\,\right] \\
&\leq C_2 \exp(-c_1 m) + C_2 m^{-10} + C_5 m^{-9} + 2\exp(-3m/10),
\end{aligned}$$

provided that m is larger than some m_0. Consolidating bounds, we may conclude that correct recovery occurs with probability at least $1 - C_9 m^{-9}$, choosing C_9 such that the bound is nontrivial only for $m > m_0$. This completes the proof of Theorem 2.1. □

We close by sketching the proof of Proposition 7.1:

Proof of Proposition 7.1 (sketch). Under the hypotheses, the bound $\|\mathcal{P}_\Omega \mathcal{P}_T\| \leq 1/2$ follows immediately from Corollary 2.7 of [CLMW11]. [CLMW11] constructs a certificate $\hat{\boldsymbol{\Lambda}}$ in three parts as

$$\boldsymbol{\Lambda}_{\text{PCP}} = \boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}^L + \boldsymbol{W}^S. \tag{7.17}$$

The two terms $\boldsymbol{W}^L$, $\boldsymbol{W}^S$ will both be elements of $T^\perp$, so

$$\|\mathcal{P}_T \boldsymbol{\Lambda}_{\text{PCP}} - \boldsymbol{U}\boldsymbol{V}^*\|_F = 0. \tag{7.18}$$

Moreover, the term $\boldsymbol{W}^S$ will satisfy $\mathcal{P}_\Omega \boldsymbol{W}^S = \lambda \operatorname{sign}(\boldsymbol{S}_0)$. So, we have

$$\|\mathcal{P}_\Omega \boldsymbol{\Lambda}_{\text{PCP}} - \lambda \operatorname{sign}(\boldsymbol{S}_0)\|_F = \|\mathcal{P}_\Omega[\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}^L]\|_F. \tag{7.19}$$

We can therefore take $\alpha = \|\mathcal{P}_\Omega[\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}^L]\|_F$. To prove Proposition 7.1, it is therefore enough to show that with high probability the following properties are satisfied:

- **(I) Structure constraint:** $\|\mathcal{P}_\Omega[\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}^L]\|_F \leq m^{-2}$.
- **(II) Dual norm constraints:**

$$\|\mathcal{P}_{T^\perp} \boldsymbol{W}^L\|_{2,2} \leq 1/8, \qquad \|\mathcal{P}_{\Omega^c}[\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}^L]\|_\infty \;\leq\; \frac{\lambda}{8}, \tag{7.20}$$

and

$$\|\mathcal{P}_{T^\perp} \boldsymbol{W}^S\|_{2,2} \leq 1/8, \qquad \|\mathcal{P}_\Omega^c \boldsymbol{W}^S\|_\infty \;\leq\; \frac{\lambda}{8}. \tag{7.21}$$

- **(III) Frobenius norm bounds:** We have $\|\mathbf{\Lambda}_{\rm PCP}\|_F \le \|\boldsymbol{U}\boldsymbol{V}^*\|_F + \|\boldsymbol{W}^L\|_F + \|\boldsymbol{W}^S\|_F$. The first term is simply $\sqrt{r}$. We will show that

$$\begin{aligned} \|\boldsymbol{W}^L\|_F &\le 3\sqrt{r}, & (7.22)\\ \|\boldsymbol{W}^S\|_F &\le \frac{4}{3}\lambda\sqrt{\|\boldsymbol{S}_0\|_0}. & (7.23) \end{aligned}$$

In paragraph (i) below, we review the construction of $\boldsymbol{W}^L$ from [CLMW11], and show that with slight changes in the constants, (I) and (7.20) are satisfied with high probability. In paragraph (ii) we review the construction of $\boldsymbol{W}^S$ and show that with high probability (7.21) is satisfied. Together, this implies that $\mathbf{\Lambda}_{\rm PCP}$ is an $(m^{-2}, 1/4)$-certificate for the PCP problem. Paragraph (ii) will also show (7.23). Finally, in paragraph (iii), we show (7.22). This step involves the most additional work. Taken together, this establishes the theorem.

(i) Constructing $\boldsymbol{W}^L$. The term $\boldsymbol{W}^L$ is constructed to lie in $T^\perp$ and satisfy

$$\mathcal{P}_\Omega[\,\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}^L\,] \approx \mathbf{0}.$$

This is accomplished via a golfing argument that writes the complement Ω^c as a union of j_0 subsets $\Upsilon_1, \ldots, \Upsilon_{j_0}$, with $\Upsilon_j \sim_{iid} \mathrm{Ber}(q)$.[7] The parameter q is set so that $\rho = (1-q)^{j_0}$, which ensures that Ω is indeed $\mathrm{Ber}(\rho)$. Notice that with this setting, we have $q \ge (1-\rho)/j_0$.

The certificate is generated inductively, starting with $\boldsymbol{Y}_0 = \mathbf{0}$. The error at step j is

$$\boldsymbol{Z}_j = \mathcal{P}_T \boldsymbol{Y}_j - \boldsymbol{U}\boldsymbol{V}^*$$

and the corrective update

$$\boldsymbol{Y}_j = \boldsymbol{Y}_{j-1} - q^{-1}\mathcal{P}_{\Upsilon_j}\boldsymbol{Z}_{j-1}.$$

This leads to a recursive expression for the error $\boldsymbol{Z}_j$:

$$\boldsymbol{Z}_j = \mathcal{P}_T(\mathcal{I} - q^{-1}\mathcal{P}_{\Upsilon_j})\mathcal{P}_T\boldsymbol{Z}_{j-1}, \tag{7.24}$$

which implies that the error decays quickly in both ℓ^∞ and Frobenius norm:

$$\begin{aligned} \|\boldsymbol{Z}_j\|_\infty &\le \tfrac{1}{2}\|\boldsymbol{Z}_{j-1}\|_\infty \le 2^{-j}\|\boldsymbol{Z}_0\|_\infty,\\ \|\boldsymbol{Z}_j\|_F &\le \tfrac{1}{2}\|\boldsymbol{Z}_{j-1}\|_F \le 2^{-j}\|\boldsymbol{Z}_0\|_F. \end{aligned}$$

These inequalities are [CLMW11] (3.4)-(3.6). Using

$$\|\boldsymbol{Z}_0\|_\infty \le \sqrt{\frac{\mu r}{mn}}, \quad \text{and} \quad \|\boldsymbol{Z}_0\|_F = \|\boldsymbol{U}\boldsymbol{V}^*\|_F = \sqrt{r}, \tag{7.25}$$

we obtain

$$\sum_{j=0}^{j_0}\|\boldsymbol{Z}_j\|_\infty \le 2\sqrt{\frac{\mu r}{mn}}, \quad \text{and} \quad \sum_{j=0}^{j_0}\|\boldsymbol{Z}_j\|_F \le 2\sqrt{r}. \tag{7.26}$$

[7] In [CLMW11], Υ_j is denoted Ω_j; we use the notation Υ to avoid confusion between the support Ω and the subsets Υ of the complement of the support.

From arguments of [CLMW11], these bounds hold simultaneously on an event $\mathcal{E}_Z$ of probability at least $1 - C_1 m^{-10}$. After j_0 steps, the component $\boldsymbol{W}^L$ is generated as

$$\boldsymbol{W}^L \doteq \mathcal{P}_{T^\perp} \boldsymbol{Y}_{j_0} = \mathcal{P}_{T^\perp} \sum_{j=1}^{j_0} q^{-1} \mathcal{P}_{\Upsilon_j} \boldsymbol{Z}_{j-1}. \tag{7.27}$$

The proof of Lemma 2.8(b) of [CLMW11] shows that

$$\|\mathcal{P}_\Omega[\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W}^L]\|_F \quad \leq \quad \|\boldsymbol{Z}_{j_0}\|_F. \tag{7.28}$$

In [CLMW11], j_0 was chosen to ensure that $\|\boldsymbol{Z}_{j_0}\|_F \leq 1/4\sqrt{m}$. Here, we set $j_0 = \lceil 3 \log_2 m \rceil$, ensuring that $\|\boldsymbol{Z}_{j_0}\|_F \leq 2^{-m}\sqrt{r} \leq m^{-2}$. The arguments of [CLMW11] establish the following:

$$\|\mathcal{P}_{T^\perp} \boldsymbol{W}^L\|_{2,2} \leq 2C_0' \sqrt{\frac{m \log m}{q}} \|\boldsymbol{U}\boldsymbol{V}^*\|_\infty, \tag{7.29}$$

$$\|\mathcal{P}_{\Omega^c} \boldsymbol{W}^L\|_\infty \leq \|\boldsymbol{Z}_{j_0}\|_\infty + 2q^{-1}\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty. \tag{7.30}$$

where C_0' is numerical. Moreover, we know that $q > \frac{c}{\log m}$, where $c > 0$ is numerical. Hence, we have

$$\|\mathcal{P}_{T^\perp} \boldsymbol{W}^L\|_{2,2} \leq C \sqrt{\frac{\mu r \log^2 m}{n}}, \tag{7.31}$$

$$\|\mathcal{P}_{\Omega^c} \boldsymbol{W}^L\|_\infty \leq \frac{1}{\sqrt{m}} \left(\sqrt{\frac{\mu r}{n}} + \frac{2}{c} \sqrt{\frac{\mu r \log^2 m}{n}} \right). \tag{7.32}$$

Recalling that by assumption $r \leq c_{\text{rank}} n / \mu \log^2 m$, we have the desired bounds (7.20), provided the constant c_{rank} is sufficiently small.

(ii) Constructing $\boldsymbol{W}^S$. The term $\boldsymbol{W}^S$ is constructed to satisfy $\boldsymbol{W}^S \in T^\perp$, $\mathcal{P}_\Omega \boldsymbol{W}^S = \lambda \operatorname{sign}(\boldsymbol{S}_0)$. More precisely, [CLMW11] set this term to be the minimum Frobenius norm solution to the system of equations

$$\mathcal{P}_T \boldsymbol{W}^S = \boldsymbol{0}, \quad \mathcal{P}_\Omega \boldsymbol{W}^S = \lambda \operatorname{sign}(\boldsymbol{S}_0). \tag{7.33}$$

This solution is given by the Neumann series

$$\boldsymbol{W}^S = \mathcal{P}_{T^\perp} \sum_{j=0}^{\infty} (\mathcal{P}_\Omega \mathcal{P}_T \mathcal{P}_\Omega)^j [\lambda \operatorname{sign}(\boldsymbol{S}_0)]. \tag{7.34}$$

So,

$$\|\boldsymbol{W}^S\|_F \leq \frac{\|\lambda \operatorname{sign}(\boldsymbol{S}_0)\|_F}{1 - \|\mathcal{P}_\Omega \mathcal{P}_T\|_{F,F}^2} \leq \tfrac{4}{3} \lambda \sqrt{\|\boldsymbol{S}_0\|_0}. \tag{7.35}$$

Similar to paragraph (i) above, one can quickly check that by ensuring ρ is smaller than some fixed constant, (7.21) is satisfied with high probability.

(iii) Bounding $\|\boldsymbol{W}^L\|_F$. We use the fact that Υ_j and $\boldsymbol{Z}_{j-1}$ are independent random variables. By (7.27), it is enough to control the Frobenius norm $q^{-1}\mathcal{P}_{\Upsilon_j}\boldsymbol{Z}_{j-1}$ for each j. Notice that

$$\|q^{-1}\mathcal{P}_{\Upsilon_j}\boldsymbol{Z}_{j-1}\|_F^2 \;=\; \|\boldsymbol{Z}_{j-1}\|_F^2 + \sum_{kl}(q^{-1}\delta_{kl}-1)[\boldsymbol{Z}_{j-1}]_{kl}^2 \;\doteq\; \|\boldsymbol{Z}_{j-1}\|_F^2 + \sum_{kl} H_{kl},$$

where δ_{kl} is an indicator for the event $(k,l)\in\Upsilon_j$. Then $\mathbb{E}[H_{kl}]=0$, $|H_{kl}|\le q^{-1}\|\boldsymbol{Z}_{j-1}\|_\infty^2$ almost surely, and $\mathbb{E}[H_{kl}^2]\le q^{-1}[\boldsymbol{Z}_{j-1}]_{kl}^4$. Summing, we have

$$\sum_{kl}\mathbb{E}[H_{kl}^2] \;\le\; q^{-1}\|\boldsymbol{Z}_{j-1}\|_\infty^2\|\boldsymbol{Z}_{j-1}\|_F^2. \tag{7.36}$$

By Bernstein's inequality,

$$\mathbb{P}\left[\,\|q^{-1}\mathcal{P}_{\Upsilon_j}\boldsymbol{Z}_{j-1}\|_F^2 \;>\; \|\boldsymbol{Z}_{j-1}\|_F^2 + t\,\right] \;\le\; \exp\left(-\frac{t^2}{2q^{-1}\|\boldsymbol{Z}_{j-1}\|_\infty^2\left(\|\boldsymbol{Z}_{j-1}\|_F^2+\frac{t}{3}\right)}\right).$$

By setting

$$t_j = C_2\max\left\{\|\boldsymbol{Z}_{j-1}\|_\infty^2 q^{-1}\log m\,,\; \|\boldsymbol{Z}_{j-1}\|_\infty\|\boldsymbol{Z}_{j-1}\|_F\sqrt{q^{-1}\log m}\right\}, \tag{7.37}$$

with appropriate numerical constant C_2, we can ensure that for each j,

$$\mathbb{P}\left[\,\|q^{-1}\mathcal{P}_{\Upsilon_j}\boldsymbol{Z}_{j-1}\|_F^2 \;>\; \|\boldsymbol{Z}_{j-1}\|_F^2 + t_j\,\right] \;\le\; m^{-11}. \tag{7.38}$$

Since we have $q > \frac{c}{\log m}$ for some positive numerical constant c, using $\sqrt{s+t}\le\sqrt{s}+\sqrt{t}$, on an event with overall probability at least $1-j_0m^{-11}$,

$$\begin{aligned}
\|\boldsymbol{W}^L\|_F &\le \sum_{j=1}^{j_0}\|q^{-1}\mathcal{P}_{\Upsilon_j}\boldsymbol{Z}_{j-1}\|_F \\
&\le \sum_{j=1}^{j_0}\|\boldsymbol{Z}_{j-1}\|_F + \sqrt{t_j} \\
&\le \sum_{j=1}^{j_0}\|\boldsymbol{Z}_{j-1}\|_F + C_3\|\boldsymbol{Z}_{j-1}\|_\infty\log m + C_4\sqrt{\|\boldsymbol{Z}_{j-1}\|_\infty\|\boldsymbol{Z}_{j-1}\|_F\log m} \\
&\le 2\sqrt{r} + 2C_3\log(m)\sqrt{\frac{\mu r}{mn}} + C_4\sqrt{\log m}\times\sum_{j=1}^{j_0}2^{-j}\|\boldsymbol{Z}_0\|_\infty^{1/2}\|\boldsymbol{Z}_0\|_F^{1/2} \\
&\le 2\sqrt{r} + 2C_3\sqrt{\frac{1}{m}\frac{\mu r\log^2 m}{n}} + 2C_4\sqrt[4]{\frac{r}{m}\frac{\mu r\log^2 m}{n}}.
\end{aligned}$$

Recalling again the assumption $r\le c_{\rm rank}n/\mu\log^2 m$, and ensuring that $c_{\rm rank}$ is sufficiently small, the final two terms above are bounded by constants. In particular, we can conclude that $\|\boldsymbol{W}^L\|_F\le 3\sqrt{r}$. This completes the proof. □

8 Numerical Experiments

In this section, we corroborate our theoretical results with simulations on both synthetic data and real images. Before proceeding to the simulation results, we provide a brief description of the optimization algorithm below.

8.1 Efficient Algorithm

Algorithm 1 TFOCS-inspired solution for (1.7)

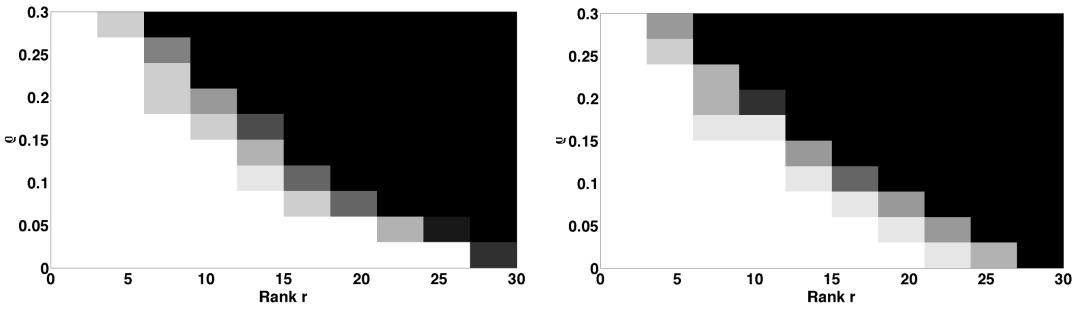
while not converged ($j = 1, 2, \ldots$) **do**
  Initialize $\boldsymbol{R}_1 \leftarrow \boldsymbol{\Theta}_1$, $\xi_1 \leftarrow 1$
  while not converged ($k = 1, 2, \ldots$) **do**
    $\boldsymbol{M}_k \leftarrow (1 - \xi_k)\boldsymbol{R}_k + \xi_k \boldsymbol{\Theta}_k$
    $(\boldsymbol{U}_k, \boldsymbol{\Sigma}_k, \boldsymbol{V}_k) \leftarrow \text{svd}\left(\boldsymbol{L}_j^{(0)} + \mu_j^{-1}\mathcal{P}_Q \boldsymbol{M}_k\right)$
    $\boldsymbol{L}_{k+1} \leftarrow \boldsymbol{U}_k \cdot \text{shrink}\left(\boldsymbol{\Sigma}_k, \mu_j^{-1}\right) \cdot \boldsymbol{V}_k^*$
    $\boldsymbol{S}_{k+1} \leftarrow \text{shrink}\left(\boldsymbol{S}_j^{(0)} + \mu_j^{-1}\mathcal{P}_Q \boldsymbol{M}_k, \lambda\mu_j^{-1}\right)$
    $\boldsymbol{\Theta}_{k+1} \leftarrow \boldsymbol{\Theta}_k + \frac{\mu_j}{2}(\boldsymbol{Y} - \mathcal{P}_Q(\boldsymbol{L}_{k+1} + \boldsymbol{S}_{k+1}))$
    $\boldsymbol{R}_{k+1} \leftarrow (1 - \xi_k)\boldsymbol{R}_k + \xi_k \boldsymbol{\Theta}_{k+1}$
    $\xi_{k+1} \leftarrow 2/(1 + \sqrt{1 + 4/\xi_k^2})$
  end while
  (Let $(\boldsymbol{L}^\dagger, \boldsymbol{S}^\dagger)$ be the converged solution)
  $\left(\boldsymbol{L}_{j+1}^{(0)}, \boldsymbol{S}_{j+1}^{(0)}\right) \leftarrow (\boldsymbol{L}^\dagger, \boldsymbol{S}^\dagger)$
  $\mu_{j+1} \leftarrow \gamma \cdot \mu_j$
end while

The problem (1.7) is a convex program, and can be solved by interior-point methods. These algorithms are known to have excellent iteration complexity and are available as part of public software packages. However, the main disadvantage of these methods is that they do not scale very well with problem size. Fortunately, there has been a flurry of work recently on fast first-order methods for nuclear-norm minimization. In this work, we adopt the framework described in [BCG10], called TFOCS, to solve (1.7) efficiently. The main idea is to apply an optimal gradient method to a smoothed version of the dual of problem (1.7). We provide a brief description of the algorithm here.

The objective function in (1.7) is continuous but not smooth. Hence, the first step is to modify it by adding a smooth regularization term as follows:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 + \frac{\mu}{2}(\|\boldsymbol{L} - \boldsymbol{L}^{(0)}\|_F^2 + \|\boldsymbol{S} - \boldsymbol{S}^{(0)}\|_F^2) \quad \text{subj. to} \quad \boldsymbol{Y} = \mathcal{P}_Q(\boldsymbol{L} + \boldsymbol{S}), \tag{8.1}$$

where $\boldsymbol{L}^{(0)}$ and $\boldsymbol{S}^{(0)}$ are fixed, and $\mu > 0$. While (8.1) remains nonsmooth, this modification has the effect of smoothing the Lagrange dual function. We can then apply fast gradient methods to the smoothed dual problem. If μ is very small, then the solution to (8.1) is close to that of the original problem (1.7). In fact, recent results suggest that these solutions may coincide exactly for small (nonzero) values of μ [ZCZ12].



(a) $p = 0$ (b) $p = 500$

(c) $p = 2,000$ (d) $p = 5,000$

Figure 2: **Compressive sensing of low-rank and sparse matrices via convex program** (1.7)**.** We use the TFOCS algorithm [BCG10] to solve the convex program for recovering an $m \times m$ matrix $\boldsymbol{M} = \boldsymbol{L}_0 + \boldsymbol{S}_0$ with $m = 100$ from $q = m^2 - p$ random linear measurements $\mathcal{P}_Q[\boldsymbol{M}]$. For each subplot: the x-axis is the rank r of the matrix $\boldsymbol{L}_0$ and the y-axis is the percentage of non-zero entries in $\boldsymbol{S}_0$. The intensity is proportional to the probability of success with pure white color meaning 100% (out of 10 random trials).

Let us now focus on the smoothed problem. The Lagrangian of this problem is given by

$$\mathcal{L}(\boldsymbol{L}, \boldsymbol{S}, \boldsymbol{\Theta}) = \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 + \frac{\mu}{2}(\|\boldsymbol{L} - \boldsymbol{L}^{(0)}\|_F^2 + \|\boldsymbol{S} - \boldsymbol{S}^{(0)}\|_F^2) - \langle \boldsymbol{\Theta}, \mathcal{P}_Q(\boldsymbol{L} + \boldsymbol{S}) - \boldsymbol{Y} \rangle, \tag{8.2}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{m \times n}$ is a matrix of Lagrange multipliers. The dual function can then be defined as

$$g(\boldsymbol{\Theta}) = \inf_{\boldsymbol{L}, \boldsymbol{S}} \mathcal{L}(\boldsymbol{L}, \boldsymbol{S}, \boldsymbol{\Theta}). \tag{8.3}$$

The dual function has some attractive properties that can be exploited for optimization. We note that $g(\cdot)$ is concave in $\boldsymbol{\Theta}$ and its gradient is given by

$$\nabla g(\boldsymbol{\Theta}) = \boldsymbol{Y} - \mathcal{P}_Q(\hat{\boldsymbol{L}}(\boldsymbol{\Theta}) + \hat{\boldsymbol{S}}(\boldsymbol{\Theta})),$$

where $(\hat{\boldsymbol{L}}(\boldsymbol{\Theta}), \hat{\boldsymbol{S}}(\boldsymbol{\Theta}))$ is the unique optimal solution to the problem (8.3) (with $\boldsymbol{\Theta}$ fixed). This pair is given in closed form by

$$\begin{aligned} \hat{\boldsymbol{L}}(\boldsymbol{\Theta}) &= \boldsymbol{U}' \cdot \text{shrink}\left(\boldsymbol{\Sigma}', \mu^{-1}\right) \cdot \boldsymbol{V}'^{*}, \\ \hat{\boldsymbol{S}}(\boldsymbol{\Theta}) &= \text{shrink}\left(\boldsymbol{S}^{(0)} + \mu^{-1}\mathcal{P}_Q\boldsymbol{\Theta}, \lambda\mu^{-1}\right), \end{aligned} \tag{8.4}$$

where $\boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^*$ is the reduced Singular Value Decomposition (SVD) of $\boldsymbol{L}^{(0)} + \mu^{-1}\mathcal{P}_Q\boldsymbol{\Theta}$, and the *shrinkage* operator, denoted by $\text{shrink}(\cdot,\cdot)$, is defined as follows:

$$\text{shrink}(x, \alpha) = \text{sign}(x) \cdot \max\{|x| - \alpha, 0\},$$

where $x \in \mathbb{R}$ and $\alpha \geq 0$. The shrinkage operator is extended to vectors and matrices by applying it elementwise.

Let $\hat{\boldsymbol{\Theta}}$ be the point of maxima of $g(\cdot)$. Then, it is possible to compute $\hat{\boldsymbol{\Theta}}$ by a simple gradient ascent iteration as follows:

$$\begin{aligned}
\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^* &= \text{svd}(\boldsymbol{L}^{(0)} + \mu^{-1}\mathcal{P}_Q\boldsymbol{\Theta}_k), \\
\boldsymbol{L}_{k+1} &= \boldsymbol{U}_k \cdot \text{shrink}\left(\boldsymbol{\Sigma}_k, \mu^{-1}\right) \cdot \boldsymbol{V}_k^*, \\
\boldsymbol{S}_{k+1} &= \text{shrink}\left(\boldsymbol{S}^{(0)} + \mu^{-1}\mathcal{P}_Q\boldsymbol{\Theta}_k, \lambda\mu^{-1}\right), \\
\boldsymbol{\Theta}_{k+1} &= \boldsymbol{\Theta}_k + t_k(\boldsymbol{Y} - \mathcal{P}_Q(\boldsymbol{L}_{k+1} + \boldsymbol{S}_{k+1})),
\end{aligned}$$

where $t_k > 0$ is the step size satisfying $t_k \leq \mu/2$. Once we obtain the dual optimal solution $\hat{\boldsymbol{\Theta}}$, the primal optimal solution can be computed using (8.4). While this iteration provably converges to an optimal solution to the dual problem (and hence yields a solution to the primal problem), faster convergence can be obtained by replacing simple gradient ascent method with an accelerated gradient method, such as Nesterov's optimal method [Nes83]. This modification yields Algorithm 1.

The above iterative scheme has two free parameters, μ and $(\boldsymbol{L}^{(0)}, \boldsymbol{S}^{(0)})$. While a small μ is desirable for accuracy, it leads to slower convergence in practice. This is particularly true if $\boldsymbol{L}^{(0)}$ and $\boldsymbol{S}^{(0)}$ are far from the optimal solution. Thus, we follow a continuation strategy suggested in [BCG10], in which we successively solve (8.1) while monotonically decreasing the value of μ. The optimal solution from one complete iteration is used to initialize the subsequent one. The entire algorithm is summarized as Algorithm 1. For our experiments, we initialize $(\boldsymbol{L}_1^{(0)}, \boldsymbol{S}_1^{(0)})$ to $(\boldsymbol{Y}, \boldsymbol{0})$. We also fix $\mu_1 = 100$ and $\gamma = 0.9$. We terminate the outer loop of this algorithm when the relative change in $\boldsymbol{L}^{(0)}, \boldsymbol{S}^{(0)}$ is sufficiently small:

$$\left\|\boldsymbol{L}_{j+1}^{(0)} - \boldsymbol{L}_j^{(0)}\right\|_F^2 + \left\|\boldsymbol{S}_{j+1}^{(0)} - \boldsymbol{S}_j^{(0)}\right\|_F^2 \;<\; \varepsilon^2 \times \left(\left\|\boldsymbol{L}_j^{(0)}\right\|_F^2 + \left\|\boldsymbol{S}_j^{(0)}\right\|_F^2\right).$$

In our numerical experiments, we set $\varepsilon = 0.001$.

The major computational burdens in Algorithm 1 are (a) computing the partial singular value decomposition and (b) applying the projection operator $\mathcal{P}_Q$. The cost of the singular value decomposition can be ameliorated using rank prediction and partial singular value decomposition. This heuristic can dramatically reduce the computational complexity of practical implementations. We should note, however, that we are currently not aware of any algorithm that provably solves (1.7) (or its close variants) and provably controls the rank of the iterates. As noted above, the other main cost is that of storing and applying the operator $\mathcal{P}_Q$. For imaging applications, this operator is often either almost full rank, or has special structure (such as acting column-wise) that renders its contribution to the computational complexity negligible compared to the SVD. However, when Q is unstructured (e.g., an isotropic random subspace) the costs of representing it in memory and applying it to matrices are both proportional to $mn \cdot \dim(Q)$. Reducing this burden using structured random projections is another interesting direction for future work.

8.2 Simulation Results

We test our algorithm on some randomly generated data to corroborate our theoretical results. Given n, r, k and p, we generate the data as follows:

- $\boldsymbol{L}_0 = \boldsymbol{U}\boldsymbol{V}^*$, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n \times r}$ are random matrices each of whose entries is i.i.d. according to the standard normal distribution.
- $\boldsymbol{S}_0 \in \mathbb{R}^{n \times n}$ is a sparse matrix with exactly k non-zero entries whose support is uniformly distributed among all possible sets of cardinality k and whose non-zero entries are i.i.d. uniformly in the range $[-10, 10]$.
- Q is generated as the linear subspace spanned by p independent random matrices $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_p \in \mathbb{R}^{n \times n}$. Each of the $\boldsymbol{G}_i$'s have entries that are i.i.d. according to $\mathcal{N}(0, 1/n^2)$.

We fix $n = 100$ in all our simulations. To observe the recovery behavior as a function of r, k and p, we plot our results in the following fashion. We fix p and subsequently the subspace Q. Then, we vary r and k, and empirically observe the probability of recovery success. For each pair (r, k), we carry out 10 independent trials. A trial is considered successful if

$$\frac{\|\hat{\boldsymbol{L}} - \boldsymbol{L}_0\|_F}{\|\boldsymbol{L}_0\|_F} \leq 10^{-2},$$

where $\hat{\boldsymbol{L}}$ is the recovered low-rank component.

The simulation results are summarized in a series of plots in Figure 2. The case $p = 0$ is basically the PCP scenario when all the entries of the corrupted matrix are directly observed. We note that as p increases, the region of reliable recovery shrinks. Notice that when $p = 5,000$, the number of linear measurements is only half of the number of entries and there remains a small region where the convex program succeeds.

8.3 Removing shadows from faces images

We next show the result of applying Algorithm 1 to a dataset of real faces images. It is known from [CLMW11] that faces images can often be decomposed as the low-rank part, capturing the face appearances under different illuminations, and a sparse error part, representing shadows and specularities. In this task, we prepare the face images as in [CLMW11], with one key difference: only the matrix $\boldsymbol{D} = \mathcal{P}_Q[\boldsymbol{M}]$ is observed, where $\boldsymbol{M}$ is the corrupted face data matrix, and Q is a subspace we generate randomly.

For $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, we let the dimension of the random subspace $\dim(Q) = 0.75mn$, reducing 25% measurements of the original face data. Nevertheless, from Figure 3 we can see that both the low-rank and sparse parts are successfully recovered from the reduced measurements $\boldsymbol{D}$. The main reason we can only discard 25% of the measurements in this case is that n is relatively small (20 images in this example). For larger n more significant reductions are possible.

Acknowledgements

JW thanks the Rice group (Andrew Waters, Aswin Sankaranarayanan and Richard Baraniuk) for discussions and clarifications related to this work and [WSB11]. He would also like to thank Xiaodong

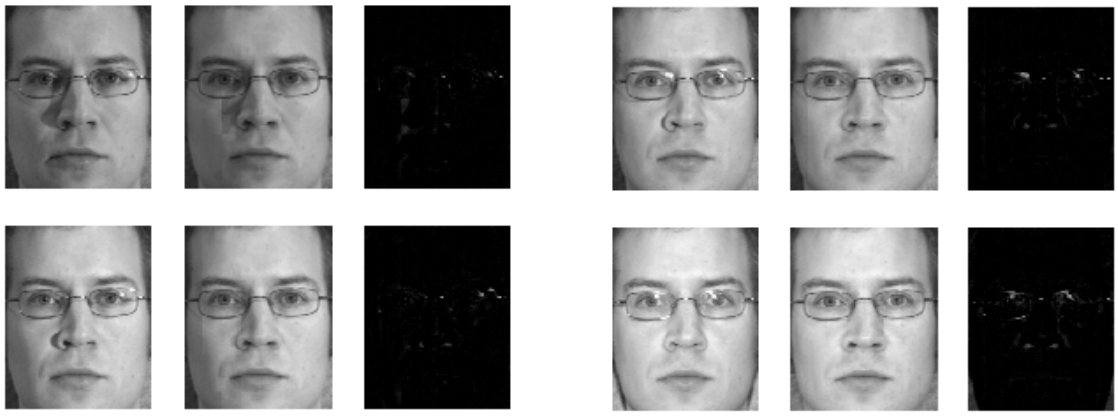


Figure 3: **Removing shadows from faces images.** The first column corresponds to the input images. The second column $\boldsymbol{L}$ is obtain by solving our convex optimization (1.7). The last column $\boldsymbol{S}$ corresponds to the shadows and specularities obtained by the Compressive PCP decomposition.

Li of Stanford for discussions related to this work. Finally, the authors thank the anonymous reviwers for their detailed feedback on the paper. This work was partially supported by the funding of ONR N00014-09-1-0230, NSF CCF 09-64215, NSF IIS 11-16012, and DARPA KECoM 10036-100471. JW was also supported by startup funding from Columbia University.

References

[ANW11] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *preprint*, 2011.

[Bac10] F. Bach. Structured sparsity-inducing norms through submodular functions. *preprint*, 2010.

[BCG10] S. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. preprint, 2010.

[BJ03] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

[BSE07] J. Bobin, Y. Moudden J. L. Starck, and M. Elad. Sparsity and morphological diversity and source separation. *IEEE Transactions on Image Processing*, 16(11):2662–2674, 2007.

[CJSC13] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 2013.

[CLMW11] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(7), May 2011.

[CR08] E. Candès and B. Recht. Exact matrix completion via convex optimzation. *Foundations of Computational Mathematics*, 9(6):717–772, 2008.

[CR11] E. Candès and B. Recht. Simple bounds for low-complexity model reconstruction. *preprint*, 2011.

[CRPW10] V. Chandrasekaran, B. Recht, P. Parillo, and A. Wilsky. The convex geometry of linear inverse problems. *preprint*, 2010.

[CSPW11] V. Chandrasekaran, S. Sanghavi, P. Parillo, and A. Wilsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[CT05] E. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Information Theory*, 51(12), 2005.

[DG03] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.

[Don06] D. Donoho. Compressive sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[DS01] K. Davidson and S. Szarek. Local operator theory, random matrices and banach spaces. In *Handbook of the Geometry of Banach Spaces, Volume 1*, 2001.

[GLW+10] A. Ganesh, X. Li, J. Wright, E. Candes, and Y. Ma. Dense error correction for low-rank matrices via principal component pursuit. In *Proc. of International Symposium on Information Theory*, 2010.

[GMWM12] A. Ganesh, K. Min, J. Wright, and Y. Ma. Principal component pursuit with reduced linear measurements. Available at `http://arxiv.org/pdf/1202.6445`, 2012.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3), 2011.

[GV11] M. Golbabaee and P. Vandergheynst. Hyperspectral image compressed sensing via low-rank and joint sparse matrix recovery. In *Indernational Conference on Acoustics, Speech and Signal Processing ICASSP*, 2011.

[HJ90] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[HKZ11] D. Hsu, S. Kakade, and T. Zhang. Robust matrix decomposition with outliers. *IEEE Transactions on Information Theory*, 57:7221–7234, 2011.

[Led01] M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

[Li11] X. Li. Compressed sensing and matrix completion with constant proportion of corruptions. Available at `http://arxiv.org/abs/1104.1041`, 2011.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic function by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

[MT11] M. McCoy and J. Tropp. Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics*, 2011.

[Nes83] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[NRWY10] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for analyzing m-estimators with decomposible regularizers. Available at `http://arxiv.org/abs/1010.2731v1`, 2010.

[NT08] D. Needell and J. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comp. Harmonic Anal.*, 26:301–321, 2008.

[PGW+12] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2233–2246, 2012.

[Rec11] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[SA11] X. Shu and N. Ahuja. Imaging via three-dimensional compressive sampling (3DCS). In *International Conference on Computer Vision ICCV*, 2011.

[Ver09] R. Vershynin. Lectures on geometric functional analysis. *http://www-personal.umich.edu/~romanv/papers/GFA-book/GFA-book.pdf*, 2009.

[Ver11] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2011. Available at `http://www-personal.umich.edu/ romanv/papers/non-asymptotic-rmt-plain.pdf`.

[Wat92] G. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Applications*, 170:33–45, 1992.

[WGS+10] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Asian Conference on Computer Vision*, 2010.

[WSB11] A. Waters, A. Sankaranarayanan, and R. Baraniuk. Sparcs: Recovering low-rank and sparse matrices from compressive measurements. In *Proc. Neural Information Processing Systems (NIPS)*, 2011.

[XSC11] H. Xu, S. Sanghavi, and C. Caramanis. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 2011.

[ZCZ12] H. Zhang, J. Cai, and J. Zhu. Strongly convex programming for exact matrix completion and robust principal component analysis. *preprint*, 2012.

[ZGLM12] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma. Tilt: Transform invariant low-rank textures. *International Journal of Computer Vision*, 99:1–24, 2012.

[ZLW+10] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma. Stable principal component pursuit. In *Proc. of International Symposium on Information Theory*, 2010.

A Decomposable Norms

In this appendix, we briefly compare the notion of local decomposability in Definition 4.1 (and [CR11]) to the notion of decompsability in [NRWY10]. In the sense of [NRWY10], a norm $\|\cdot\|$ is decomposable over a subspace pair $(T, T^\perp)$ if for all $\boldsymbol{x} \in T$, $\boldsymbol{y} \in T^\perp$, $\|\boldsymbol{x}+\boldsymbol{y}\| = \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$. If $\boldsymbol{x} \in T$, and $\|\cdot\|$ is decomposable over $(T, T^\perp)$ in the sense of [NRWY10], and the restriction of $\|\cdot\|$ to T is differentiable at $\boldsymbol{x}$, then $\|\cdot\|$ is locally decomposable at $\boldsymbol{x}$. However, there exist norms that are decomposable in the sense of [NRWY10], but are not locally decomposable. For example, for $\boldsymbol{x} \in \mathbb{R}^4$, let

$$\Gamma = \{(s,t,0,0) \mid s,t \in \mathbb{R}\},$$

and

$$\|\boldsymbol{x}\|_\diamondsuit = \|P_\Gamma \boldsymbol{x}\|_\infty + \|P_{\Gamma^\perp}\boldsymbol{x}\|_\infty .$$

In the sense of [NRWY10], this norm is evidently decomposable over $T = \Gamma$, $T^\perp = \Gamma^\perp$. At the point $\boldsymbol{x} = (1,1,0,0)$, the subdifferential is

$$\partial \|\cdot\|_\diamondsuit (\boldsymbol{x}) = \{(a,b,c,d) \mid a \geq 0, b \geq 0, a+b=1, |c|+|d| \leq 1\} .$$

Since the projection of the subdifferential onto T is not a singleton, this norm is not locally decomposable at $\boldsymbol{x}$.

On the other hand, if a norm $\|\cdot\|$ is locally decomposable at $\boldsymbol{x}$, it is easy to show that for any $\boldsymbol{y} \in T^\perp$, we have

$$\|\boldsymbol{x}+\boldsymbol{y}\| = \|\boldsymbol{x}\| + \|\boldsymbol{y}\| .$$

So, if the norm is locally decomposable at every $\boldsymbol{x} \in T$, it is decomposable over $(T, T^\perp)$, in the sense of [NRWY10].

B Proof of Lemma 4.2: Optimality Conditions

Proof. Let f denote the objective function. Consider a feasible perturbation $\boldsymbol{\delta} = (\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_\tau)$, so $\mathcal{P}_Q \sum_i \boldsymbol{\Delta}_i = \boldsymbol{0}$. Then for any $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_\tau$ such that $\forall\, i,\ \boldsymbol{W}_i \in \partial \|\cdot\|_{(i)}(\boldsymbol{X}_{i,\star})$, we have

$$f(\boldsymbol{x}_\star + \boldsymbol{\delta}) \ \geq\ f(\boldsymbol{x}_\star) + \sum_i \lambda_i \langle \boldsymbol{W}_i, \boldsymbol{\Delta}_i \rangle. \tag{B.1}$$

By duality of norms, for each i there exists $\boldsymbol{H}_i \in \mathbb{R}^{m\times n}$ with $\|\boldsymbol{H}_i\|_{(i)}^* \leq 1$ and

$$\langle \boldsymbol{H}_i, \mathcal{P}_{T_i^\perp} \boldsymbol{\Delta}_i \rangle \ =\ \|\mathcal{P}_{T_i^\perp} \boldsymbol{\Delta}_i\|_{(i)}. \tag{B.2}$$

Set $\boldsymbol{W}_i = \boldsymbol{S}_i + \mathcal{P}_{T_i^\perp}\boldsymbol{H}_i$. From our definition of a locally decomposable norm, $\mathcal{P}_{T_i^\perp}$ is nonexpansive, and so $\|\mathcal{P}_{T_i^\perp}\boldsymbol{H}_i\|_{(i)}^* \leq 1$, and $\boldsymbol{W}_i \in \partial\|\cdot\|_{(i)}(\boldsymbol{X}_{i,\star})$. Moreover,

$$\begin{aligned}
\langle \boldsymbol{W}_i, \boldsymbol{\Delta}_i\rangle &= \langle \mathcal{P}_{T_i}\boldsymbol{W}_i, \boldsymbol{\Delta}_i\rangle + \langle \mathcal{P}_{T_i^\perp}\boldsymbol{W}_i, \boldsymbol{\Delta}_i\rangle \\
&= \langle \mathcal{P}_{T_i}\boldsymbol{W}_i, \mathcal{P}_{T_i}\boldsymbol{\Delta}_i\rangle + \langle \mathcal{P}_{T_i^\perp}\boldsymbol{W}_i, \mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\rangle \\
&= \langle \boldsymbol{S}_i, \mathcal{P}_{T_i}\boldsymbol{\Delta}_i\rangle + \langle \boldsymbol{H}_i, \mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\rangle \\
&= \langle \boldsymbol{S}_i, \mathcal{P}_{T_i}\boldsymbol{\Delta}_i\rangle + \|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)} \qquad \text{(B.3)}
\end{aligned}$$

Plugging in to (B.1), we have

$$\begin{aligned}
f(\boldsymbol{x}_\star + \boldsymbol{\delta}) &\geq f(\boldsymbol{x}_\star) + \sum_i \langle \lambda_i \boldsymbol{S}_i, \mathcal{P}_{T_i}\boldsymbol{\Delta}_i\rangle + \lambda_i\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)} \\
&= f(\boldsymbol{x}_\star) + \sum_i \langle \mathcal{P}_{T_i}\boldsymbol{\Lambda}, \mathcal{P}_{T_i}\boldsymbol{\Delta}_i\rangle + \lambda_i\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)} \\
&= f(\boldsymbol{x}_\star) + \sum_i \langle \boldsymbol{\Lambda}, \mathcal{P}_{T_i}\boldsymbol{\Delta}_i\rangle + \lambda_i\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)} \\
&= f(\boldsymbol{x}_\star) + \sum_i \langle \boldsymbol{\Lambda}, \boldsymbol{\Delta}_i\rangle - \langle \boldsymbol{\Lambda}, \mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\rangle + \lambda_i\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)} \\
&\geq f(\boldsymbol{x}_\star) + \Big\langle \boldsymbol{\Lambda}, \sum_j \boldsymbol{\Delta}_j\Big\rangle + \sum_i -\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^*\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)} + \lambda_i\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)} \\
&= f(\boldsymbol{x}_\star) + \Big\langle \mathcal{P}_{Q^\perp}\boldsymbol{\Lambda}, \sum_j \boldsymbol{\Delta}_j\Big\rangle + \sum_i \left(\lambda_i - \|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^*\right)\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)} \\
&= f(\boldsymbol{x}_\star) + \sum_i \left(\lambda_i - \|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^*\right)\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i\|_{(i)}, \qquad \text{(B.4)}
\end{aligned}$$

where we have used that $\langle \mathcal{P}_Q\boldsymbol{\Lambda}, \sum_j \boldsymbol{\Delta}_j\rangle = \langle \boldsymbol{\Lambda}, \mathcal{P}_Q\sum_j \boldsymbol{\Delta}_j\rangle = 0$, since $\boldsymbol{\delta}$ is feasible. Since each of the $\|\mathcal{P}_{T_i^\perp}\boldsymbol{\Lambda}\|_{(i)}^*$ is strictly smaller than λ_i, if any of the $\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i$ are nonzero, then $f(\boldsymbol{x}_\star+\boldsymbol{\delta}) > f(\boldsymbol{x}_\star)$. If, on the other hand, all of the $\mathcal{P}_{T_i^\perp}\boldsymbol{\Delta}_i$ are zero, then $\boldsymbol{\Delta}_i \in T_i$ for all i, and the constraint $\mathcal{P}_Q\sum_i \boldsymbol{\Delta}_i = \boldsymbol{0}$ implies that $\sum_i \boldsymbol{\Delta}_i \in (T_1 + \cdots + T_\tau) \cap Q^\perp$. If $\sum_i \boldsymbol{\Delta}_i \neq \boldsymbol{0}$, this contradicts independence of $(T_1, \ldots, T_\tau, Q^\perp)$. If $\sum_i \boldsymbol{\Delta}_i = \boldsymbol{0}$, this contradicts independence of $T_1, \ldots, T_\tau$ (which follows from independence of $(T_1, \ldots, T_\tau, Q^\perp)$). So, we conclude that for any feasible perturbation $\boldsymbol{\delta}$, $f(\boldsymbol{x}_\star + \boldsymbol{\delta})$ is strictly larger than $f(\boldsymbol{x}_\star)$. □

C Proof of Lemma 5.1: Operator Approximations

Proof. Fix an 1/4-net Γ for the unit ball of the Frobenius norm, restricted to S. By [Ver11] Lemma 5.2, there exists such a net of size at most $\exp(\dim(S)\log 9)$. Let $\mathcal{H} : \mathbb{R}^\gamma \to \mathbb{R}^{m\times n}$ via $\mathcal{H}\boldsymbol{x} = \sum_{i=1}^\gamma \boldsymbol{H}_i x_i$, and let $\psi : \mathbb{R}^\gamma \to \mathbb{R}^{m\times n}$ via $\psi\boldsymbol{x} = \sum_{i=1}^\gamma \bar{\boldsymbol{H}}_i x_i$, where $(\bar{\boldsymbol{H}}_i)$ is an orthonormal sequence of matrices that span R. By the Bartlett decomposition, we may assume that $\left[\mathrm{vec}\left[\bar{\boldsymbol{H}}_1\right] \mid \cdots \mid \mathrm{vec}\left[\bar{\boldsymbol{H}}_\gamma\right]\right] \in \mathbb{R}^{mn\times\gamma}$ is distributed according to the Haar measure on the Stiefel manifold of $mn \times \gamma$ matrices with orthonormal columns. Moreover, we have $\mathcal{A} = \mathcal{H}\mathcal{H}^*$ and $\mathcal{P}_R = \psi\psi^*$.

A standard argument (see [Ver11] Lemma 5.4) gives that

$$\left\| \mathcal{P}_S \frac{mn}{\gamma} \mathcal{A} \mathcal{P}_S - \mathcal{P}_S \right\|_{F,F} = \sup_{\substack{\boldsymbol{X} \in S \\ \|\boldsymbol{X}\|_F = 1}} \left| \frac{mn}{\gamma} \|\mathcal{H}^* \boldsymbol{X}\|_2^2 - 1 \right| \leq 2 \sup_{\boldsymbol{X} \in \Gamma} \left| \frac{mn}{\gamma} \|\mathcal{H}^* \boldsymbol{X}\|_2^2 - 1 \right|. \quad \text{(C.1)}$$

Notice that $\sqrt{\frac{mn}{\gamma}} \mathcal{H}^* \boldsymbol{X}$ is distributed as an iid $\mathcal{N}(0, 1/\gamma)$ random vector. Using Lemma 1 of [LM00],

$$\mathbb{P}\left[\left| \frac{mn}{\gamma} \|\mathcal{H}^* \boldsymbol{X}\|_2^2 - 1 \right| \geq 2\sqrt{\frac{t}{\gamma}} + 2\frac{t}{\gamma} \right] \leq 2\, e^{-t}. \quad \text{(C.2)}$$

Choose $t = c_1 \gamma$, with c_1 small enough that $4\sqrt{c_1} + 4c_1 \leq 1/2$. Take a union bound over all $\exp(\dim(S) \log 9)$ elements of Γ to get

$$\mathbb{P}\left[\left\| \mathcal{P}_S \frac{mn}{\gamma} \mathcal{A} \mathcal{P}_S - \mathcal{P}_S \right\|_{F,F} \geq \frac{1}{2} \right] \leq 2 \exp\left(-c_1 \gamma + \dim(S) \log 9 \right). \quad \text{(C.3)}$$

Using the assumption that $\gamma > C_1 \dim(S)$, and ensuring that C_1 is large enough that $c_1 > \frac{\log 9}{C_1}$ completes the proof of (5.2).

For the second term, we repeat the argument, noting that

$$\left\| \frac{mn}{\gamma} \mathcal{P}_S \mathcal{P}_R \mathcal{P}_S - \mathcal{P}_S \right\|_{F,F} \leq 2 \sup_{\boldsymbol{X} \in \Gamma} \left| \frac{mn}{\gamma} \|\psi^* \boldsymbol{X}\|_2^2 - 1 \right|. \quad \text{(C.4)}$$

Note that $\|\psi^* \boldsymbol{X}\|_2^2 = \|(\mathrm{vec} \circ \psi)^* \mathrm{vec}\,[\boldsymbol{X}]\,\|_2^2$. The operator $\mathrm{vec} \circ \psi : \mathbb{R}^\gamma \to \mathbb{R}^{mn}$ can be identified with an $mn \times \gamma$ matrix $\boldsymbol{U}$, which per the above discussion can be taken to be distributed according to the Haar measure. By orthogonal invariance, for any fixed $\boldsymbol{x}$, $\boldsymbol{U}^* \boldsymbol{x}$ is equal in distribution to the restriction of a uniformly distributed random unit vector $\boldsymbol{r} \in \mathbb{S}^{mn-1}$ to its first γ coordinates. Lemma 2.2 of [DG03] provides convenient tail bounds for the norm of such a coordinate restriction. Applying that lemma, we have that for every $t > 0$, there exists $c_t > 0$ such that

$$\mathbb{P}\left[\left| \frac{mn}{\gamma} \|\psi^* \boldsymbol{X}\|_2^2 - 1 \right| > t \right] \leq \exp\left(-c_t \gamma \right). \quad \text{(C.5)}$$

Set $t = 1/32$. As above, ensuring that C_1 is larger than $\frac{\log 9}{c_t}$ and taking a union bound shows that with the desired probability $\left\| \frac{mn}{\gamma} \mathcal{P}_S \mathcal{P}_R \mathcal{P}_S - \mathcal{P}_S \right\|_{F,F} \leq 1/16$. Rescaling gives the bound quoted in the statement of the lemma. □

D Derandomizing the Signs

In this section, we show how the assumption that the signs of the sparse term $\boldsymbol{S}_0$ are random can be removed. To be clear, one can replace the assumtion that $\mathrm{sign}\,(\boldsymbol{S}_0)$ follows a Bernoulli-Rademacher model with nonzero probability ρ, with the assumption that $\mathrm{sign}\,(\boldsymbol{S}_0) = \mathcal{P}_\Omega \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is *any* fixed matrix of signs, and $\mathcal{P}_\Omega$ is an independent Bernoulli subset with nonzero probability $\rho/2$. The argument follows very directly that of [CLMW11].

Lemma D.1. *Suppose that* $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ *is the unique optimal solution to the Compressive Principal Component Pursuit problem*

$$\text{minimize} \quad \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \mathcal{P}_Q[\boldsymbol{L} + \boldsymbol{S}] = \boldsymbol{D} \tag{D.1}$$

with input data $\boldsymbol{D} = \mathcal{P}_Q[\boldsymbol{L}_0 + \boldsymbol{S}_0]$. *Then for any subset* $W \subset [m] \times [n]$, *if we set* $\boldsymbol{S}_0' = \mathcal{P}_W \boldsymbol{S}_0$, *the pair* $(\boldsymbol{L}_0, \boldsymbol{S}_0')$ *is the unique optimal solution to the Compressive PCP problem with input data* $\boldsymbol{D}' = \mathcal{P}_Q[\boldsymbol{L}_0 + \boldsymbol{S}_0']$.

Proof. Let $(\hat{\boldsymbol{L}}, \hat{\boldsymbol{S}})$ denote an optimal solution with input $\boldsymbol{D}'$. Then

$$\left\|\hat{\boldsymbol{L}}\right\|_* + \lambda \left\|\hat{\boldsymbol{S}}\right\|_1 \;\leq\; \|\boldsymbol{L}_0\|_* + \lambda \left\|\boldsymbol{S}_0'\right\|_1 .$$

Consider the pair $(\hat{\boldsymbol{L}}, \hat{\boldsymbol{S}} + \boldsymbol{S}_0 - \boldsymbol{S}_0')$. By linearity,

$$\begin{aligned} \mathcal{P}_Q[\hat{\boldsymbol{L}} + \hat{\boldsymbol{S}} + \boldsymbol{S}_0 - \boldsymbol{S}_0'] &= \mathcal{P}_Q[\hat{\boldsymbol{L}} + \hat{\boldsymbol{S}}] + \mathcal{P}_Q[\boldsymbol{S}_0 - \boldsymbol{S}_0'] \\ &= \mathcal{P}_Q[\boldsymbol{L}_0 + \boldsymbol{S}_0] + \mathcal{P}_Q[\boldsymbol{S}_0 - \boldsymbol{S}_0'] \;=\; \boldsymbol{D}, \end{aligned}$$

and so this pair is feasible for the original problem.

By the triangle inequality, we have

$$\begin{aligned} \left\|\hat{\boldsymbol{L}}\right\|_* + \lambda \left\|\hat{\boldsymbol{S}} + \boldsymbol{S}_0 - \boldsymbol{S}_0'\right\|_1 &\leq \left\|\hat{\boldsymbol{L}}\right\|_* + \lambda \left\|\hat{\boldsymbol{S}}\right\|_1 + \lambda \left\|\boldsymbol{S}_0 - \boldsymbol{S}_0'\right\|_1 \\ &\leq \|\boldsymbol{L}_0\|_+ \lambda \left\|\boldsymbol{S}_0'\right\|_1 + \lambda \left\|\boldsymbol{S}_0 - \boldsymbol{S}_0'\right\|_1 \\ &= \|\boldsymbol{L}_0\|_+ \lambda \|\boldsymbol{S}_0\|_1 . \end{aligned}$$

In the final equality, we have used the fact that for any coordinate subspace W and matrix $\boldsymbol{X}$, $\|\mathcal{P}_W \boldsymbol{X}\|_1 + \|\mathcal{P}_{W^\perp} \boldsymbol{X}\|_1 = \|\boldsymbol{X}\|_1$. By unique optimality of $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ for the problem with input $\boldsymbol{D}$, we have $\hat{\boldsymbol{L}} = \boldsymbol{L}_0$, and $\hat{\boldsymbol{S}} = \boldsymbol{S}_0'$. □

Lemma D.2. *Suppose that the solution to the CPCP problem with input data* $\boldsymbol{D} = \mathcal{P}_Q[\boldsymbol{L}_0 + \boldsymbol{S}_0]$ *is unique and equal to* $(\boldsymbol{L}_0, \boldsymbol{S}_0)$, *with probability* p_0 *in the model in which the entries of* $\operatorname{sign}(\boldsymbol{S}_0)$ *are iid Bernoulli-Rademacher random variables with nonzero probability* ρ. *Then with the same probability* p_0, *the solution is unique and equal to* $(\boldsymbol{L}_0, \boldsymbol{S}_0)$, *in the model in which* $\operatorname{sign}(\boldsymbol{S}_0) = \mathcal{P}_W \boldsymbol{\Sigma}$, *where* $\mathcal{P}_W$ *is an iid* Bernoulli$(\rho/2)$ *random matrix, and* $\boldsymbol{\Sigma}$ *is any fixed matrix of signs.*

Proof. Let $\boldsymbol{\Sigma}_0$ be iid Bernoulli-Rademacher, with nonzero probability ρ, as in Theorem 2.1. Set

$$W = \{(i,j) \mid [\boldsymbol{\Sigma}_0]_{i,j} = [\boldsymbol{\Sigma}]_{i,j}\} .$$

By Lemma D.1, whenever $(\boldsymbol{L}_0, \boldsymbol{S}_0)$ is uniquely recovered by CPCP, so is $(\boldsymbol{L}_0, \mathcal{P}_W \boldsymbol{S}_0)$. The result follows by observing that $(i,j) \in W$ independently with probability $\rho/2$. □