Compressed sensing of low-rank plus sparse matrices

Jared Tanner[a,b], Simon Vary[a,*]

[a] *Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK*
[b] *The Alan Turing Institute, The British Library, London NW1 2DB, UK*

Abstract

Expressing a matrix as the sum of a low-rank matrix plus a sparse matrix is a flexible model capturing global and local features in data. This model is the foundation of robust principle component analysis [1, 2], and popularized by dynamic-foreground/static-background separation [3] amongst other applications. Compressed sensing, matrix completion, and their variants [4, 5] have established that data satisfying low complexity models can be efficiently measured and recovered from a number of measurements proportional to the model complexity rather than the ambient dimension. This manuscript develops similar guarantees showing that $m \times n$ matrices that can be expressed as the sum of a rank-r matrix and a s-sparse matrix can be recovered by computationally tractable methods from $\mathcal{O}(r(m+n-r)+s)\log(mn/s)$ linear measurements. More specifically, we establish that the restricted isometry constants for the aforementioned matrices remain bounded independent of problem size provided p/mn, s/p, and $r(m+n-r)/p$ reman fixed. Additionally, we show that semidefinite programming and two hard threshold gradient descent algorithms, NIHT and NAHT, converge to the measured matrix provided the measurement operator's RIC's are sufficiently small. Numerical experiments illustrating these results are shown for synthetic problems, dynamic-foreground/static-background separation, and multispectral imaging.

Keywords: matrix sensing, low-rank plus sparse matrix, restricted isometry property, non-convex methods, robust PCA

2010 MSC: 15A29, 41A29, 62H25, 65F10, 65J20, 68Q25, 90C22, 90C26

1. Introduction

Data with a known underlying low-dimensional structure can often be estimated from a number of measurements proportional to the degrees of freedom of the underlying model, rather than what

*Corresponding author
Email addresses: tanner@maths.ox.ac.uk (Jared Tanner), vary@maths.ox.ac.uk (Simon Vary)
[1] This publication is based on work partially supported by: the EPSRC I-CASE studentship (voucher 15220165) in partnership with Leonardo, The Alan Turing Institute through EPSRC (EP/N510129/1) and the Turing Seed Funding grant SF019.

its ambient dimension would suggests. Examples of such low-dimensional structures for which the aforementioned is true include: compressed sensing [6, 7, 8], matrix completion [9, 10, 11], sparse measures [12, 13, 14], and atomic decompositions [15] more generally. Our work extends these results to the matrices which are formed as a sum of a low-rank matrix and a sparse matrix, a model popularized by the work on robust principle component anaysis (RPCA) [1, 2]. Specifically, we consider matrices $X \in \mathbb{R}^{m\times n}$ of the form $X = L + S$, where L is of rank at most r, and S has at most s non-zero entries, $\|S\|_0 \leq s$. The low-rank plus sparse model is a rich model with the low rank component modeling global correlations, while the additive sparse component allows a fixed number of entries to deviate from this global model in an arbitrary way. Among applications of this model are image restoration [16], hyperspectral image denoising [17, 18, 19], face detection [20, 21], acceleration of dynamic MRI data acquisition [22], analysis of medical imagery [23], separation of moving objects in at otherwise static scene [3], and target detection [24].

Unlike RPCA where X is directly available, we consider the compressed sensing setting where X is measured through a linear operator $\mathcal{A}(\cdot)$, where $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, $b \in \mathbb{R}^p$ and typically $p \ll mn$. Our contributions extend existing results on *restricted isometry constants* (RIC) for Gaussian and other measurement operators for sparse vectors [25] or low-rank matrices [11] to the sets of low-rank plus sparse matrices. For the set of matrices which are the sum of a low-rank plus a sparse matrix the results differ subtly due to the space not being closed, in that there are matrices X for which there does not exist a nearest projection to the set of low-rank plus sparse matrices [26]. To overcome this, we introduce the set of low-rank plus sparse matrices with a constraint on the Frobenius norm of its low-rank component, see Definition 1.1.

Definition 1.1 (Low-rank plus sparse set $\mathrm{LS}_{m,n}(r, s, \tau)$)**.** *Denote the set of $m \times n$ real matrices that are the sum of a rank r matrix and a s sparse matrix as*

$$\mathrm{LS}_{m,n}(r, s, \tau) = \left\{ L + S \in \mathbb{R}^{m\times n} : \mathrm{rank}(L) \leq r, \; \|S\|_0 \leq s, \; \|L\|_F \leq \tau \right\}, \tag{1}$$

where the rank r matrix has its Frobenius norm upper bounded by τ.

The natural generalization of the RIC definition from sparse vectors and low-rank matrices to the space $\mathrm{LS}_{m,n}(r, s, \tau)$ is given in Definition 1.2.

Definition 1.2 (RIC for $\mathrm{LS}_{m,n}(r, s, \tau)$)**.** *Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ be a linear map. For every pair of integers (r, s) and every $\tau > 0$, define the (r, s, τ)-restricted isometry constant to be the smallest $\Delta_{r,s,\tau}(\mathcal{A}) > 0$ such that*

$$(1 - \Delta_{r,s,\tau}(\mathcal{A})) \|X\|_F^2 \leq \| \mathcal{A}(X)\|_2^2 \leq (1 + \Delta_{r,s,\tau}(\mathcal{A})) \|X\|_F^2, \tag{2}$$

for all matrices $X \in \mathrm{LS}_{m,n}(r, s, \tau)$.

Linear maps $\mathcal{A}$ which have a sufficient concentration of measure phenomenon can overcome the dimensionality of $\mathrm{LS}_{m,n}(r,s,\tau)$ to achieve $\Delta_{r,s,\tau}$ which is bounded by a fixed value independent of dimension size provided the number of measurements p is proportional to the degrees of freedom of a rank-r plus sparsity-s matrix $r(m+n-r)+s$. A suitable class of random linear maps is captured in the following definition.

Definition 1.3 (Nearly isometrically distributed map)**.** *Let $\mathcal{A}$ be a random variable that takes values in linear maps $\mathbb{R}^{m\times n} \to \mathbb{R}^p$. We say that $\mathcal{A}$ is nearly isometrically distributed if, for $\forall X \in \mathbb{R}^{m\times n}$,*

$$\mathbb{E}\left[\|\mathcal{A}(X)\|^2\right] = \|X\|_F^2 \tag{3}$$

and for all $\varepsilon \in (0,1)$, we have

$$\Pr\left(\left|\|\mathcal{A}(X)\|_2^2 - \|X\|_F^2\right| \geq \varepsilon\|X\|_F^2\right) \leq 2\exp\left(-\frac{p}{2}\left(\varepsilon^2/2 - \varepsilon^3/3\right)\right), \tag{4}$$

and there exists some constant $\gamma > 0$ such that for all $t > 0$, we have

$$\Pr\left(\|\mathcal{A}\| \geq 1 + \sqrt{\frac{mn}{p}} + t\right) \leq \exp\left(-\gamma p t^2\right). \tag{5}$$

There are two crucial properties for a random map to be nearly isometric. Firstly, it needs to be isometric in expectation as in (3), and exponentially concentrated around the expected value as in (4). Secondly, the probability of large distortions of length must be exponentially small as in (5). This ensures that even after taking a union bound over an exponentially large covering number for $\mathrm{LS}_{m,n}(r,s,\tau)$, see Lemma 2.3, the probability of distortion remains small [25, 11].

In addition to developing RIC bounds as in Definition 1.2 we also show that the RIC of an operator implies uniqueness of the decomposition and that exact recovery is possible with computationally efficient algorithms such as convex relaxations or gradient descent methods. The following subsection summarizes our main contributions. The rest of the paper is organized as

- In Section 2, we prove that the RICs of $\mathrm{LS}_{m,n}(r,s,\tau)$ for Gaussian and fast Johnson-Lindenstrauss transform (FJLT) measurement operators remain bounded independent of problem size provided the number of measurements p is proportional to $\mathcal{O}\left(r(m+n-r)+s\right)$.

- In Section 3, we prove that when the RICs of $\mathcal{A}(\cdot)$ are suitably bounded then the solution to a linear system $\mathcal{A}(X_0) = b$ has a unique decomposition in $\mathrm{LS}_{m,n}(r,s,\tau)$ that can be recovered using computationally efficient convex optimization solvers and hard thresholding gradient descent algorithms which are natural extensions of algorithms developed for compressed sensing [27] and matrix completion [28].

- In Section 4, we empirically study the average case of recovery on synthetic data by solving convex optimization and by the proposed gradient descent methods and observe a phase transition in the space of parameters for which the methods succeed. We also give an example of two practical applications of the low-rank plus sparse matrix recovery in the form of a subsampled dynamic-foreground/static-background video separation and robust recovery of multispectral imagery.

1.1. Main contribution

The foundational analytical tool for our results is the RIC for $\mathrm{LS}_{m,n}(r,s,\tau)$, which as for other RICs [25, 11], follows from balancing a covering number for the space $\mathrm{LS}_{m,n}(r,s,\tau)$ and the measurement operator being a near isometry as defined in Definition 1.3.

Theorem 1 (RIC for $\mathrm{LS}_{m,n}(r,s,\tau)$)**.** *For a given m,n,p and $\Delta \in (0,1)$ and a random linear transform $\mathcal{A}:\mathbb{R}^{m\times n}\to\mathbb{R}^p$ satisfying the concentration of measure inequalities in Definition 1.3, there exist constants $c_0, c_1 > 0$ such that the RIC for $\mathrm{LS}_{m,n}(r,s,\tau)$ is upper bounded with $\Delta_{r,s,\tau}(\mathcal{A}) \leq \Delta$ provided*

$$p > c_0\left(r(m+n-r)+s\right)\log\left((1+\tau)\frac{mn}{s}\right), \tag{6}$$

with probability at least $1-\exp(-c_1 p)$, where c_0, c_1 are constants that depend only on Δ.

Theorem 1 establishes that for random ensembles of linear transformations that satisfy the concentration of measure inequalities in Definition 1.3, the RIC for $\mathrm{LS}_{m,n}(r,s,\tau)$ is upper bounded in the asymptotic regime as m,n and p approach infinity at appropriate rates. Specifically, the RIC remains bounded independent of the problem dimensions m and n provided p to be taken proportional to the order of degrees of freedom of the rank-r plus sparsity-s matrices times a logarithmic factor as in (6).

Examples of random ensembles of $\mathcal{A}$ which satisfy the conditions of Definition 1.3 include random Gaussian ensemble which acquires the information about the matrix X through p linear measurements of the form

$$b_\ell := \mathcal{A}(X)_\ell = \langle A^{(\ell)}, X\rangle \quad \text{for} \quad \ell = 1,2,\ldots,p, \tag{7}$$

where the p distinct sensing matrices $A^{(\ell)} \in \mathbb{R}^{m\times n}$ are the sensing operators defining $\mathcal{A}$ and have entries sampled from the Gaussian distribution as $A^{(\ell)}_{i,j} \sim \mathcal{N}(0,1/p)$. Other notable examples include symmetric Bernoulli ensembles, and Fast Johnson-Lindenstrauss Transform (FJLT) [29, 30].

For a linear transform $\mathcal{A}$ which has RIC suitably upper bounded and a given vector of samples $b=\mathcal{A}(X_0)$, the matrix X_0 is the only matrix in the set $\mathrm{LS}_{m,n}(r,s,\tau)$ that satisfies the linear constraint.

Theorem 2 (Existence of a unique solution for $\mathcal{A}$ with RIC)**.** *Suppose that $\Delta_{2r,2s,2\tau}(\mathcal{A}) < 1$ for some integers $r, s \geq 1$ and a non-negative real number $\tau > 0$. Let $b = \mathcal{A}(X_0)$, then X_0 is the only matrix in the set $\mathrm{LS}_{m,n}(r,s,\tau)$ satisfying $\mathcal{A}(X) = b$.*

Proof. Assume, on the contrary, that there exists a matrix $X \in \mathrm{LS}_{m,n}(r,s,\tau)$ such that $\mathcal{A}(X) = b$ and $X \neq X_0$. Then $Z := X_0 - X$ is a non-zero matrix and $Z \in \mathrm{LS}_{m,n}(2r, 2s, 2\tau)$ and $\mathcal{A}(Z) = 0$. But then by the RIC we would have $0 = \|\mathcal{A}(Z)\|_2^2 \geq (1 - \Delta_{2r,2s,2\tau}) \|Z\|_F^2 > 0$, which is a contradiction. □

As in compressed sensing and matrix completion, it is in general NP-hard to recovery $X_0 = L_0 + S_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ from $\mathcal{A}(X_0)$ for minimal r, s when $p \ll mn$. This follows from the non-convexity of the feasible set $\mathrm{LS}_{m,n}(r,s,\tau)$. However, we show that if the linear transformation $\mathcal{A}$ has sufficiently small RIC over the set $\mathrm{LS}_{m,n}(r,s,\tau)$, then the solution can be obtained with computationally tractable methods such as by solving the semidefinite program

$$\min_{X=L+S\in\mathbb{R}^{m\times n}} \|L\|_* + \lambda \|S\|_1 , \qquad \text{s.t.} \quad \|\mathcal{A}(X) - b\|_2 \leq \varepsilon_{\mathrm{b}}, \quad \|L\|_F \leq \tau \tag{8}$$

where $\|\cdot\|_*$ is the Schatten 1-norm and $\|\cdot\|_1$ is the sum of the absolute value of the entries[2] and ε_{b} is the model misfit.

Theorem 3 (Guaranteed recovery by the convex relaxation)**.** *Let $b = \mathcal{A}(X_0)$ and suppose that $r, s \geq 1$ and $\tau > 0$ are such that the restricted isometry constant $\Delta_{4r,3s,2\tau}(\mathcal{A}) \leq \frac{1}{5}$. Let $X_* = L_* + S_*$ be the solution of* (8) *with $\lambda = \sqrt{r/s}$, then $\|X_* - X_0\|_F \leq 67\,\varepsilon_{\mathrm{b}}$.*

Alternatively, X_0 can be obtained from its compressed measurements $\mathcal{A}(X_0)$ by iterative gradient descent methods that are guaranteed to converge to a global minimizer of the non-convex optimization problem

$$\min_{X=L+S\in\mathbb{R}^{m\times n}} \|\mathcal{A}(X) - b\|_2 , \qquad \text{s.t.} \quad X \in \mathrm{LS}_{m,n}(r,s,\tau). \tag{9}$$

We introduce two natural extensions of the simple yet effective Normalized Iterative Hard Thresholding (NIHT) for compressed sensing [27] and matrix completion [28] algorithms, here called NIHT and Normalized Alternative Hard Thresholding (NAHT) for low-rank plus sparse matrices, Algorithms 1 and 2 respectively. In both cases we establish that if the measurement operator has suitably small RICs then NIHT and NAHT provably converge to the global minimum of the non-convex problem formulated in (9) and recover $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ for which $b = \mathcal{A}(X_0)$.

Theorem 4 (Guaranteed recovery by NIHT)**.** *Suppose that $r, s \geq 1$ and $\tau > 0$ are such that the restricted isometry constant $\Delta_3 := \Delta_{3r,3s,2\tau}(\mathcal{A}) < \frac{1}{5}$. Then NIHT applied to $b = \mathcal{A}(X_0)$ as described*

[2] Our use of $\|\cdot\|_1$ as the sum of the modulus of the entries of a matrix differs from the vector induced 1-norm of a matrix.

Algorithm 1 Normalized Iterative Hard Thresholding (NIHT) for LS recovery

Input: $b = \mathcal{A}(X_0), \mathcal{A}, r, s$, and termination criteria

Set: $(L^0, S^0) = \mathcal{P}\left(\mathcal{A}^*(b); r, s, \tau, \varepsilon\right),\ X^0 = L^0 + S^0,\ j = 0$

$\Omega^0 = \mathrm{supp}(S^0)$ and U^0 as the top r left singular vectors of L^0

```
while not converged do
    Compute the residual R^j = A^* (b − A(X^j))
    Compute the stepsize: α_j = ‖Proj_(U^j,Ω^j) (R^j)‖_F^2 / ‖A (Proj_(U^j,Ω^j) (R^j))‖_2^2
    Set W^j = X^j + α_j R^j
    Compute (L^{j+1}, S^{j+1}) = RPCA_{r,s,τ}(W^j, ε_p) and set X^{j+1} = L^{j+1} + S^{j+1}
    Let Ω^{j+1} = supp(S^{j+1}) and U^{j+1} be the top r left singular vectors of L^{j+1}
    j = j + 1
end while
```

Output: X^j

in Algorithm 1 will linearly converge as

$$\|X^{j+1} - \widehat{X}\|_F \leq 8\frac{\Delta_3\,(1-3\Delta_3)}{(1-\Delta_3)^2}\|X^j - \widehat{X}\|_F + \frac{1-5\Delta_3}{1-\Delta_3}\,\varepsilon_{\mathrm{p}}, \tag{10}$$

where ε_{p} *is the accuracy of the Robust PCA oblique projection that performs projection on the set of the low-rank plus sparse matrices* $\mathrm{LS}_{m,n}(r,s,\tau)$ *and* $\widehat{X}$ *is a matrix in proximity of* X_0

$$\|\widehat{X} - X_0\|_F \leq \varepsilon_{\mathrm{p}}\,\frac{1-\Delta_3}{1-5\Delta_3}. \tag{11}$$

Theorem 5 (Guaranteed recovery by NAHT)**.** *Suppose that* $r, s \geq 1$ *and* $\tau > 0$ *are such that the restricted isometry constant* $\Delta_3 := \Delta_{3r,3s,2\tau}(\mathcal{A}) < \frac{1}{9}$*. Then NAHT applied to* $b = \mathcal{A}(X_0)$ *as described in Algorithm 2 will linearly converge to* $X_0 = L_0 + S_0$ *as*

$$\left\|L_0 - L^{j+1}\right\|_F + \left\|S_0 - S^{j+1}\right\|_F \leq \frac{6\Delta_3}{1-3\Delta_3}\left(\left\|L^j - L_0\right\|_F + \left\|S^j - S_0\right\|_F\right). \tag{12}$$

Note that the projection used in computing the stepsize is defined as $\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right) := P_{U^j}R^j + \mathbb{1}_{\Omega^j} \circ (R^j - P_{U^j}R^j)$, where $P_{U^j} := U^j\left(U^j\right)^*$, $\mathbb{1}_{\Omega^j}$ is a matrix with ones at indices Ω^j, and $\circ$ denotes the entry-wise Hadamard product. This corresponds to first projecting the left singular vectors of R^j on the subspace spanned by columns of U^j and then setting entries at indices Ω^j to be equal to the entries of R^j at indices Ω^j. One can repeat this process to achieve better more precise projection of R^j in the low-rank plus sparse matrix set defined by $\left(U^j, \Omega^j\right)$.

The hard thresholding projection in Algorithm 1 is performed by computing Robust PCA which is solved to an accuracy proportional to ε_{p} as given by (11). The RPCA projection of a matrix $W \in \mathbb{R}^{m\times n}$ on the set of $\mathrm{LS}_{m,n}(r,s,\tau)$ with precision ε_{p} returns a matrix $X \in \mathrm{LS}_{m,n}(r,s,\tau)$ such that

Algorithm 2 Normalized Alternating Hard Thresholding (NAHT) for LS recovery

Input: $b = \mathcal{A}(X_0), \mathcal{A}, r, s,$ and termination criteria

Set: $(L^0, S^0) = \mathcal{P}\left(\mathcal{A}^*(b); r, s, \tau, \varepsilon\right), \; X^0 = L^0 + S^0, \; j = 0$

$\Omega^0 = \mathrm{supp}(S^0)$ and U^0 as the top r left singular vectors of L^0

while not converged **do**

  Compute the residual $R_L^j = \mathcal{A}^*\left(\mathcal{A}(X^j) - b\right)$

  Compute the stepsize $\alpha_j^L = \left\|\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right)\right\|_F^2 \Big/ \left\|\mathcal{A}\left(\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right)\right)\right\|_2^2$

  Set $V^j = L^j - \alpha_j^L R_L^j$

  Set $L^{j+1} = \mathrm{HT}(V^j; r)$ and let U^{j+1} be the left singular vectors of L^{j+1}

  Set $X^{j+\frac{1}{2}} = L^{j+1} + S^j$

  Compute the residual $R_S^j = \mathcal{A}^*\left(\mathcal{A}(X^{j+\frac{1}{2}}) - b\right)$

  Compute the stepsize $\alpha_j^S = \left\|\mathrm{Proj}_{(U^{j+1},\Omega^j)}\left(R^j\right)\right\|_F^2 \Big/ \left\|\mathcal{A}\left(\mathrm{Proj}_{(U^{j+1},\Omega^j)}\left(R^j\right)\right)\right\|_2^2$

  Set $W^j = S^j - \alpha_j^S R_S^j$

  Set $S^{j+1} = \mathrm{HT}(W^j; s)$ and let $\Omega^{j+1} = \mathrm{supp}(S^{j+1})$

  Set $X^{j+1} = L^{j+1} + S^{j+1}$

  $j = j + 1$

end while

Output: $X^j = L^j + S^j$

$$X \leftarrow \mathrm{RPCA}_{r,s,\tau}(W, \varepsilon_{\mathrm{p}}) \qquad \text{s.t.} \qquad \left\|X - W_{\mathrm{rpca}}\right\|_F \leq \varepsilon_{\mathrm{p}}, \tag{13}$$

where $W_{\mathrm{rpca}} := \arg\min_{Y \in \mathrm{LS}_{m,n}(r,s,\tau)} \|Y - W\|_F$ is the optimal projection of the matrix W on the set $\mathrm{LS}_{m,n}(r, s, \tau)$.

1.2. Relation to prior work

It is well known that the low-rank plus sparse matrix decomposition solved by RPCA does not need to have a unique solution without further constraints, such as the singular vectors of the low-rank component being uncorrelated with the canonical basis as quantified by the incoherence condition [9, 11] with parameter μ

$$\max_{i \in \{1,\ldots,r\}} \|U^* e_i\|_2 \leq \sqrt{\frac{\mu r}{m}}, \qquad \max_{i \in \{1,\ldots,r\}} \|V^* e_i\|_2 \leq \sqrt{\frac{\mu r}{n}}, \tag{14}$$

where $L = U \Sigma V^*$ is the singular value decomposition of the rank r component L of size $m \times n$. The incoherence condition for small values of μ ensures that the left and the right singular vectors are well spread out and not sparse. It is therefore sensible to expect that the problem of recovering $X_0 \in \mathrm{LS}_{m,n}(r, s, \tau)$ from subsampled measurements should obey the same conditions. The incoherence assumption is directly used in NIHT in the RPCA projection step in Algorithm 1, Line 5, the solution of which is dependent on the incoherence of L. The incoherence assumption in the convergence analysis of NAHT and the convex recovery is implicitly assumed in the form of RIC, where we restrict the energy of τ which constraints the correlation of L and S.

The results presented here extend the well developed literature on compressed sensing and matrix completion/sensing [4, 5] to the setting of low-rank plus sparse matrices as defined in Def. 1.1. These foundational RIC bound results allow for further extension to other non-convex algorithms, such as [31], further model based constraints as in [32] and other additive models.

The recovery result by convex relaxation in Theorem 3 controls the measurement error and/or model mismatch ε_{b}. In the proof of NIHT convergence in Theorem 4 we consider exact measurements but we control the error of the Robust PCA projection which is assumed to be solved only within prescribed precision ε_{p}. The convergence result of NAHT in Theorem 5 alternates between projecting of the low-rank and the sparse component. The non-convex algorithms are also stable to error ε_{b}, but we omit the stability analysis for clarity in the proofs.

2. Restricted Isometry Constants for $\mathrm{LS}_{m,n}(r, s, \tau)$

This section presents a proof of Theorem 1, that linear maps $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ sampled from a class of probability distributions obeying concentration of measure and large deviation inequalities, have bounded RIC for sets of low-rank plus sparse matrices with bounded energy as defined in Definition 1.2. More precisely, that the RIC of $\mathcal{A}$ remains bounded independent of dimension once $p \geq \mathcal{O}\left((r(m+n-r)+s)\log\left((1+\tau)\frac{mn}{s}\right)\right)$. Examples of linear maps which satisfy these bounds include random Gaussian matrices and the Fast Johnson-Lindenstrauss transform (FJLT) [29, 30]. We extend the method of proof used in the context of sparse vectors by [25] and its alteration for the low-rank matrix recovery by [11].

The proof of Theorem 1 begins with the derivation of an RIC for a single subspace $\Sigma_{m,n}(V, W, T, \tau)$ of $\mathrm{LS}_{m,n}(r, s, \tau)$ when the column space of $\mathcal{C}(L)$ is restricted in the subspace V, the row space $\mathcal{C}(L^T)$ in the subspace W and the sparse component S is in the subspace T

$$\Sigma_{m,n}(V, W, T, \tau) = \left\{X = L + S \in \mathbb{R}^{m\times n} : \mathcal{C}(L) \subseteq V,\, \mathcal{C}(L^T) \subseteq W,\, \mathrm{supp}(S) \subseteq T,\, \|L\|_F \leq \tau\right\}. \quad (15)$$

Following this, we show that the isometry constant of $\mathcal{A}$ is robust to a perturbation of the column and the row subspaces (V, W) of the low-rank component. Finally, we use a covering argument over all possible column and row subspaces (V, W) of the low-rank component and count over all possible sparsity subspaces T of the sparse component to derive an exponentially small probability bound for the event that $\mathcal{A}(\cdot)$ satisfies RIC with constant Δ for sets

$$\mathrm{LS}_{m,n}(r,s,\tau) = \{\Sigma_{m,n}(V,W,T,\tau) \,:\, V \in \mathcal{G}(m,r),\, W \in \mathcal{G}(n,r),\, T \in \mathcal{V}(mn,s)\}\,, \tag{16}$$

where $\mathcal{G}(m,r)$ is the Grassmannian manifold – the set of all r-dimensional subspaces of $\mathbb{R}^m$, and $\mathcal{V}(mn,s)$ is the set of all possible supports sets of an $m \times n$ matrix that has s elements. Thus proving RIC for sets of low rank plus sparse matrices given the energy bound on the low-rank component L.

Our proof of Theorem 1 follows from proving the alternative form of (2) defined without the squared norms by

$$\left(1 - \widehat{\Delta}_{r,s,\tau}(\mathcal{A})\right) \|X\|_F \leq \|\,\mathcal{A}(X)\|_2 \leq \left(1 + \widehat{\Delta}_{r,s,\tau}(\mathcal{A})\right) \|X\|_F, \tag{17}$$

which we denote as $\widehat{\Delta}$. The discrepancy between (17) and (2) is due to (17) being more direct to derive and (2) allowing for more concise derivation of Theorem 3, 4, and 5.

The following result describes the behavior of $\mathcal{A}$ when constrained to a single fixed column and a row space (V,W) and a single sparse matrix space T.

Lemma 2.1 (RIC for a fixed LS subspace $\Sigma_{m,n}(V,W,T,\tau)$)**.** *Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ be a nearly isometric random linear map from Definition 1.3 and $\Sigma_{m,n}(V,W,T,\tau)$ as defined in* (15) *is fixed for some $(V,W), T$ and $\tau \in (0,1)$. Then for any $\widehat{\Delta} \in (0,1)$*

$$\forall X \in \Sigma\,(V,W,T,\tau): \quad (1-\widehat{\Delta})\|X\|_F \leq \|\,\mathcal{A}(X)\| \leq (1+\widehat{\Delta})\|X\|_F, \tag{18}$$

with probability at least

$$1 - 2\left(\frac{24}{\widehat{\Delta}}\tau\right)^{\dim V \cdot \dim W} \left(\frac{24}{\widehat{\Delta}}(1+\tau)\right)^{\dim T} \exp\left(-\frac{p}{2}\left(\frac{\widehat{\Delta}^2}{8} - \frac{\widehat{\Delta}^3}{24}\right)\right). \tag{19}$$

The proof follows the same argument as the one for sparse vectors [25, Lemma 5.1] and for low-rank matrices in [11, Lemma 4.3]. Our variant of the proof for low-rank plus sparse matrices is presented in Appendix B on page 36.

To establish the impact of a perturbation of the spaces (U,V) on the $\widehat{\Delta}$ in Lemma 2.1 we define a metric $\rho(\cdot,\cdot)$ on $\mathcal{G}(D,d)$ as follows

$$U_1, U_2 \in \mathcal{G}(D,d): \quad \rho(U_1,U_2) := \|P_{U_1} - P_{U_2}\|. \tag{20}$$

The Grassmannian manifold $\mathcal{G}\,(D,d)$ combined with distance $\rho(\cdot,\cdot)$ as in (20) defines a metric space $(\mathcal{G}\,(D,d)\,, \rho\,(\cdot,\cdot))$, where P_U denotes an orthogonal projection associated with the subspace U. Let us also denote a set of matrices whose column and row space is a subspace of V and W respectively

$$(V,W) = \left\{X;\, \mathcal{C}(X) \subseteq V,\, \mathcal{C}(X^T) \subseteq W\right\}, \tag{21}$$

and $P_{(V,W)}$ is an orthogonal projection that ensures that the column space and row space of $P_{(V,W)}X$ lies within V and W. The distance between $\Sigma_1 := \Sigma_{m,n}(V_1, W_1, T, \tau)$ and $\Sigma_2 := \Sigma_{m,n}(V_2, W_2, T, \tau)$ that have a fixed T is given by

$$\rho\left(\left(V_1, W_1\right), \left(V_2, W_2\right)\right) = \|P_{(V_1,W_1)} - P_{(V_2,W_2)}\|. \tag{22}$$

Lemma 2.2 (Variation of $\widehat{\Delta}$ in RIC in respect to a perturbation of (V, W))**.** *Let* $\Sigma_1 := \Sigma_{m,n}(V_1, W_1, T, \tau)$ *and* $\Sigma_2 := \Sigma_{m,n}(V_2, W_2, T, \tau)$ *be two low-rank plus sparse subspaces with the same fixed subspace* T*. Suppose that for* $\widehat{\Delta} > 0$*, the linear operator* $\mathcal{A}$ *satisfies*

$$\forall X \in \Sigma_1 : \quad (1 - \widehat{\Delta})\|X\|_F \leq \|\,\mathcal{A}(X)\| \leq (1 + \widehat{\Delta})\|X\|_F. \tag{23}$$

Then

$$\forall Y \in \Sigma_2 : \quad (1 - \widehat{\Delta}')\|Y\|_F \leq \|\,\mathcal{A}(Y)\| \leq (1 + \widehat{\Delta}')\|Y\|_F, \tag{24}$$

with $\widehat{\Delta}' := \widehat{\Delta} + \tau\rho\left(\left(V_1, W_1\right), \left(V_2, W_2\right)\right)\left(1 + \widehat{\Delta} + \|\,\mathcal{A}\,\|\right)$ *with* ρ *as defined in* (20)*.*

The proof is similar to the line of argument made in [11, Lemma 4.4], see Appendix B on page 39. The notable exception is the term τ appearing in the expression for $\widehat{\Delta}'$, which is a result of the set $\mathrm{LS}_{m,n}(r, s)$ not being closed without the constraint $\|L\|_F \leq \tau$ [26, Theorem 1.1].

To establish the proof of Theorem 1 we combine Lemma 2.1 and Lemma 2.2 with an ε-covering of $\mathrm{LS}_{m,n}(r, s, \tau)$, where ε will be picked to control the maximal allowed perturbation between the subspaces $\rho\left(\left(V_1, W_1\right), \left(V_2, W_2\right)\right)$. The *covering number* $\mathfrak{R}(\varepsilon)$ of $\mathrm{LS}_{m,n}(r, s, \tau)$ at resolution ε is the smallest number of subspaces (V_i, W_i, T_i) such that, for any triple of $V \in \mathcal{G}(m, r), W \in \mathcal{G}(n, r), T \in \mathcal{V}(mn, s)$ there exists i with $\rho\left(\left(V, W\right), \left(V_i, W_i\right)\right) \leq \varepsilon$ and $T = T_i$. The following Lemma gives an upper bound on the cardinality of ε-covering.

Lemma 2.3 (Covering number of $\mathrm{LS}_{m,n}(r, s, \tau)$)**.** *The covering number* $\mathfrak{R}(\varepsilon)$ *of the set* $\mathrm{LS}_{m,n}(r, s, \tau)$ *is bounded above by*

$$\mathfrak{R}(\varepsilon) \leq \binom{mn}{s}\left(\frac{4\pi}{\varepsilon}\right)^{r(m+n-2r)}. \tag{25}$$

The proof comes by counting the possible support sets with cardinality s and by the work of Szarek on ε-covering of the Grassmannian [33, Theorem 8], for completeness the proof is given in Appendix B, page 38.

Bounds on the RIC for the set of low-rank plus sparse matrices then follow a proof technique that uses the covering number argument in combination with the concentration of measure inequalities as was done before for sparse vectors [25] and subsequently for low-rank matrices [11].

Proof of Theorem 1 (RIC for $\mathrm{LS}_{m,n}(r, s, \tau)$), stated on page 4.

Proof. By linearity of $\mathcal{A}$ we can assume without loss of generality that $\|X\|_F = 1$. Let (V_i, W_i, T_i) be an ε-covering of $\mathrm{LS}_{m,n}(r, s, \tau)$ with the covering number $\mathfrak{N}(\varepsilon)$ bounded by Lemma 2.3. For every triple (V_i, W_i, T_i) define a subset of matrices

$$\mathrm{B}_i = \left\{X \in \Sigma_{m,n}\left(V, W, T_i, \tau\right) \,:\, \rho\left(\left(V, W\right), \left(V_i, W_i\right)\right) \leq \varepsilon\right\}. \tag{26}$$

By (V_i, W_i, T_i) being an ε-covering $\mathrm{LS}_{m,n}(r, s, \tau) \subseteq \bigcup_i \mathrm{B}_i$. Therefore, if for all B_i

$$(\forall X \in \mathrm{B}_i): \quad (1 - \widehat{\Delta})\|X\|_F \leq \|\,\mathcal{A}(X)\| \leq (1 + \widehat{\Delta})\|X\|_F \tag{27}$$

holds, then necessarily $\widehat{\Delta}_{r,s,\tau} \leq \widehat{\Delta}$, proving that

$$\Pr(\widehat{\Delta}_{r,s,\tau} \leq \widehat{\Delta}) = \Pr\Big(\forall X \in \mathrm{LS}_{m,n}(r, s, \tau) \,:\, (1 - \widehat{\Delta})\|X\|_F \leq \|\,\mathcal{A}(X)\| \leq (1 + \widehat{\Delta})\|X\|_F\Big) \tag{28}$$

$$\geq \Pr\Big((\forall i), (\forall X \in \mathrm{B}_i) \,:\, (1 - \widehat{\Delta})\|X\|_F \leq \|\,\mathcal{A}(X)\| \leq (1 + \widehat{\Delta})\|X\|_F\Big), \tag{29}$$

where the inequality comes from the fact that $\mathrm{LS}_{m,n}(r, s, \tau)$ is a subset of the ε-covering and therefore the statement holds with less or equal probability. It remains to derive a lower bound on the probability in the equation (29) which in turn proves the theorem.

In the case that $\|\,\mathcal{A}\,\| \geq \frac{\widehat{\Delta}}{2\tau\varepsilon} - 1$, which we show in (32) occurs with exponentially small probability, rearranging the terms gives that $\varepsilon \leq \widehat{\Delta}/\left(2\tau\left(1 + \|\,\mathcal{A}\,\|\right)\right)$ which guarantees also that

$$\tau\varepsilon(1 + \widehat{\Delta}/2 + \|\,\mathcal{A}\,\|) \leq \widehat{\Delta}/2. \tag{30}$$

If the RIC holds for a fixed (V_i, W_i, T_i) with $\widehat{\Delta}/2$, then by Lemma 2.2 in combination with (30) yields

$$(\forall X \in \mathrm{B}_i) \,:\, (1 - \widehat{\Delta})\|X\|_F \leq \|\,\mathcal{A}\,\| \leq (1 + \widehat{\Delta})\|X\|_F. \tag{31}$$

Therefore, using the probability union bound on (29) over all i's and the probability of $\|\,\mathcal{A}\,\|$ satisfying the bound $\varepsilon \leq \widehat{\Delta}/\left(2\tau\left(1 + \|\,\mathcal{A}\,\|\right)\right)$.

$$\Pr\Big((\forall i), (\forall X \in \mathrm{B}_i) \,:\, (1 - \widehat{\Delta})\|X\|_F \leq \|\,\mathcal{A}(X)\| \leq (1 + \widehat{\Delta})\|X\|_F\Big) \tag{32}$$

$$\geq 1 - \sum_i \Pr\Big(\exists Y \in \Sigma_{m,n}(V_i, W_i, T_i, \tau), \text{ s.t. } \|\,\mathcal{A}(Y)\| < (1 - \widehat{\Delta}/2) \text{ or } \|\,\mathcal{A}(Y)\| > (1 + \widehat{\Delta}/2)\Big) \tag{33}$$

$$- \Pr\Big(\ \|\mathcal{A}\| \geq \frac{\widehat{\Delta}}{2\tau\varepsilon} - 1\Big). \tag{34}$$

The probability in (33) is bounded from above as

$$\sum_i \Pr\Big(\exists Y \in \Sigma_{m,n}(V_i, W_i, T_i, \tau), \text{ s.t. } \|\,\mathcal{A}(Y)\| < (1 - \widehat{\Delta}/2) \text{ or } \|\,\mathcal{A}(Y)\| > (1 + \widehat{\Delta}/2)\Big) \tag{35}$$

$$\leq 2\,\mathfrak{N}(\varepsilon)\left(\frac{48}{\widehat{\Delta}}\tau\right)^{r^2}\left(\frac{48}{\widehat{\Delta}}(1 + \tau)\right)^{s}\exp\left(-\frac{p}{2}\left(\frac{\widehat{\Delta}^2}{32} - \frac{\widehat{\Delta}^3}{192}\right)\right) \tag{36}$$

$$\leq 2\binom{mn}{s}\left(\frac{4\pi}{\varepsilon}\right)^{r(m+n-2r)}\left(\frac{48}{\widehat{\Delta}}\tau\right)^{r^2}\left(\frac{48}{\widehat{\Delta}}(1 + \tau)\right)^{s}\exp\left(-\frac{p}{2}\left(\frac{\widehat{\Delta}^2}{32} - \frac{\widehat{\Delta}^3}{192}\right)\right), \tag{37}$$

where in the first inequality we used Lemma 2.1 and in the second inequality the bound on the ε-covering by Lemma 2.3.

In order to complete the lower bound in (32) it remains to upper bound (34) which we obtain by selecting the covering resolution ε sufficiently small so that the $\Pr\left(\|\mathcal{A}\| \geq \frac{\widehat{\Delta}}{2\tau\varepsilon} - 1\right)$ is exponentially small with exponent proportional to the bound in (37). From condition (5) of Definition 1.3 we have that the linear sensing map satisfies

$$(\exists \gamma > 0): \quad \Pr\left(\|\mathcal{A}\| \geq 1 + \sqrt{\frac{mn}{p}} + t\right) \leq \exp\left(-\gamma p t^2\right), \tag{38}$$

in particular

$$\Pr\left(\|\mathcal{A}\| \geq \frac{\widehat{\Delta}}{2\tau\varepsilon} - 1\right) \leq \exp\left(-\gamma p \left(\frac{\widehat{\Delta}}{2\tau\varepsilon} - \sqrt{\frac{mn}{p}} - 2\right)^2\right). \tag{39}$$

Selecting the covering resolution ε

$$\varepsilon < \frac{\widehat{\Delta}}{4\tau(\sqrt{mn/p} + 1)}, \tag{40}$$

obtains the following exponentially small upper bound

$$\Pr\left(\|\mathcal{A}\| \geq \frac{\widehat{\Delta}}{2\tau\varepsilon} - 1\right) \leq \exp\left(-\gamma mn\right). \tag{41}$$

Returning to the inequality (32), combined with the bound on the first term in (37) and the bound on the second term in (41) for ε satisfying (40) we have that

$$2\left(\frac{emn}{s}\right)^s \left(\frac{32\pi(\sqrt{mn/p}+1)}{\widehat{\Delta}}\right)^{r(m+n-2r)} \left(\frac{48}{\widehat{\Delta}}\tau\right)^{r^2} \left(\frac{48}{\widehat{\Delta}}(1+\tau)\right)^s \exp\left(-\frac{p}{2}\left(\frac{\widehat{\Delta}^2}{32} - \frac{\widehat{\Delta}^3}{192}\right)\right) \tag{42}$$

$$= \exp\Bigg(- pa(\widehat{\Delta}) + r\,(m+n-2r)\log\left(\sqrt{\frac{mn}{p}} + 1\right) + r\,(m+n-2r)\log\left(\frac{32\pi\tau}{\widehat{\Delta}}\right)$$
$$+ r^2 \log\left(\frac{48}{\widehat{\Delta}}\tau\right) + s\log\left(\frac{48}{\widehat{\Delta}}(1+\tau)\right) + s\log\left(\frac{emn}{s}\right)\Bigg), \tag{43}$$

where we used the inequality $\binom{mn}{s} \leq \left(\frac{emn}{s}\right)^s$ and we define $a(\widehat{\Delta}) := \widehat{\Delta}^2/64 - \widehat{\Delta}^3/384$. The 2^{nd}, 3^{rd} and 4^{th} terms in (43) can be bounded as

$$(\exists c_2 > 0): \quad 2^{nd} + 3^{rd} + 4^{th} \leq \left(c_2/a(\widehat{\Delta})\right) r(m+n-r)\log(mn/p), \tag{44}$$

and the 5^{th} and 6^{th} term of (43) as

$$(\exists c_3 > 0): \quad 5^{th} + 6^{th} \leq \left(c_3/a(\widehat{\Delta})\right) s\log(mn/s), \tag{45}$$

where c_2 and c_3 are dependent only on $\widehat{\Delta}$. Therefore there exists a positive constant c dependent only on $\widehat{\Delta}$ such that if $p \geq c_0\left(r(m+n-r)+s\right)\log(mn/s)$, then RIC holds with the constant $\widehat{\Delta}$ with probability at least $e^{-c_0 p}$. □

3. Provable recovery guarantees using computationally efficient algorithms

This section contains the proofs of our main algorithmic contributions that a low-rank plus sparse matrix $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ can be efficiently recovered from subsampled measurements taken by a linear mapping $\mathcal{A}(\cdot)$ which satisfies given bounds on its RIC. Subsection 3.1 presents the proof of Theorem 3 which shows that the convex relaxation (8) of (9) robustly recovers X_0. Subsection 3.2 states the proofs of Theorem 4 and Theorem 5 for the simple yet efficient hard thresholding algorithms NIHT and NAHT, described in Alg. 1 and Alg. 2 respectively.

3.1. Recovery of $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ *using the convex relaxation* (8).

Let $X_* = L_* + S_*$ be the solution of the convex optimization problem formulated in (8). Here it is shown that if the RICs of the measurement operator $\mathcal{A}(\cdot)$ are sufficient small, then $X_* = X_0$ when the linear constraint in the convex optimization problem (8) is satisfied exactly, or alternatively that $\|X_* - X_0\|_F$ is proportional to $\|\mathcal{A}(X^*) - b\|_2$.

Proof of Theorem 3 (Guaranteed recovery by the convex relaxation (8)), stated on page 5.

Proof. Let $R = X_* - X_0 = (L_* - L_0) + (S_* - S_0) = R^L + R^S$ be the residual split into the low-rank component $R^L = L_* - L_0$ and the sparse component $R^S = S_* - S_0$. We treat R^L and R^S separately, combining the method of proof used in the context of compressed sensing by [34] and its extension for the low-rank matrix recovery by [11].

By Lemma Appendix B.2 on page 40 there exist matrices $R^L_0, R^L_c \in \mathbb{R}^{m\times n}$ such that $R^L = R^L_0 + R^L_c$ and

$$\mathrm{rank}(R^L_0) \leq 2r \tag{46}$$

$$L_0 (R^L_c)^T = 0_{m\times m} \quad \text{and} \quad L_0^T R^L_c = 0_{n\times n}. \tag{47}$$

Similarly, by the argument made in the proof of [34, Theorem 1], stated in the appendices as Lemma Appendix B.3 on page 41, there exist matrices $R^S_0, R^S_c \in \mathbb{R}^{m\times n}$ such that $R^S = R^S_0 + R^S_c$ and

$$\left\|R^S_0\right\|_0 \leq s \tag{48}$$

$$\mathrm{supp}\left(S_0\right) \cap \mathrm{supp}\left(R^S_c\right) = \emptyset. \tag{49}$$

By (L_*, S_*) being a minimum and X_0 being feasible of the convex optimization problem (8)

$$\|L_0\|_* + \lambda \|S_0\|_1 \geq \|L_*\|_* + \lambda \|S_*\|_1 \tag{50}$$

$$= \left\|L_0 + R^L_0 + R^L_c\right\|_* + \lambda \left\|S_0 + R^S_0 + R^S_c\right\|_1 \tag{51}$$

$$\geq \left\|L_0 + R^L_c\right\|_* - \left\|R^L_0\right\|_* + \lambda \left\|S_0 + R^S_c\right\|_1 - \lambda \left\|R^S_0\right\|_1 \tag{52}$$

$$= \|L_0\|_* + \left\|R^L_c\right\|_* - \left\|R^L_0\right\|_* + \lambda \|S_0\|_1 + \lambda \left\|R^S_c\right\|_1 - \lambda \left\|R^S_0\right\|_1, \tag{53}$$

where the second line comes from $L_* - L_0 = R_0^L + R_c^L$ and $S_* - S_0 = R_0^S + R_c^S$, the inequality in the third line comes from the reverse triangle inequality, and the fourth line comes from the construction of R_c^L and R_c^S combined with [11, Lemma 2.3], restated as Corollary Appendix B.1, and by $\mathrm{supp}(R_c^S) \cap \mathrm{supp}(R_0^S) = \emptyset$. Subtracting $\|L_0\|_*$ and $\|S_0\|_1$ from both sides of (53) and rearranging terms yields

$$\left\|R_c^L\right\|_* + \lambda \left\|R_c^S\right\|_1 \leq \left\|R_0^L\right\|_* + \lambda \left\|R_0^S\right\|_1 . \tag{54}$$

We proceed by decomposing the remainder terms R_c^L and R_c^S as sums of matrices with decreasing energy as was done in [11] for low-rank matrices and in [34] for sparse vectors. Let $R_c^L = U\mathrm{diag}(\sigma)V^T$ be the singular value decomposition of R_c^L and split the indices of the singular values into sets of size M_r as

$$I_i := \{(i-1)M_r + 1, \ldots, iM_r\} . \tag{55}$$

Constructing $R_i^L := U_{I_i}\mathrm{diag}(\sigma_{I_i})V_{I_i}^T$ decomposes R_c^L into a sum $R_c^L = R_1^L + R_2^L + \ldots$ such that

$$\mathrm{rank}\left(R_i^L\right) \leq M_r, \qquad \forall i \geq 1 \tag{56}$$

$$R_i^L \left(R_j^L\right)^T = 0_{m\times m} \quad \text{and} \quad \left(R_i^L\right)^T R_j^L = 0_{n\times n}, \quad \forall i \neq j \tag{57}$$

$$\sigma_k \leq \frac{1}{M_r} \sum_{j\in I_i} \sigma_j, \qquad \forall k \in I_{i+1} \tag{58}$$

where the inequality (58) implies that $\left\|R_{i+1}^L\right\|_F^2 \leq \frac{1}{M_r} \left\|R_i^L\right\|_*^2$.

Similarly, order the indices of R_c^S as $v_1, v_2, \ldots, v_{mn} \in [m] \times [n]$ in decreasing order of magnitude of the entries of R_c^S and split the indices of the entries into sets of size M_s as

$$T_i := \{v_\ell \, : \, (i-1)M_s \leq \ell \leq iM_s\} , \tag{59}$$

Constructing $R_i^S := \left(R_c^S\right)_{T_i}$ decomposes R_c^S into a sum $R_c^S = R_1^S + R_2^S + \ldots$ such that

$$\left\|R_i^S\right\|_0 \leq M_s, \qquad \forall i \geq 1 \tag{60}$$

$$\emptyset = T_i \cap T_j, \qquad \forall i \neq j \tag{61}$$

$$\left|R_c^S\right|_{(v)} \leq \frac{1}{\sqrt{M_s}} \sum_{j\in T_i} \left|R_i^S\right|_{(j)} , \qquad \forall v \in T_{i+1} \tag{62}$$

where the inequality (62) implies that $\left\|R_{i+1}^S\right\|_F^2 \leq \frac{1}{M_s} \left\|R_i^S\right\|_1^2$. Combining the two decompositions of

R_c^L and R_c^S gives the following bound

$$
\begin{aligned}
\left\|R_c^L + R_c^S\right\|_F &\leq \sum_{j\geq 2} \left\|R_j^L\right\|_F + \sum_{j\geq 2} \left\|R_j^S\right\|_F && (63)\\
&\leq \sqrt{\frac{1}{M_r}} \sum_{j\geq 1} \left\|R_j^L\right\|_* + \sqrt{\frac{1}{M_s}} \sum_{j\geq 1} \left\|R_j^S\right\|_1 && (64)\\
&= \sqrt{\frac{1}{M_r}} \left\|R_c^L\right\|_* + \sqrt{\frac{1}{M_s}} \left\|R_c^S\right\|_1 && (65)\\
&\leq \frac{1}{\sqrt{M_r}} \left(\left\|R_0^L\right\|_* + \sqrt{\frac{M_r}{M_s}} \left\|R_0^S\right\|_1 \right) && (66)\\
&\leq \sqrt{\frac{2r}{M_r}} \left\|R_0^L\right\|_F + \sqrt{\frac{s}{M_s}} \left\|R_0^S\right\|_F , && (67)
\end{aligned}
$$

where the inequality in the first line comes from the triangle inequality, the second inequality comes as a consequence of (58) and (62), the third line comes from (57) combined with [11, Lemma 2.3], restated as Corollary Appendix B.1, and from (61), the fourth inequality comes from (54) with $\lambda = \sqrt{M_r/M_s}$, and the last fifth line is a property of ℓ_1 and Schatten-1 norms. Choosing $M_r = r$ and $M_s = s$ in (67) gives

$$
\left\|R_c^L + R_c^S\right\|_F \leq \sqrt{2}\left\|R_0^L\right\|_F + \left\|R_0^S\right\|_F , \tag{68}
$$

and also that $\lambda = \sqrt{r/s}$ as stated in the theorem.

By feasibility of X^* and linearity of $\mathcal{A}$ we have

$$
\varepsilon_{\mathrm{b}} \geq \left\|\mathcal{A}\left(X^*\right) - b\right\|_2 = \left\|\mathcal{A}\left(X^* - X_0\right)\right\|_2 = \left\|\mathcal{A}\left(R\right)\right\|_2 . \tag{69}
$$

Let $\Delta := \Delta_{4r,3s,2\tau}$ be an RIC with squared norms for $\mathrm{LS}_{m,n}(4r, 3s, 2\tau)$. Then

$$
\begin{aligned}
(1-\Delta)\|R_0^L + R_1^L\|_F^2 &\leq \left\|\mathcal{A}\left(R_0^L + R_1^L\right)\right\|_2^2 = \left|\left\langle \mathcal{A}(R_0^L + R_1^L),\, \mathcal{A}(R_0^L + R_1^L - R + R)\right\rangle\right| && (70)\\
&= \left|\left\langle \mathcal{A}(R_0^L + R_1^L),\, \mathcal{A}(R_0^L + R_1^L - R)\right\rangle + \left\langle \mathcal{A}(R_0^L + R_1^L),\, \mathcal{A}(R)\right\rangle\right| && (71)\\
&\leq \left|\left\langle \mathcal{A}\left(R_0^L + R_1^L\right),\, \mathcal{A}\left(-R_0^S - R_1^S - \sum_{j\geq 2} R_j\right)\right\rangle\right| + \left|\left\langle \mathcal{A}(R_0^L + R_1^L),\, \mathcal{A}(R)\right\rangle\right| && (72)\\
&\leq \Delta \left\|R_0^L + R_1^L\right\|_F \left(\left\|R_0^S + R_1^S\right\|_F + \sum_{j\geq 2} \left\|R_j\right\|_F \right) + \left\|\mathcal{A}\left(R_0^L + R_1^L\right)\right\|_2 \left\|\mathcal{A}\left(R\right)\right\|_2 , && (73)
\end{aligned}
$$

where the inequality in the first line comes from $R_0^L + R_1^L \in \mathrm{LS}_{m,n}(4r, 3s, 2\tau)$ satisfying the RIC, the second line is a consequence of feasibility in (69), and the third line comes from Lemma Appendix B.4 and by sums of individual pairs in the inner product being in $\mathrm{LS}_{m,n}(4r, 3s, 2\tau)$.

The first term in (73) can be bounded as

$$\Delta \left\|R_0^L + R_1^L\right\|_F \left(\left\|R_0^S + R_1^S\right\|_F + \sum_{j\geq 2} \|R_j\|_F\right) \tag{74}$$

$$\leq \Delta \left\|R_0^L + R_1^L\right\|_F \left(\left\|R_0^S + R_1^S\right\|_F + \sqrt{2}\left\|R_0^L\right\|_F + \left\|R_0^S\right\|_F\right) \tag{75}$$

$$\leq \Delta \left\|R_0^L + R_1^L\right\|_F \left(2\left\|R_0^S + R_1^S\right\|_F + \sqrt{2}\left\|R_0^L + R_1^L\right\|_F\right) \tag{76}$$

where the second line comes as a consequence of optimality in (68) with $M_r = r$ and $M_s = s$, and the third line comes from $\left\|R_0^L\right\|_F \leq \left\|R_0^L + R_1^L\right\|_F$ and $\left\|R_1^L\right\|_F \leq \left\|R_0^L + R_1^L\right\|_F$. The upper bound of the second term in (73) is a consequence of feasibility bound in (69) and of the RIC for $R_0^L + R_1^L \in \mathrm{LS}_{m,n}(4r, 3s, 2\tau)$

$$\left\|\mathcal{A}\left(R_0^L + R_1^L\right)\right\|_2 \|\,\mathcal{A}\,(R)\,\|_2 \leq \varepsilon_{\mathrm{b}}(1+\Delta)\|R_0^L + R_1^L\|_F. \tag{77}$$

Combining inequality (76) and (77) yields an upper bound of (73)

$$(1-\Delta)\left\|R_0^L + R_1^L\right\|_F^2 \leq \Delta \left\|R_0^L + R_1^L\right\|_F \left(2\left\|R_0^S + R_1^S\right\|_F + \sqrt{2}\left\|R_0^L + R_1^L\right\|_F + \varepsilon_{\mathrm{b}}\,\frac{1+\Delta}{\Delta}\right), \tag{78}$$

which after dividing both sides by $(1-\Delta)\left\|R_0^L + R_1^L\right\|_F$ gives

$$\left\|R_0^L + R_1^L\right\|_F \leq \frac{\Delta}{1-\Delta}\left(2\left\|R_0^S + R_1^S\right\|_F + \sqrt{2}\left\|R_0^L + R_1^L\right\|_F\right) + \varepsilon_{\mathrm{b}}\,\frac{1+\Delta}{1-\Delta}. \tag{79}$$

Mutatis mutandis, the same argument applies to $\left\|R_0^S + R_1^S\right\|_F$ (for details, see Remark Appendix B.1)

$$\left\|R_0^S + R_1^S\right\|_F \leq \frac{\Delta}{1-\Delta}\left(\left(1+\sqrt{2}\right)\left\|R_0^L + R_1^L\right\|_F + \left\|R_0^S + R_1^S\right\|_F\right) + \varepsilon_{\mathrm{b}}\,\frac{1+\Delta}{1-\Delta}. \tag{80}$$

Adding (79) and (80) together

$$\left\|R_0^L + R_1^L\right\|_F + \left\|R_0^S + R_1^S\right\|_F \leq \frac{\Delta}{1-\Delta}\left(\left(1+2\sqrt{2}\right)\left\|R_0^L + R_1^L\right\|_F + 3\left\|R_0^S + R_S^L\right\|_F\right) + 2\,\varepsilon_{\mathrm{b}}\,\frac{1+\Delta}{1-\Delta}. \tag{81}$$

For $\Delta \leq \frac{1}{5}$ the prefactor $\frac{\Delta}{1-\Delta} \leq \frac{1}{4}$ and (81) is upper bounded as

$$\left\|R_0^L + R_1^L\right\|_F + \left\|R_0^S + R_1^S\right\|_F \leq \frac{1+2\sqrt{2}}{4}\left\|R_0^L + R_1^L\right\|_F + \frac{3}{4}\left\|R_0^S + R_1^S\right\|_F + 3\,\varepsilon_{\mathrm{b}}, \tag{82}$$

The maximum of $\left\|R_0^L + R_1^L\right\|_F + \left\|R_0^S + R_1^S\right\|_F$ over the constraints given by the inequality in (82) is attained when

$$\left\|R_0^L + R_1^L\right\|_F + \left\|R_0^S + R_1^S\right\|_F = \frac{3\,\varepsilon_{\mathrm{b}}}{3-2\sqrt{2}}. \tag{83}$$

By orthogonality from construction in (57) and (61) we have that

$$\frac{3\,\varepsilon_{\mathrm{b}}}{3-2\sqrt{2}} \geq \left\|R_0^L + R_1^L\right\|_F + \left\|R_0^S + R_1^S\right\|_F = \left\|R_0^L\right\|_F + \left\|R_0^S\right\|_F \tag{84}$$

$$\geq \frac{1}{\sqrt{2}}\sum_{j\geq 2}\|R_j\|_F \geq \frac{1}{\sqrt{2}}\|R_c\|_F\,, \tag{85}$$

where the inequality in the second line comes from (68). Applying the triangle inequality on $R = R_0^L + R_0^S + R_c$ concludes the proof

$$\|R\|_F = \left\|R_0^L + R_0^S + R_c\right\|_F \le \left\|R_0^L\right\|_F + \left\|R_0^S\right\|_F + \|R_c\|_F \le 3\,\varepsilon_{\mathrm{b}}\,\frac{1+\sqrt{2}}{3-2\sqrt{2}} \le 67\,\varepsilon_{\mathrm{b}}\,. \tag{86}$$

□

3.2. Recovery of $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ by Alg. 1 and Alg. 2.

This section presents the proofs of Theorem 4 and 5, that NIHT and NAHT respectively recover $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ from $\mathcal{A}(X_0)$ and knowledge of (r,s,τ) provided the RICs of $\mathcal{A}(\cdot)$ are sufficiently bounded.

The proof of NIHT follows the same line of thought as the one for low-rank matrix completion [28], with the only difference of the hard thresholding projection, in the form of RPCA, being an imprecise projection with accuracy ε_{p} as stated in (13). The proof consists of deriving an inequality where $\|X^{j+1} - X_0\|_F$ is bounded by a factor multiplying $\|X^j - X_0\|_F$, and then showing that this multiplicative factor is strictly less then one if $\mathcal{A}$ satisfies RIC with $\Delta_3 := \Delta_{r,s,\tau}(\mathcal{A}) < 1/5$.

Proof of Theorem 4 (Guaranteed recovery by NIHT, Alg. 1).

Proof. Let $b = \mathcal{A}(X_0)$ be the vector of measurements of the matrix $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ and $W^j = X^j - \alpha_j^L\,\mathcal{A}^*\left(\mathcal{A}(X^j) - b\right)$ to be the update of X^j before the oblique Robust PCA projection step $X^{j+1} = \mathrm{RPCA}_{r,s,\tau}(W^j, \varepsilon_{\mathrm{p}})$. By X^{j+1} being within an ε_{p} distance in the Frobenius norm of the optimal RPCA projection $X^{j+1}_{\mathrm{rpca}} := \mathrm{RPCA}_{r,s,\tau}(W^j, 0)$ defined in (13)

$$\left\|W^j - X^{j+1}\right\|_F^2 = \left\|W^j - X^{j+1}_{\mathrm{rpca}} + X^{j+1}_{\mathrm{rpca}} - X^{j+1}\right\|_F^2 \tag{87}$$

$$\le \left(\left\|W^j - X^{j+1}_{\mathrm{rpca}}\right\|_F + \left\|X^{j+1} - X^{j+1}_{\mathrm{rpca}}\right\|_F\right)^2 \tag{88}$$

$$\le \left(\left\|W^j - X_0\right\|_F + \varepsilon_{\mathrm{p}}\right)^2, \tag{89}$$

where in the second line we used the triangle inequality, and the third line comes from X^{j+1}_{rpca} being the optimal projection thus being the closest matrix in $\mathrm{LS}_{m,n}(r,s,\tau)$ to W^j in the Frobenius norm and by X^{j+1} being within ε_{p} distance of X^{j+1}_{rpca}. By expansion of the left hand side of (87)

$$\left\|W^j - X^{j+1}\right\|_F^2 = \left\|W^j - X_0 + X_0 - X^{j+1}\right\|_F^2 \tag{90}$$

$$= \left\|W^j - X_0\right\|_F^2 + \left\|X_0 - X^{j+1}\right\|_F^2 + 2\left\langle W^j - X_0,\, X_0 - X^{j+1}\right\rangle \tag{91}$$

$$= \left(\left\|W^j - X_0\right\|_F + \varepsilon_{\mathrm{p}}\right)^2 \le \left\|W^j - X_0\right\|_F^2 + 2\,\varepsilon_{\mathrm{p}}\left\|W^j - X_0\right\|_F + {\varepsilon_{\mathrm{p}}}^2 \tag{92}$$

where in the last line (92) follows from the inequality in (89). Subtracting $\|W^j - X_0\|_F^2$ from both sides of (92) gives

$$\left\|X^{j+1} - X_0\right\|_F^2 \le 2\left\langle W^j - X_0,\, X^{j+1} - X_0\right\rangle + 2\,\varepsilon_{\mathrm{p}}\left\|W^j - X_0\right\|_F + {\varepsilon_{\mathrm{p}}}^2\,. \tag{93}$$

The matrix W^j in the inner product on the right hand side of (93) can be expressed using the update rule $W^j = X^j - \alpha_j \mathcal{A}^* \left(\mathcal{A}\left(X^j\right) - b\right)$

$$
\begin{aligned}
2\left\langle W^j - X_0,\, X^{j+1} - X_0 \right\rangle &= 2\langle X^j - X_0, X^{j+1} - X_0\rangle - 2\,\alpha_j \left\langle \mathcal{A}^* \mathcal{A}\left(X^j - X_0\right),\, X^{j+1} - X_0 \right\rangle && (94)\\
&= 2\left\langle X^j - X_0,\, X^{j+1} - X_0 \right\rangle - 2\,\alpha_j \left\langle \mathcal{A}\left(X^j - X_0\right),\, \mathcal{A}\left(X^{j+1} - X_0\right)\right\rangle && (95)\\
&\leq 2\left\|I - \alpha_j\, A_Q^* A_Q\right\|_2 \left\|X^j - X_0\right\|_F \left\|X^{j+1} - X_0\right\|_F, && (96)
\end{aligned}
$$

where in the first line[3] we use $b = \mathcal{A}(X_0)$ and linearity of $\mathcal{A}$, in the second line we split the inner product into two inner products by linearity of $\mathcal{A}$, and the inequality in the third line is a consequence of Lemma Appendix B.5.

The matrix W^j can be expressed using the update rule $W^j = X^j - \alpha_j \mathcal{A}^* \left(\mathcal{A}\left(X^j\right) - b\right)$ in the second term of the right hand side of (93) and upper bounded by Lemma Appendix B.5

$$
\begin{aligned}
\left\|W^j - X_0\right\|_F &= \left\|X^j - X_0 + \alpha_j \mathcal{A}^* \left(\mathcal{A}\left(X^j - X_0\right)\right)\right\|_2 && (97)\\
&\leq \left\|I - \alpha_j\, A_Q^* A_Q\right\|_2 \left\|X^j - X_0\right\|_F. && (98)
\end{aligned}
$$

By Lemma Appendix B.5, the eigenvalues of $\left(I - \alpha_j\, A_Q^* A_Q\right)$ are bounded by

$$
1 - \alpha_j\left(1 + \Delta_3\right) \leq \lambda\left(I - \alpha_j\, A_Q^* A_Q\right) \leq 1 + \alpha_j\left(1 - \Delta_3\right), \tag{99}
$$

where $\Delta_3 := \Delta_{3r,3s,2\tau}$.

Consider the stepsize computed in Algorithm 1, Line 3 inspired by the previous work on NIHT in the context of compressed sensing [27] and low-rank matrix sensing [28]

$$
\alpha_j = \frac{\left\|\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right)\right\|_F}{\left\|\mathcal{A}\left(\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right)\right)\right\|_2} \tag{100}
$$

where the projection $\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right)$ ensures that the residual R^j is projected onto the set $\mathrm{LS}_{m,n}(r,s,\tau)$. Then we can bound α_j using the RIC of $\mathcal{A}$ as

$$
\frac{1}{1 + \Delta_1} \leq \alpha_j \leq \frac{1}{1 - \Delta_1}, \tag{101}
$$

where $\Delta_1 := \Delta_{r,s,\tau}$. Combining (99) with (101) gives

$$
1 - \frac{1 + \Delta_3}{1 - \Delta_1} \leq \lambda\left(I - \alpha_j\, A_Q^* A_Q\right) \leq 1 - \frac{1 - \Delta_3}{1 + \Delta_1}. \tag{102}
$$

Since $\Delta_3 \geq \Delta_1$, the magnitude of the lower bound in (102) is greater than the upper bound. Therefore

$$
\eta := 2\left(\frac{1 + \Delta_3}{1 - \Delta_1} - 1\right) \geq 2\left\|I - \alpha_j\, A_Q^* A_Q\right\|_2, \tag{103}
$$

[3]Here it would be possible to extend the result to be stable under measurement error ε_{b} as done in Theorem 3 by adding an error term in (94).

where the constant η is strictly smaller than 1 if $\Delta_3 < 1/5$.

Finally, the error in (93) can be upper bounded by (96) combined with (98) with η being the upper bound on the operator norm in (103)

$$\left\|X^{j+1} - X_0\right\|_F^2 \leq \eta \left\|X^j - X_0\right\|_F \left\|X^{j+1} - X_0\right\|_F + \eta\, \varepsilon_{\mathrm{p}} \left\|X^j - X_0\right\|_F + {\varepsilon_{\mathrm{p}}}^2 . \tag{104}$$

It remains to show the inequality (104) implies the update rule contracts the error and the iterates X^j converge to a matrix $\hat{X}$ that is within a small in the Frobenius norm from X_0 depending on the precision ε_{p} of the RPCA. Rewrite (104) using the notation $e^j := \|X^j - X_0\|_F$

$$\left(e^{j+1}\right)^2 \leq \eta\, e^j\, e^{j+1} + \eta\ \varepsilon_{\mathrm{p}}\ e^j + {\varepsilon_{\mathrm{p}}}^2 \tag{105}$$

$$e^{j+1} \leq \eta\, e^j + \eta\ \varepsilon_{\mathrm{p}}\ \frac{e^j}{e^{j+1}} + \frac{{\varepsilon_{\mathrm{p}}}^2}{e^{j+1}}. \tag{106}$$

In the following, assume $\frac{\varepsilon_{\mathrm{p}}}{1-\eta} \leq e^{j+1}$ and upper bound the right hand side of (106) as

$$\begin{aligned} e^{j+1} &\leq \eta\, e^j + \eta\ \varepsilon_{\mathrm{p}}\ \frac{e^j}{e^{j+1}} + \frac{{\varepsilon_{\mathrm{p}}}^2}{e^{j+1}} & (107)\\ &= \eta\, e^j + \eta(1-\eta)e^j \frac{\varepsilon_{\mathrm{p}}}{(1-\eta)e^{j+1}} + \varepsilon_{\mathrm{p}}(1-\eta)\frac{\varepsilon_{\mathrm{p}}}{(1-\eta)e^{j+1}} & (108)\\ &\leq \eta\, e^j + \eta(1-\eta)e^j + (1-\eta)\,\varepsilon_{\mathrm{p}} & (109)\\ &< e^j \eta, & (110) \end{aligned}$$

where the inequality (109) is a consequence of $e^{j+1} \geq \varepsilon_{\mathrm{p}}/(1-\eta)$ and the inequality (110) holds if $e^j > \varepsilon_{\mathrm{p}}/(1-\eta)$. Therefore, if $e^j > \varepsilon_{\mathrm{p}}/(1-\eta)$ then the error sequence is contractive because $\eta < 1$ by $\Delta_3 < 1/5$ and has a fixed point $e^* = \varepsilon_{\mathrm{p}}/(1-\eta) = \varepsilon_{\mathrm{p}}\, \frac{1-\Delta_3}{1-5\Delta_3}$. Moreover, by $\Delta_3 \geq \Delta_1$, the equation in (109) is upper bounded as

$$e^{j+1} \leq 8\frac{\Delta_3\,(1-3\Delta_3)}{(1-\Delta_3)^2} e^j + \frac{1-5\Delta_3}{1-\Delta_3}\,\varepsilon_{\mathrm{p}}, \tag{111}$$

which gives an upper bound on the rate of convergence. □

Proof of Theorem 5 (Guaranteed recovery of NAHT, Alg. 2).

Proof. Let $b = \mathcal{A}(X_0)$ be the vector of measurements[4] of the matrix $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ and $V^j = L^j - \alpha_j^L\, \mathcal{A}^* \left(\mathcal{A}(X^j) - b\right)$ to be the update of L^j before the rank r projection $L^{j+1} = \mathrm{HT}_r(V^j)$. As a consequence of L^{j+1} being the closest rank r matrix to V^j in the Frobenius norm we have that

$$\begin{aligned} \left\|V^j - L_0\right\|_F^2 \geq \left\|V^j - L^{j+1}\right\|_F^2 &= \left\|V^j - L_0 + L_0 - L^{j+1}\right\|_F^2 \\ &= \left\|V^j - L_0\right\|_F^2 + \left\|L_0 - L^{j+1}\right\|_F^2 + 2\left\langle V^j - L_0,\, L_0 - L^{j+1}\right\rangle . \end{aligned} \tag{112}$$

[4]Again, it is possible to extend the result to the case when there is a measurement error ε_{b} as done in Theorem 3 by having $b = \mathcal{A}(X_0) + e$, with $\|e\|_2 \leq \varepsilon_{\mathrm{b}}$.

Subtracting $\left\|V^j - L_0\right\|_F^2$ from both sides of (112) and rearranging terms gives

$$\begin{aligned}
\left\|L_0 - L^{j+1}\right\|_F^2 &\leq 2\left\langle V^j - L_0,\, L^{j+1} - L_0\right\rangle & (113)\\
&= 2\left\langle L^j - \alpha_j^L\,\mathcal{A}^*\left(\mathcal{A}\left(X^j - X_0\right)\right) - L_0,\, L^{j+1} - L_0\right\rangle & (114)\\
&= 2\left\langle L^j - L_0 - \alpha_j^L\,\mathcal{A}^*\left(\mathcal{A}\left(L^j - L_0 + S^j - S_0\right)\right),\, L^{j+1} - L_0\right\rangle & (115)\\
&= 2\left\langle L^j - L_0,\, L^{j+1} - L_0\right\rangle - 2\,\alpha_j^L\left\langle\mathcal{A}\left(L^j - L_0\right),\,\mathcal{A}\left(L^{j+1} - L_0\right)\right\rangle\\
&\quad - 2\,\alpha_j^L\left\langle\mathcal{A}\left(S^j - S_0\right),\,\mathcal{A}\left(L^{j+1} - L_0\right)\right\rangle & (116)\\
&\leq 2\left\|I - \alpha_j^L\,A_Q^*A_Q\right\|\left\|L^j - L_0\right\|_F\left\|L^{j+1} - L_0\right\|_F\\
&\quad + 2\,\alpha_j^L\,\Delta_3\left\|S^j - S_0\right\|_F\left\|L^{j+1} - L_0\right\|_F\,, & (117)
\end{aligned}$$

where in the second line we expanded V^j using the update rule $V^j = L^j - \alpha_j^L\,\mathcal{A}\left(\mathcal{A}(X^j) - b\right)$ and $b = \mathcal{A}(X_0)$, in the third line we expanded $X^j = L^j + S^j$, in the fourth line we split the inner product into two inner products by linearity of $\mathcal{A}$, and in the last line the inequality comes from Lemma Appendix B.5 bounding the first two terms and Appendix B.4 bounding the third term with $\Delta_3 := \Delta_{3r,3s,2\tau}$. Dividing both sides of (117) by $\left\|L_0 - L^{j+1}\right\|_F$ gives

$$\left\|L_0 - L^{j+1}\right\|_F \leq 2\left\|I - \alpha_j^L\,A_Q^*A_Q\right\|\,\left\|L^j - L_0\right\|_F + 2\,\alpha_j^L\,\Delta_3\left\|S^j - S_0\right\|_F. \tag{118}$$

Let $W^j = S^j - \alpha_j^S\,\mathcal{A}^*\left(\mathcal{A}(X^{j+\frac{1}{2}}) - b\right)$ be the subsequent update of S^j before the s sparse projection $S^{j+1} = \mathrm{HT}_s(W^j)$. By S^{j+1} being the closest s sparse matrix to W^j in the Frobenius norm and by $\|S_0\|_0 \leq s$, it follows that

$$\begin{aligned}
\left\|W^j - S_0\right\|_F^2 \geq \left\|W^j - S^{j+1}\right\|_F^2 &= \left\|W^j - S_0 + S_0 - S^{j+1}\right\|_F^2\\
&= \left\|W^j - S_0\right\|_F^2 + \left\|S_0 - S^{j+1}\right\|_F^2 + 2\left\langle W^j - S_0,\, S_0 - S^{j+1}\right\rangle. & (119)
\end{aligned}$$

Subtracting $\left\|W^j - S_0\right\|_F^2$ from both sides in (119) and rearranging terms gives

$$\begin{aligned}
\left\|S_0 - S^{j+1}\right\|_F^2 &\leq 2\left\langle W^j - S_0,\, S^{j+1} - S_0\right\rangle & (120)\\
&= 2\left\langle S^j - \alpha_j^S\,\mathcal{A}^*\left(\mathcal{A}\left(X^{j+\frac{1}{2}} - X_0\right)\right) - S_0,\, S^{j+1} - S_0\right\rangle & (121)\\
&= 2\left\langle S^j - S_0 - \alpha_j^S\,\mathcal{A}^*\left(\mathcal{A}\left(L^{j+1} - L_0 + S^j - S_0\right)\right),\, S^{j+1} - S_0\right\rangle & (122)\\
&= 2\left\langle S^j - S_0,\, S^{j+1} - S_0\right\rangle - 2\,\alpha_j^S\left\langle\mathcal{A}\left(S^j - S_0\right),\,\mathcal{A}\left(S^{j+1} - S_0\right)\right\rangle\\
&\quad - 2\,\alpha_j^S\left\langle\mathcal{A}\left(L^{j+1} - L_0\right),\,\mathcal{A}\left(S^{j+1} - S_0\right)\right\rangle & (123)\\
&\leq 2\left\|I - \alpha_j^S\,A_Q^*A_Q\right\|\left\|S^j - S_0\right\|_F\left\|S^{j+1} - S_0\right\|_F\\
&\quad + 2\,\alpha_j^S\,\Delta_3\left\|L^{j+1} - L_0\right\|_F\left\|S^{j+1} - S_0\right\|_F\,, & (124)
\end{aligned}$$

where in the second line we express W^j using the update rule $W^j = S^j - \alpha_j^L\,\mathcal{A}\left(\mathcal{A}(X^{j+\frac{1}{2}}) - b\right)$ and $b = \mathcal{A}(X_0)$, in the third line we expanded $X^{j+\frac{1}{2}} = L^{j+1} + S^j$, in the fourth line we split the inner product into two inner products by linearity of $\mathcal{A}$, and the inequality in the last line comes from Lemma Appendix B.5 bounding the first two terms and Appendix B.4 bounding the third term with $\Delta_3 := \Delta_{3r,3s,2\tau}$. Dividing both sides of (124) by $\left\|L_0 - L^{j+1}\right\|_F$ gives

$$\left\|S_0 - S^{j+1}\right\|_F \leq 2\left\|I - \alpha_j^S\, A_Q^* A_Q\right\| \left\|S^j - S_0\right\|_F + 2\,\alpha_j^L\, \Delta_3 \left\|L^{j+1} - L_0\right\|_F . \tag{125}$$

Adding together (118) and (125)

$$\begin{aligned}\left\|L_0 - L^{j+1}\right\|_F + \left\|S_0 - S^{j+1}\right\|_F \leq 2\left\|I - \alpha_j^L\, A_Q^* A_Q\right\| \left\|L^j - L_0\right\|_F + 2\,\alpha_j^L\, \Delta_3 \left\|S^j - S_0\right\|_F \\ + 2\left\|I - \alpha_j^S\, A_Q^* A_Q\right\| \left\|S^j - S_0\right\|_F + 2\,\alpha_j^L\, \Delta_3 \left\|L^{j+1} - L_0\right\|_F ,\end{aligned} \tag{126}$$

which after rearranging terms in (126) becomes

$$\begin{aligned}\left(1 - 2\,\alpha_j^S\, \Delta_3\right) \|L_0 - L^{j+1}\|_F + \|S_0 - S^{j+1}\|_F \leq 2\|I - \alpha_j^L\, A_Q^* A_Q\|\, \|L^j - L_0\|_F \\ +2\left(\|I - \alpha_j^S\, A_Q^* A_Q\| + \alpha_j^L\, \Delta_3\right) \|S^j - S_0\|_F\end{aligned} \tag{127}$$

and because $\alpha_j^S, \alpha_j^L, \Delta_3 \geq 0$, subtracting $2\,\alpha_j^S\, \Delta_3 \|S_0 - S^{j+1}\|_F$ on the left does not increase the left hand side while adding $2\,\alpha_j^L\, \Delta_3 \left\|L^j - L_0\right\|_F$ on the right does not decrease the right hand side of (127), therefore

$$\begin{aligned}\left(1 - 2\,\alpha_j^S\, \Delta_3\right) \left(\|L_0 - L^{j+1}\|_F + \|S_0 - S^{j+1}\|_F\right) \\ \leq 2\left(\|I - \alpha_j^S\, A_Q^* A_Q\| + \alpha_j^L\, \Delta_3\right) \left(\|L^j - L_0\|_F + \|S^j - S_0\|_F\right) ,\end{aligned} \tag{128}$$

Dividing both sides of (128) by $\left(1 - 2\,\alpha_j^S\, \Delta_3\right)$ simplifies to

$$\left\|L_0 - L^{j+1}\right\|_F + \left\|S_0 - S^{j+1}\right\|_F \leq 2\frac{\left\|I - \alpha_j^S\, A_Q^* A_Q\right\| + \alpha_j^L\, \Delta_3}{1 - 2\,\alpha_j^S\, \Delta_3} \left(\left\|L^j - L_0\right\|_F + \left\|S^j - S_0\right\|_F\right) . \tag{129}$$

By Lemma Appendix B.5, the eigenvalues of $\left(I - \alpha_j\, A_Q^* A_Q\right)$ can be bounded as

$$1 - \alpha_j \left(1 + \Delta_3\right) \leq \lambda \left(I - A_Q^* A_Q\right) \leq 1 + \alpha_j \left(1 - \Delta_3\right) , \tag{130}$$

with $\Delta_3 := \Delta_{3r,3s,2\tau}$ being the RIC of $\mathcal{A}$. By α_j^L and α_j^S being the normalized stepsizes as introduced in [27, 28]

$$\alpha_j^L = \frac{\left\|\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right)\right\|_F^2}{\left\|\mathcal{A}\left(\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right)\right)\right\|_2^2} \qquad \text{and} \qquad \alpha_j^S = \frac{\left\|\mathrm{Proj}_{(U^{j+1},\Omega^j)}\left(R^j\right)\right\|_F^2}{\left\|\mathcal{A}\left(\mathrm{Proj}_{(U^{j+1},\Omega^j)}\left(R^j\right)\right)\right\|_2^2} \tag{131}$$

where the projection $\mathrm{Proj}_{(U^j,\Omega^j)}\left(R^j\right), \mathrm{Proj}_{(U^{j+1},\Omega^j)}\left(R^{j+\frac{1}{2}}\right)$ ensures that the residual R^j and $R^{j+\frac{1}{2}}$ is projected into the set $\mathrm{LS}_{m,n}(r, s, \tau)$. Then, it follows from the RIC for $\mathcal{A}$ that the stepsizes α_j^L, α_j^S can be bounded as

$$\frac{1}{1 + \Delta_1} \leq \alpha_j^{L/S} \leq \frac{1}{1 - \Delta_1}, \tag{132}$$

where $\Delta_1 := \Delta_{r,s,\tau}$. Putting (130) and (132) together

$$1 - \frac{1+\Delta_3}{1-\Delta_1} \leq \lambda\left(I - \alpha_j^{L/S}\, A_Q^* A_Q\right) \leq 1 - \frac{1-\Delta_3}{1+\Delta_1}. \tag{133}$$

Since $\Delta_3 \geq \Delta_1$ we have that the magnitude of the lower bound in (133) is greater than the upper bound. Therefore

$$\frac{1+\Delta_3}{1-\Delta_1} - 1 \geq \left\| I - \alpha_j^{L/S}\, A_Q^* A_Q \right\|_2. \tag{134}$$

Finally, the constant on the right hand side of (129) can be upper bounded

$$\eta := 2\frac{\left\|I - \alpha_j^S\, A_Q^* A_Q\right\| + \alpha_j^L\, \Delta_3}{1 - 2\,\alpha_j^S\, \Delta_3} \tag{135}$$

$$\leq 2\frac{\left(\frac{1+\Delta_3}{1-\Delta_1} - 1\right) + \frac{\Delta_3}{1-\Delta_1}}{1 - 2\frac{\Delta_3}{1-\Delta_1}} = 2\frac{\Delta_1 + 2\Delta_3}{1 - \Delta_1 - 2\Delta_3} \tag{136}$$

$$\leq \frac{6\Delta_3}{1-3\Delta_3}, \tag{137}$$

where the inequality in the second line comes from upper bounds in (134) and in (132), and the third line follows from $\Delta_3 \geq \Delta_1$. For $\Delta_3 := \Delta_{3r,3s,2\tau} < 1/9$ the inequality in (129) implies contraction of the error

$$\left\|L_0 - L^{j+1}\right\|_F + \left\|S_0 - S^{j+1}\right\|_F \leq \eta\left(\left\|L^j - L_0\right\|_F + \left\|S^j - S_0\right\|_F\right), \tag{138}$$

because $\eta < 1$ which guarantees linear convergence of iterates L^j and S^j to L_0 and S_0 respectively. □

4. Numerical experiments

This section demonstrates the computational efficacy of recoverying a low-rank plus sparse matrix $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ from its undersampled values $\mathcal{A}(X_0)$. Section 4.1 considers synthetic examples where matrices in $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ are created, and recovery from their undersampled values attempted for the following algorithms: NIHT (Alg. 1), NAHT (Alg. 2), SpaRCS [31], and the convex relaxation (8). Figure 1 presents empirically observed phase transitions, which indicate the values of model complexity r, s, and measurements p for which recovery is possible. Figure 2 and 3 gives examples of convergence rates for NIHT, NAHT, and SpaRCS, including contrasting different methods to implement the projection NIHT, step 5 of Alg. 1. Section 4.2 presents applications to dynamic-foreground/static-background and computational multispectal imaging. An additional phase transition simulation for the convex relaxation is given in Appendix Appendix A. Software to reproduce the experiments in this section is publicly available[5].

[5] https://github.com/svary/lrps-recovery

4.1. Empirical average case performance on synthetic data

Synthetic matrices $X_0 = L_0 + S_0 \in \mathrm{LS}_{m,n}(r, s, \tau)$ are generated using the experimental setup proposed in the Robust PCA literature [35, 36, 37]. The low-rank component is formed as $L_0 = UV^T$, where $U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}$ are two random matrices having their entries drawn i.i.d. from the standard Gaussian distribution. The support set of the sparse component S_0 is generated by sampling a uniformly random subset of $[m] \times [n]$ indices of size s and each non-zero entry $(S_0)_{i,j}$ is drawn from the uniform distribution over $[-\mathbb{E}\left(|(L_0)_{i,j}|\right), \mathbb{E}\left(|(L_0)_{i,j}|\right)]$. Each synthetic matrix is measured using linear operators $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$. The random Gaussian measurement operators are constructed by p matrices $A^{(\ell)} \in \mathbb{R}^{m\times n}$ whose entries are sampled from Gaussian distribution $A^{(\ell)}_{i,j} \sim \mathcal{N}(0, 1/p)$ where p is the number of measurements. The Fast Johnson-Lindenstrauss Transform is implemented as

$$\mathcal{A}_{\mathrm{FJLT}}\left(X\right) = RHD\,\mathrm{vec}\left(X\right), \tag{139}$$

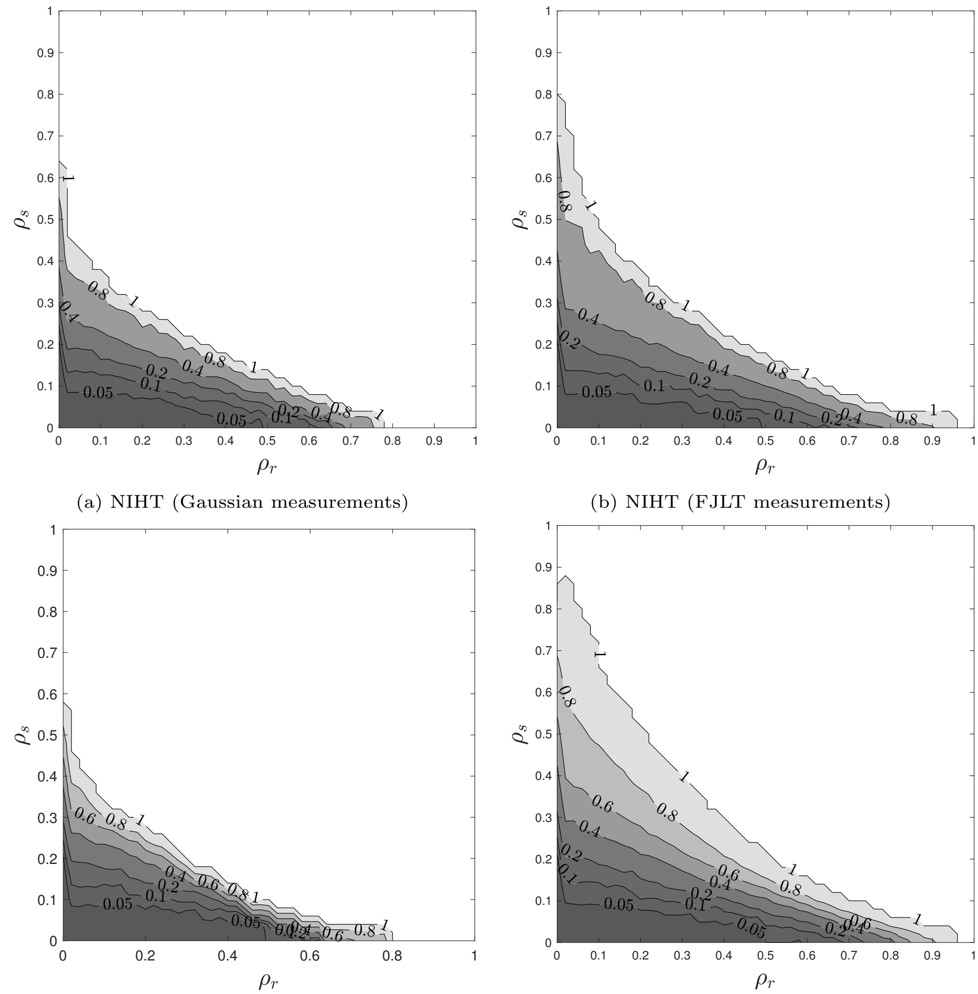
where $R \in \mathbb{R}^{p\times mn}$ is a restriction matrix constructed from a $mn \times mn$ identity matrix with p rows randomly selected, $H \in \mathbb{R}^{mn\times mn}$ is discrete cosine transform matrix, $D \in \mathbb{R}^{mn\times mn}$ is a diagonal matrix whose entries are sampled independently randomly from $\{-1, 1\}$, and $\mathrm{vec}\left(X\right) \in \mathbb{R}^{mn}$ is the vectorized matrix $X \in \mathbb{R}^{m\times n}$.

Theorems 1, 3, 4, and 5 indicate that recovery of X_0 from $\mathcal{A}(X_0)$ depends on the problem dimensions through the ratios of the number of measurements p with the ambient dimension mn, and the minimum number of measurements, $r(m+n-r)+s$, through an undersampling and two oversampling ratios

$$\delta = \frac{p}{mn} \quad \text{and} \quad \rho_r = \frac{r(m+n-r)}{p}, \quad \rho_s = \frac{s}{p}. \tag{140}$$

The matrix dimensions m and n are held fixed, while p, r and s are chosen according to varying parameters δ, ρ_r and ρ_s. For each pair of $\rho_r, \rho_s \in \{0, 0.02, 0.04, \ldots, 1\}$ where $\rho_r + \rho_s \le 1$, with the sampling ratio restricted to values $\delta \in \{0.02, 0.04, \ldots, 1\}$, 20 simulated recovery tests are conducted and we compute the critical subsampling ratio δ^* above which more than half of the experiments succeeded. For the linear transform $\mathcal{A}$ drawn from the (dense) Gaussian distribution, the highest per iteration cost in NIHT and NAHT comes from applying $\mathcal{A}$ to the residual matrix, which requires pmn scalar multiplications which scales proportionally to $(mn)^2$. For this reason, our tests are restricted to the matrix size of $m = n = 100$ in the case of NIHT and NAHT, and to a smaller size $m = n = 30$ for testing the recovery by solving the convex relaxation (8) with semidefinite programming [38] that has $\mathcal{O}\left((mn)^2\right)$ variables which is more computationally demanding[6] compared to the hard thresholding

[6] As an example, a low-rank plus sparse matrix with $m = n = 100$ with $\rho_r = \rho_s = 0.1$ undersampled and measured with Gaussian matrix with $\delta = 0.5$ takes 2.5 seconds and 2.3 seconds to recover using NIHT and NAHT respectively, while the recovery using the convex relaxation takes over 7 hours.



(a) NIHT (Gaussian measurements) (b) NIHT (FJLT measurements)

(c) NAHT (Gaussian measurements) (d) NAHT (FJLT measurements)

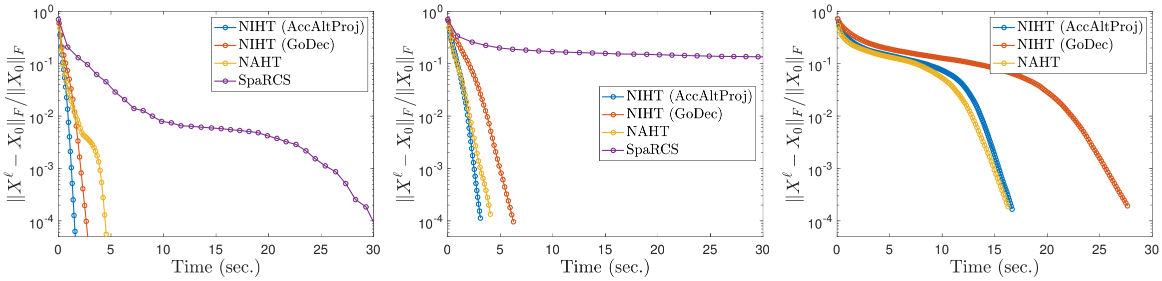
Figure 1: Phase transition level curves denoting the value of δ^* for which values of ρ_r and ρ_s *below which* are recovered for at least half of the experiments for δ, ρ_r, and ρ_s as given by (140). NIHT is observed to recover matrices of higher ranks and sparsities from FJLT than from Gaussian measurements, while the phase transitions for NIHT and NAHT are comparable. The RPCA projection in NIHT, step 5 in Alg. 1, is performed by AccAltProj [37].

gradient descent methods. Algorithms are terminated at iteration ℓ when either: the relative residual error is smaller than 10^{-6}, that is when $\| \mathcal{A}(X^\ell) - b\|_2/\|b\|_2 \leq 10^{-6}\|b\|_2$, or the relative decrease in the objective is small

$$\left(\frac{\| \mathcal{A}(X^\ell) - b\|_2}{\| \mathcal{A}(X^{\ell-15}) - b\|_2} \right)^{1/15} > 0.999, \tag{141}$$

or the maximum of 300 iterations is reached. An algorithm is considered to have successfully recovered $X_0 \in \mathrm{LS}_{m,n}(r,s,\tau)$ if it returns a matrix $X^\ell \in \mathrm{LS}_{m,n}(r,s,\tau)$ that is within 10^{-2} of X_0 in the relative Frobenius error, $\|X^\ell - X_0\|_F \leq 10^{-2}\|X_0\|_F$.

Figure 1 depicts the phase transitions of δ above which NIHT and NAHT successfully recovers

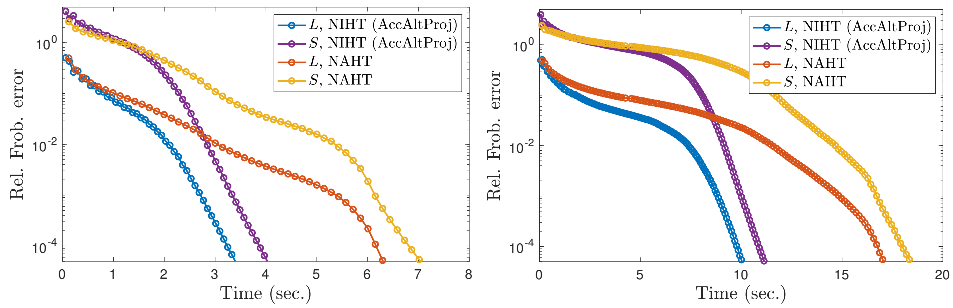


(a) $\rho_r = \rho_s = 0.05$ (b) $\rho_r = \rho_s = 0.1$ (c) $\rho_r = \rho_s = 0.2$

Figure 2: Relative error in the approximate X^ℓ as a function of time for synthetic problems with $m = n = 100$ and $p = (1/2)100^2$, $\delta = 1/2$, for Gaussian linear measurements $\mathcal{A}$. In (b), SpaRCS converged in 171 sec. (45 iterations), and in (c), SpaRCS did not converge.

X_0 in more than half of the experiments. For example, the level curve 0.4 in Fig. 1 denotes the values of ρ_r and ρ_s *below which* recovery is possible for at least half of the experiments for $p = 0.4mn$ and ρ_r, ρ_s as given by (140). Note that the bottom left portion of Fig. 1 corresponds to smaller values of model complexity (r, s) and are correspondingly easier to recover than larger values of (r, s). Both algorithms are observed to recover matrices with prevalent rank structure, $\rho_r \leq 0.6$, even from very few measurements as opposed to matrices with prevalent sparse structure requiring in general more measurements for a successful recovery. Phase transitions corresponding to the sparse-only ($\rho_r = 0$) and to the rank-only ($\rho_s = 0$) cases are roughly in agreement with phase transitions that have been observed for non-convex algorithms in compressed sensing [39] and matrix completion literature [28, 40]. We observe that NAHT achieves almost identical performance to NIHT in terms of possible recovery despite not requiring the computationally expensive Robust PCA projection in every iteration. For both algorithms we see that the successful recovery is possible for matrices with higher ranks and sparsities in the case of FJLT measurements compared to Gaussian measurements.

Equivalent experiments are conducted for the convex relaxation (8), but with smaller matrix size 30×30 and limited to 10 simulations for each set of parameters due to the added computational demands. The convex optimization is formulated using CVX modeling framework [41] and solved in Matlab by the semidefinite programming optimization package SDPT3 [38]. We observe that recovery by solving the convex relaxation is successful for somewhat lower ranks and sparsities and requiring larger sampling ratio δ compared to the non-convex algorithms. The observed phase transitions of the convex relaxation alongside phase transitions for $m = n = 30$ experiments with NIHT and NAHT are depicted in Figure A.6 in Appendix A. Comparing the phase transitions of the non-convex algorithms in Fig. 1 and Fig. A.6 show that with the increased problem size, the phase transition are independent of the dimension with only small differences due to finite dimensional effects of the smaller problem size in the case of $m = n = 30$.



(a) $\delta = 0.5, \rho_r = 0.05, \rho_s = 0.15$ (b) $\delta = 0.5, \rho_r = 0.05, \rho_s = 0.25$

Figure 3: Error between between the approximate recovered low-rank and sparse components L^ℓ and S^ℓ and the true low-rank and sparse components L_0 and S_0. Error is plotted as a function of recovery time for synthetic problems with $m = n = 100$ and $p = (1/2)100^2$, $\delta = 1/2$, for Gaussian linear measurements $\mathcal{A}$.

Figure 2 presents convergence timings of Matlab implementations of the three non-convex algorithms used for recovery of matrices with $m = n = 100$ from $p = (1/2)10^2$ ($\delta = 1/2$) measurements and three values of $\rho_r = \rho_s = \{0.05, 0.1, 0.2\}$. The convergence results are presented for two variants of NIHT with different Robust PCA algorithms Accelerated Alternating Projection (AccAltProj) [37] and Go Decomposition (GoDec) [42] in the projection step 5 of Alg. 1. Both NIHT and NAHT converge at a much faster rate than the existing non-convex algorithm for low-rank plus sparse matrix recovery SpaRCS [31]. All the algorithms take longer to recover a matrix for increased rank r and/or sparsity s.

The computational efficacy of NIHT compared to NAHT depends on the cost of computing the Robust PCA calculation in comparison to the cost of applying $\mathcal{A}$. NAHT computes two step sizes in each iteration which results into computing $\mathcal{A}$ twice per iteration in comparison to just one such computation per iteration in the case of NIHT. On the other hand, NIHT involves solving Robust PCA in every iteration for the projection step whereas NAHT performs computationally cheaper singular value decomposition (SVD) and sparse hard thresholding projection.

Figure 3 illustrates the convergence of the individual low-rank and sparse components $\|L^\ell - L_0\|_F$ and $\|S^\ell - S_0\|_F$ as a function of time. The algorithms are observed to approximate the the low-rank factor more accurately than the sparse component and that the computational time increases for larger values of sparsity fraction ρ_s. Moreover, for both NIHT and NAHT the relative error of both components decreases together.

4.2. Applications

4.2.1. Dynamic-foreground/static-background video separation

Background/foreground separation is the task of distinguishing moving objects from the static-background in a time series, e.g. a video recording. A widely used approach is to arrange frames of the

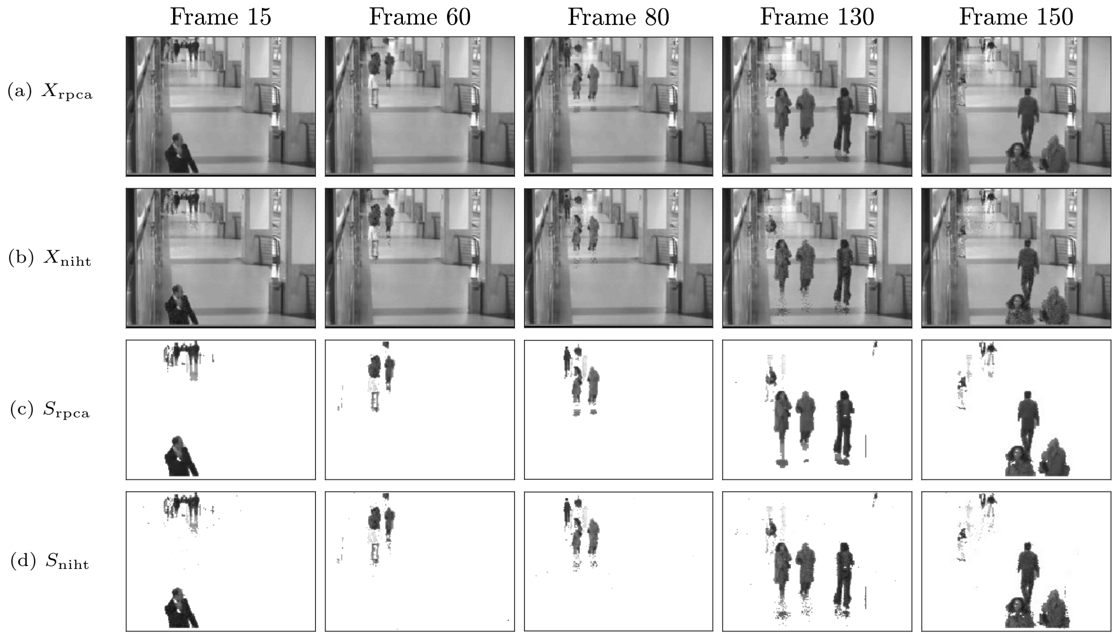


Figure 4: NIHT recovery results of a $256 \times 256 \times 150$ video sequence compared to the approximation of the complete video sequence by Robust PCA (AccAltProj [37]). The video sequence is reshaped into a $26\,600 \times 150$ matrix and either recovered from FJLT measurements with $\delta = 0.33$ using rank $r = 1$ and sparsity $s = 197\,505$ or approximated from the full video sequence by computing RPCA by AccAltProj with the same rank and sparsity parameters. Recovery by NIHT from subsampled information achieves PSNR of 34.5 dB whereas the Robust PCA approximation from the full video sequence achieves PSNR of 35.5 dB.

video sequence into an $m \times n$ matrix, where m is the number of pixels and n is the number of frames of the recording and apply Robust PCA to decompose the matrix into the sum of a low-rank and a sparse component which model the static background and dynamic foreground respectively [3]. Herein we consider the same problem but with the additional challenge of recovering the video sequence from subsampled information [31] analogous to compressed sensing.

We apply NIHT, Alg. 1, to the well studied shopping mall surveillance [43] which is $256 \times 256 \times 150$ video sequence. The video sequence is rearranged into a matrix of size $26\,600 \times 150$ and measured using subsampled FJLT (139) with one third as many meausrements as the ambient dimension, $\delta = 0.33$. The static-background is modeled with a rank-r matrix with $r = 1$ and the dynamic-foreground by an s-sparse matrix with $s = 197\,505$ ($\rho_r = 0.02$, $\rho_s = 0.15$). Figure 4 displays the reconstructed image X_{niht} and its sparse component S_{niht} alongside the results obtained from applying Robust PCA (AccAltProj [37]) which makes use of the fully sampled video sequence rather than the one-third measurements available to NIHT. NIHT accurately estimates the video sequence achieving PSNR of 34.5 dB while also separating the low-rank background from the sparse foreground. The results are of a similar visual quality to the case of Robust PCA that achieves PSNR of 35.5 dB which requires access to the full video sequence.

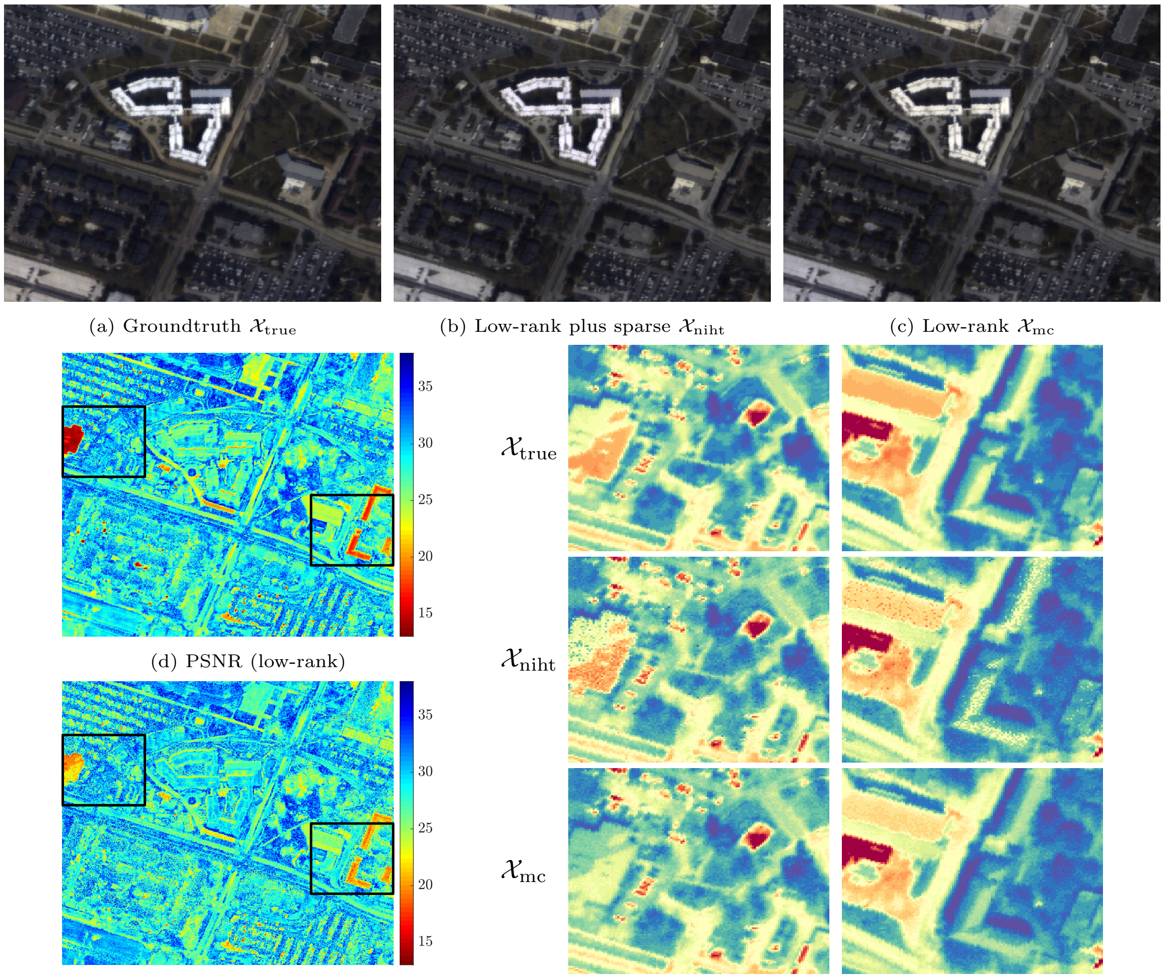
4.2.2. Computational multispectral imaging

A multispectral image captures a wide range of light spectra generating a vector of spectral responses at each image pixel thus acquiring information in the form of a third order tensor. Low-rank model has a vital role in multispectral imaging in the form of a linear spectral mixing models that assume the spectral responses of the imaged scene are well approximated as a linear combination of spectral responses of only few core materials referred to as *endmembers* [44]. As such, the low-rank structure can be exploited by computational imaging systems which acquire the image in a compressed from and use computational methods to recover a high-resolutional image [45, 46, 47]. However, when different materials are in close proximity the resulting spectrum can be highly nonlinear combination of the *endmembers* resulting in anomalies of the model [48]. Herein we propose the low-rank plus sparse matrix recovery as a way to model the spectral anomalies in the low-rank structure.

We employ NIHT on a $512 \times 512 \times 48$ airborne hyperspectral image from the GRSS 2018 Data Fusion contest [49] that is rearranged into a matrix of size $262\,144 \times 48$ and subsampled using FJLT with $\delta = 0.33$. Figure 5 demonstrates recovery by NIHT using rank $r = 3$ and sparsity $s = 150\,995$ ($\rho_r = 0.25$, $\rho_s = 0.05$) in comparison with the the low-rank model with rank $r = 3$ and $s = 0$ ($\rho_r = 0.25$, $\rho_s = 0$). Both methods recover the image well but the low-rank plus sparse recovery achieves slightly higher PSNR of 39.1 dB compared to the low-rank recovery that has PSNR of 38.9 dB and slightly better fine details. Figure 5d and Figure 5e depict the localization of the error in terms of PSNR and shows that adding the sparse component improves PSNR of a few localized parts. Although the overall gain in the PSNR is small compared to the low-rank model, the differences in the localized regions of the image can be potentially impactful when further analyzed in practical applications such as semantic segmentation [50].

5. Conclusion

The main theorems, Theorems 1, 2, 3, 4, and 5, are the natural extension of analogous results in the compressed sensing and matrix completion literature to the space of low-rank plus sparse matrices, Def. 1.1, see [4, 5] and references therein. They establish the foundational theory and provide examples of algorithms for recovery of matrices that can be expressed as a sum of a low-rank and a sparse matrix from under sampled measurements. While these results could be anticipated, with [31] being an early non-convex algorithm for this setting, these advancements had not yet been proven. We prove that the restricted isometry constants of random linear operators obeying concentration of measure inequalities, such as Gaussian measurements or the Fast Johnson-Lindenstrauss Transform, can be upper bounded



(a) Groundtruth $\mathcal{X}_{\text{true}}$ (b) Low-rank plus sparse $\mathcal{X}_{\text{niht}}$ (c) Low-rank $\mathcal{X}_{\text{mc}}$

(d) PSNR (low-rank)

(e) PSNR (low-rank plus sparse) (f) Detail 1 (694 nm) (g) Detail 2 (694 nm)

Figure 5: Recovery by NIHT from FJLT measurements with $\delta = 0.33$ using low-rank model ($\rho_r = 0.25$, $\rho_s = 0$) compared to the low-rank plus sparse model ($\rho_r = 0.25$, $\rho_s = 0.05$). Figure 5a - 5b show the color renderings of the original multispectral image and the two recovered images. Figure 5d and Figure 5e show the spatial PSNR of the recovery from the low-rank only model (overall PSNR of 38.9 dB) and the low-rank plus sparse model (overall PSNR of 39.1 dB) respectively. Figure 5f and Figure 5g show two details of size 128×128 in the 694 nm band.

when the number of measurements are of the order depending on the degrees of freedom of the low-rank plus sparse matrix. Making use of these RICs, we show that low-rank plus sparse matrices can be provably recovered by computationally efficient methods, e.g. by solving semidefinite programming or by two gradient descent algorithms, when the restricted isometry constants of the measurement operator are sufficiently bounded. Numerical experiments on synthetic data empirically demonstrate phase transitions in the parameter space for which the recovery is possible. Experiments for dynamic-foreground/static-background video separation show that the segmentation of moving objects can be obtains with similar error from only one third as many measurement as compared to the entire video sequence. The contributions here open up the possibility of other algorithms in compressed sensing and low-rank matrix completion/sensing to be extended to the case of low-rank plus sparse matrix recovery, e.g. more efficient algorithms such as those employing momentum [51, 52] or minimising over increasingly larger subspaces [40]. These results also illustrate how RICs can be developed for more complex additive data models and one can expect that similar results would be possible for new data models.

Acknowledgement

We would like to thank Robert A. Lamb and David Humphreys for useful discussions around the applications of low-rank plus sparse model to multispectral imaging.

References

[1] E. J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM 58 (3) (2011) 1–37. doi:10.1145/1970392.1970395.

[2] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, A. S. Willsky, Rank-sparsity incoherence for matrix decomposition, SIAM Journal on Optimization 21 (2) (2011) 572–596. doi:10.1137/090761793.

[3] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, E.-H. Zahzah, Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset, Computer Science Review 23 (2017) 1–71. doi:10.1016/j.cosrev.2016.11.001.

[4] Y. C. Eldar, G. Kutyniok, Compressed sensing: Theory and applications, Cambridge University Press, 2012. doi:10.1017/CBO9780511794308.

[5] S. Foucart, H. Rauhut, A Mathematical Introduction to Compressive Sensing, Applied and Numerical Harmonic Analysis, Springer New York, New York, NY, 2013. doi:10.1007/978-0-8176-4948-7.

[6] D. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (4) (2006) 1289–1306. doi:10.1109/TIT.2006.871582.

[7] E. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory 52 (2) (2006) 489–509. arXiv:0409186, doi:10.1109/TIT.2005.862083.

[8] E. Candes, T. Tao, Decoding by Linear Programming, IEEE Transactions on Information Theory 51 (12) (2005) 4203–4215. doi:10.1109/TIT.2005.858979.

[9] E. J. Candès, B. Recht, Exact matrix completion via convex optimization, Foundations of Computational Mathematics 9 (6) (2009) 717–772. doi:10.1007/s10208-009-9045-5.

[10] E. J. Candes, T. Tao, The power of convex relaxation: near-optimal matrix completion, IEEE Transactions on Information Theory 56 (5) (2010) 2053–2080. doi:10.1109/TIT.2010.2044061.

[11] B. Recht, M. Fazel, P. A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Review 52 (3) (2010) 471–501. doi:10.1137/070697835.

[12] E. J. Candès, C. Fernandez-Granda, Towards a mathematical theory of super-resolution, Communications on Pure and Applied Mathematics 67 (6) (2014) 906–956. doi:10.1002/cpa.21455.

[13] V. Duval, G. Peyré, Exact support recovery for sparse spikes Deconvolution, Foundations of Computational Mathematics 15 (5) (2015) 1315–1355. doi:10.1007/s10208-014-9228-6.

[14] A. Eftekhari, J. Tanner, A. Thompson, B. Toader, H. Tyagi, Sparse non-negative super-resolution — simplified and stabilised, Applied and Computational Harmonic Analysis 1 (2019) 1–65. doi:10.1016/j.acha.2019.08.004.

[15] Y. Chi, M. Ferreira Da Costa, Harnessing sparsity over the continuum: atomic norm minimization for superresolution, IEEE Signal Processing Magazine 37 (2) (2020) 39–57. doi:10.1109/MSP.2019.2962209.

[16] S. Gu, L. Zhang, W. Zuo, X. Feng, Weighted nuclear norm minimization with application to image denoising, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, no. 2, IEEE, 2014, pp. 2862–2869. doi:10.1109/CVPR.2014.366.

[17] A. Gogna, A. Shukla, H. K. Agarwal, A. Majumdar, Split Bregman algorithms for sparse / joint-sparse and low-rank signal recovery: application in compressive hyperspectral imaging, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 1302–1306. doi:10.1109/ICIP.2014.7025260.

[18] Y. Chen, Y. Guo, Y. Wang, D. Wang, C. Peng, G. He, Denoising of hyperspectral images using nonconvex low rank matrix approximation, IEEE Transactions on Geoscience and Remote Sensing 55 (9) (2017) 5366–5380. doi:10.1109/TGRS.2017.2706326.

[19] W. Wei, L. Zhang, Y. Zhang, C. Wang, C. Tian, Hyperspectral image denoising from an incomplete observation, in: 2015 International Conference on Orange Technologies (ICOT), IEEE, 2015, pp. 177–180. doi:10.1109/ICOT.2015.7498517.

[20] X. Luan, B. Fang, L. Liu, W. Yang, J. Qian, Extracting sparse error of robust PCA for face recognition in the presence of varying illumination and occlusion, Pattern Recognition 47 (2) (2014) 495–508. doi:10.1016/j.patcog.2013.06.031.

[21] J. Wright, A. Yang, A. Ganesh, S. Sastry, Yi Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227. doi:10.1109/TPAMI.2008.79.

[22] F. Xu, J. Han, Y. Wang, M. Chen, Y. Chen, G. He, Y. Hu, Dynamic magnetic resonance imaging via nonconvex low-rank matrix approximation, IEEE Access 5 (2017) 1958–1966. doi:10.1109/ACCESS.2017.2657645.

[23] H. Gao, J.-F. Cai, Z. Shen, H. Zhao, Robust principal component analysis-based four-dimensional computed tomography, Physics in Medicine and Biology 56 (11) (2011) 3181–3198. doi:10.1088/0031-9155/56/11/002.

[24] O. Oreifej, X. Li, M. Shah, Simultaneous video stabilization and moving object detection in turbulence, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2) (2013) 450–462. doi:10.1109/TPAMI.2012.97.

[25] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation 28 (3) (2008) 253–263. doi:10.1007/s00365-007-9003-x.

[26] J. Tanner, A. Thompson, S. Vary, Matrix Rigidity and the Ill-Posedness of Robust PCA and Matrix Completion, SIAM Journal on Mathematics of Data Science 1 (3) (2019) 537–554. doi:10.1137/18M1227846.

[27] T. Blumensath, M. Davies, Normalized iterative hard thresholding: guaranteed stability and performance, IEEE Journal of Selected Topics in Signal Processing 4 (2) (2010) 298–309. doi:10.1109/JSTSP.2010.2042411.

[28] J. Tanner, K. Wei, Normalized iterative hard thresholding for matrix completion, SIAM Journal on Scientific Computing 35 (5) (2013) S104–S125. doi:10.1137/120876459.

[29] N. Ailon, B. Chazelle, The fast Johnson–Lindenstrauss transform and approximate nearest neighbors, SIAM Journal on Computing 39 (1) (2009) 302–322. doi:10.1137/060673096.

[30] F. Krahmer, R. Ward, New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property, SIAM Journal on Mathematical Analysis 43 (3) (2011) 1269–1281. doi:10.1137/100810447.

[31] A. E. Waters, A. C. Sankaranarayanan, R. G. Baraniuk, SpaRCS: recovering low-rank and sparse matrices from compressive measurements, in: Advances in Neural Information Processing Systems 24 (NIPS 2011), no. 2, 2011, pp. 1089—-1097.

[32] R. G. Baraniuk, V. Cevher, M. F. Duarte, C. Hegde, Model-based compressive sensing, IEEE Transactions on Information Theory 56 (4) (2010) 1982–2001. doi:10.1109/TIT.2010.2040894.

[33] S. Szarek, Metric entropy of homogeneous spaces, Banach Center Publications 43 (1) (1998) 395–410. doi:10.4064/-43-1-395-410.

[34] E. J. Candès, J. K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics 59 (8) (2006) 1207–1223. doi:10.1002/cpa.20124.

[35] P. Netrapalli, U. N. Niranjan, S. Sanghavi, A. Anandkumar, P. Jain, Non-convex robust PCA, in: Advances in Neural Information Processing Systems 27 (NIPS 2014), 2014.

[36] X. Yi, D. Park, Y. Chen, C. Caramanis, Fast algorithms for robust PCA via gradient descent, in: Advances in Neural Information Processing Systems 29 (NIPS 2016), 2016.

[37] H. Cai, J.-F. Cai, K. Wei, Accelerated alternating projections for robust principal component analysis, Journal of Machine Learning Research 20 (1) (2019) 685—-717.

[38] K. C. Toh, M. J. Todd, R. H. Tütüncü, SDPT3 — A Matlab software package for semidefinite programming, Version 1.3, Optimization Methods and Software 11 (1-4) (1999) 545–581. doi:10.1080/10556789908805762.

[39] J. Blanchard, J. Tanner, Performance comparisons of greedy algorithms in compressed sensing, Numerical Linear Algebra with Applications 22 (2) (2015) 254–282. doi:10.1002/nla.1948.

[40] J. D. Blanchard, J. Tanner, K. Wei, CGIHT: Conjugate gradient iterative hard thresholding for compressed sensing and matrix completion, Information and Inference (2015) iav01doi:10.1093/imaiai/iav011.

[41] G. Stephen, M. Boyd, CVX: Matlab Software for Disciplined Convex Programming, version 2.1 (2014).

[42] T. Zhou, D. Tao, GoDec: Randomized low-rank & sparse matrix decomposition in noisy case, Proceedings of the 28th International Conference on Machine Learning 35 (1) (2011) 33–40.

[43] L. Li, W. Huang, I.-H. Gu, Q. Tian, Statistical modeling of complex backgrounds for foreground object detection, IEEE Transactions on Image Processing 13 (11) (2004) 1459–1472. doi:10.1109/TIP.2004.836169.

[44] M. Dimitris, D. Marden, G. Shaw A., Hyperspectral image processing for automatic target detection applications, Lincoln Laboratory Journal 14 (1) (2003) 79 —- 116. doi:10.1039/C4RA04655B.

[45] X. Cao, T. Yue, X. Lin, S. Lin, X. Yuan, Q. Dai, L. Carin, D. J. Brady, Computational snapshot multispectral cameras: toward dynamic capture of the spectral world, IEEE Signal Processing Magazine 33 (5) (2016) 95–108. doi:10.1109/MSP.2016.2582378.

[46] K. Degraux, V. Cambareri, L. Jacques, B. Geelen, C. Blanch, G. Lafruit, Generalized inpainting method for hyperspectral image acquisition, in: 2015 IEEE International Conference on Image Processing (ICIP), Vol. 2015-Decem, IEEE, 2015, pp. 315–319. doi:10.1109/ICIP.2015.7350811.

[47] G. A. Antonucci, S. Vary, D. Humphreys, R. A. Lamb, J. Piper, J. Tanner, Multispectral snapshot demosaicing via non-convex matrix completion, in: 2019 IEEE Data Science Workshop (DSW), IEEE, 2019, pp. 227–231. doi:10.1109/DSW.2019.8755561.

[48] D. Stein, S. Beaven, L. Hoff, E. Winter, A. Schaum, A. Stocker, Anomaly detection from hyperspectral imagery, IEEE Signal Processing Magazine 19 (1) (2002) 58–69. doi:10.1109/79.974730.

[49] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hansch, B. Le Saux, Advanced multi-sensor optical remote sensing for urban land use and land cover classification: outcome of the 2018 IEEE GRSS data fusion contest, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12 (6) (2019) 1709–1724. doi:10.1109/JSTARS.2019.2911113.

[50] R. Kemker, C. Salvaggio, C. Kanan, Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning, ISPRS Journal of Photogrammetry and Remote Sensing 145 (June 2017) (2018) 60–77. doi:10.1016/j.isprsjprs.2018.04.014.

[51] A. Kyrillidis, V. Cevher, Matrix recipes for hard thresholding methods, Journal of Mathematical Imaging and Vision 48 (2) (2014) 235–265. doi:10.1007/s10851-013-0434-7.

[52] K. Wei, Fast iterative hard thresholding for compressed sensing, IEEE Signal Processing Letters 22 (5) (2015) 593–597. doi:10.1109/LSP.2014.2364851.

[53] G. G. Lorentz, M. V. Golitschek, Y. Makovoz, Constructive approximation: Advanced problems, Springer-Verlag Berlin Heidelberg, 1996.

Appendix A. Phase transitions for synthetic problem of size $m = n = 30$

Figure A.6 depicts the phase transitions of δ above which NIHT, NAHT and solving the convex relaxation problem in (8) successfully recovers X_0 in more than half of the experiments. Comparing Fig. A.6 to Fig. 1 we see that the phase transitions roughly occur for the same parameters ρ_r, ρ_s with only small differences due to the finite dimensional effects of the smaller problem size being more pronounced when $m = n = 30$. We also observe that non-convex algorithms perform better than the convex relaxation in that they are able to recover higher ranks and sparsities from fewer samples in addition to also taking less time to converge.

Appendix B. Supporting lemmata

Proof of Lemma 2.1 (RIC for a fixed LS subspace), stated on page 9.
The proof uses similar to arguments as [25, Lemma 5.1] and [11, Lemma 4.3] with the exception that here we consider two subsets, one for the low rank and another for the sparse component.

Proof. By the linearity of $\mathcal{A}(\cdot)$ assume without loss of generality that $\|X\|_F = 1$. By the triangle inequality and $\|X\|_F = 1$, the upper bound of the Frobenius norm of the low-rank component $\|L\|_F \leq \tau$ gives also an upper bound on the Frobenius norm of the sparse component $\|S\|_F \leq 1 + \tau$. There exist two finite $(\widehat{\Delta}/8)$-coverings of the two matrix sets

$$\left\{L \in \mathbb{R}^{m\times n} \,:\, \mathcal{C}(L) \subseteq V,\, \mathcal{C}(L^T) \subseteq W,\, \|L\|_F \leq \tau\right\} \tag{B.1}$$

$$\left\{S \in \mathbb{R}^{m\times n} \,:\, S \subseteq T,\, \|S\|_F \leq 1+\tau\right\}, \tag{B.2}$$

that we denote Λ^L, Λ^S and by [53, Chap. 16] their covering numbers are upper bounded as

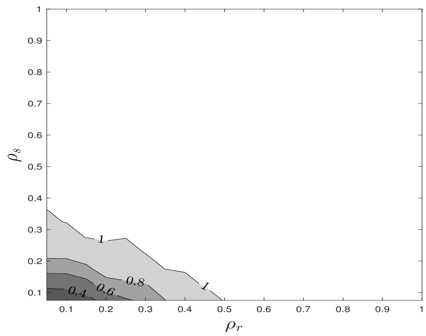
$$\left|\Lambda^L\right| \leq \left(\frac{24}{\widehat{\Delta}}\tau\right)^{\dim V \cdot \dim W} \qquad\qquad \left|\Lambda^S\right| \leq \left(\frac{24}{\widehat{\Delta}}(1+\tau)\right)^{\dim T}. \tag{B.3}$$

Let $\Lambda := \left\{Q^L + Q^S : Q^L \in \Lambda^L,\, Q^S \in \Lambda^S\right\}$ be the set of sums of all possible pairs of the two coverings. The set Λ is a $(\widehat{\Delta}/4)$-covering of the set $\Sigma_{m,n}(V, W, T, \tau)$ since for all $X \in \Sigma_{m,n}(V, W, T, \tau)$ there exists a pair $Q \in \Lambda$ such that

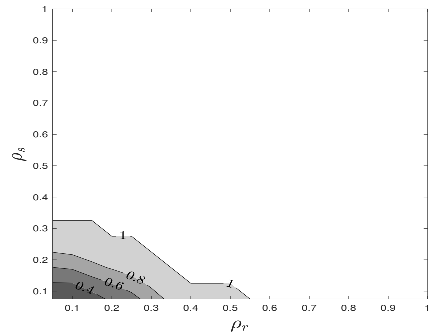
$$\|X - Q\|_F = \left\|L + S - \left(Q^L + Q^S\right)\right\|_F \tag{B.4}$$

$$\leq \left\|L - Q^L\right\|_F + \left\|S - Q^S\right\|_F \leq \frac{\widehat{\Delta}}{8} + \frac{\widehat{\Delta}}{8}, \tag{B.5}$$

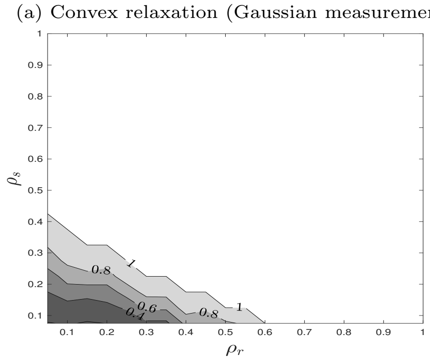
where in the first line we used the fact that X can be expressed as $L + S$, and in the second line we applied the triangular inequality combined with the Q^L, Q^S being $(\widehat{\Delta}/8)$-coverings of the matrix sets for the low-rank component and the sparse component respectively.



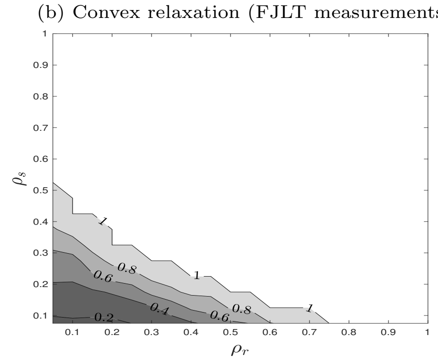
(a) Convex relaxation (Gaussian measurements)



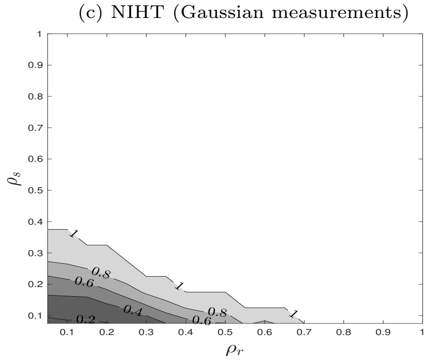
(b) Convex relaxation (FJLT measurements)



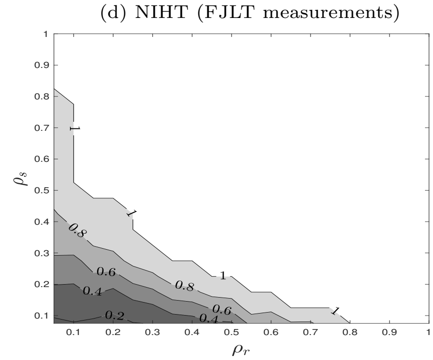
(c) NIHT (Gaussian measurements)



(d) NIHT (FJLT measurements)



(e) NAHT (Gaussian measurements)



(f) NAHT (FJLT measurements)

Figure A.6: Phase transition level curves denoting the value of δ^* for which values of ρ_r and ρ_s *below which* are recovered for at least 5 out of 10 experiments for δ, ρ_r, and ρ_s as given by (140). The convex optimization problem is solved by SDPT3 [38]. NIHT and NAHT are observed to recover matrices of higher ranks and sparsities compared to solving the convex relaxation.

Applying the probability union bound on concentration of measure of $\mathcal{A}$ as in (4) with $\varepsilon = \widehat{\Delta}/2$ gives that

$$(\forall Q \in \Lambda) : \left(1 - \frac{\widehat{\Delta}}{2}\right) \|Q\|_F \leq \|\,\mathcal{A}(Q)\| \leq \left(1 + \frac{\widehat{\Delta}}{2}\right) \|Q\|_F, \tag{B.6}$$

holds with the probability at least

$$1 - 2\left(\frac{24}{\widehat{\Delta}}\tau\right)^{\dim V \cdot \dim W} \left(\frac{24}{\widehat{\Delta}}(1+\tau)\right)^{\dim T} \exp\left(-\frac{p}{2}\left(\frac{\widehat{\Delta}^2}{8} - \frac{\widehat{\Delta}^3}{24}\right)\right). \tag{B.7}$$

By $\Sigma_{m,n}(V,W,T,\tau)$ being a closed set the maximum

$$M = \max_{Y \in \Sigma_{m,n}(V,W,T,\tau),\, \|Y\|_F = 1} \|\mathcal{A}(Y)\|_2, \tag{B.8}$$

is attained. Then there exists $Q \in \Lambda$ such that

$$\|\mathcal{A}(X)\|_2 \leq \|\mathcal{A}(X)\|_2 + \|\mathcal{A}(X-Q)\|_2 \leq 1 + \frac{\widehat{\Delta}}{2} + M\frac{\widehat{\Delta}}{4}, \tag{B.9}$$

where the first inequality comes from applying the reverse triangle inequality to X and $Q - X$ and in the second inequality we used (B.6) to upper bound $\|\mathcal{A}(X)\|_2$ and the upper bound of $\|X - Q\|$ comes from $Q \in \Lambda$ combined with Λ being a $(\widehat{\Delta}/4)$-covering. Note that the inequality (B.9) holds for all $X \in \Sigma_{m,n}(V,W,T,\tau)$ whose Frobenius norm $\|X\|_F$ so also for a matrix $\widehat{X}$ for which the maximum in (B.8) is attained. The inequality in (B.9) applied to $\widehat{X}$ yields

$$M \leq 1 + \frac{\widehat{\Delta}}{2} + M\frac{\widehat{\Delta}}{4} \quad \Longrightarrow \quad M \leq 1 + \widehat{\Delta}. \tag{B.10}$$

The lower bound follows from the reverse triangle inequality

$$\|\mathcal{A}(X)\| \geq \|\mathcal{A}(Q)\| - \|\mathcal{A}(X-Q)\| \geq \left(1 - \frac{\widehat{\Delta}}{2}\right) - (1+\widehat{\Delta})\frac{\widehat{\Delta}}{4} \geq 1 - \widehat{\Delta} \tag{B.11}$$

where the second inequality comes from $\|\mathcal{A}(X-Q)\| \leq M\|X-Q\|_F \leq (1+\widehat{\Delta})\frac{\widehat{\Delta}}{4}$ by (B.8) combined with Q being an element of a $(\widehat{\Delta}/4)$-covering.

Combining (B.9) with the bound on M in (B.10) gives the upper bound and (B.11) gives the lower bound on $\|\mathcal{A}(X)\|$ completing the proof. □

Lemma Appendix B.1 (ε-covering of the Grassmannian [33, Theorem 8]). *Let $(\mathcal{G}(D,d), \rho(\cdot,\cdot))$ be a metric space on a Grassmannian manifold $\mathcal{G}(D,d)$ with the metric ρ as defined in* (20)*. Then there exists ε-covering $\mathcal{G}(D,d)$ with $\Lambda = \{U_i\}_{i=1}^N \subset \mathcal{G}(D,d)$ such that*

$$\forall U \in \mathcal{G}(D,d): \quad \min_{\widehat{U} \in \Lambda} \rho(U, \widehat{U}) \leq \varepsilon, \tag{B.12}$$

and $N \leq \left(\frac{C_0}{\varepsilon}\right)^{d(D-d)}$ with C_0 independent of ε, bounded by $C_0 \leq 2\pi$.

The above bound on the covering number of the Grassmannian is used in the following lemma to bound the covering number of the set $\mathrm{LS}_{m,n}(r,s,\tau)$.

Proof of Lemma 2.3 (Covering number of $\mathrm{LS}_{m,n}(r,s,\tau)$), stated on page 10.

Proof. By Lemma Appendix B.1 there exist two finite $(\varepsilon/2)$-coverings $\Lambda_1 := \{V_i\}_{i=1}^{|\Lambda_1|} \subseteq \mathcal{G}(m,r)$ and $\Lambda_2 := \{W_i\}_{i=1}^{|\Lambda_2|} \subseteq \mathcal{G}(n,r)$, with their covering numbers upper bounded as

$$|\Lambda_1| \leq \left(\frac{4\pi}{\varepsilon}\right)^{r(m-r)} \qquad |\Lambda_2| \leq \left(\frac{4\pi}{\varepsilon}\right)^{r(n-r)}, \tag{B.13}$$

as given in [11, (4.18)] that uses [33, Theorem 8]. By Λ_1, Λ_2 being $(\varepsilon/2)$-coverings

$$\forall V \in \mathcal{G}(m,r): \quad \exists V_i \in \Lambda_1, \quad \rho(V, V_1) \leq \varepsilon/2, \tag{B.14}$$

$$\forall W \in \mathcal{G}(n,r): \quad \exists W_i \in \Lambda_2, \quad \rho(W, W_1) \leq \varepsilon/2. \tag{B.15}$$

Let $\Lambda_3 = \mathcal{V}(mn, s)$ where $\mathcal{V}(mn, s)$ is the set of all possible support sets of an $m \times n$ matrix that has s elements. Thus the cardinality of Λ_3 is $\binom{mn}{s}$.

Construct $\Lambda = (\Lambda_1 \times \Lambda_2 \times \Lambda_3)$ where $\times$ denotes the Cartesian product. Choose any $V \in \mathcal{G}(m,r), W \in \mathcal{G}(n,r)$ and $T \in \mathcal{V}(mn,s)$ for which we now show there exists $\left(\widehat{V}, \widehat{W}, \widehat{T}\right) \in \Lambda$ such that $\rho\left((V,W), \left(\widehat{V}, \widehat{W}\right)\right) \leq \varepsilon$ and $T = \hat{T}$, thus showing that the set Λ is an ε-covering of $\mathrm{LS}_{m,n}(r,s,\tau)$.

Satisfying $T = \widehat{T}$ comes from $\Lambda_3 = \mathcal{V}(mn,s)$ containing all support sets with at most s entries. The projection operator onto the pair (V,W) can be written as $P_{(V,W)} = P_V \otimes P_W$, so for the two pairs of subspaces (V,W) and $(\widehat{V}, \widehat{W})$ we have the following

$$\rho\left((V,W), \left(\widehat{V}, \widehat{W}\right)\right) = \|P_{(V,W)} - P_{(\widehat{V},\widehat{W})}\| \tag{B.16}$$

$$= \|P_V \otimes P_W - P_{\widehat{V}} \otimes P_{\widehat{W}}\| \tag{B.17}$$

$$= \|\left(P_V - P_{\widehat{V}}\right) \otimes P_W + P_{\widehat{V}}\left(P_W - P_{\widehat{W}}\right)\| \tag{B.18}$$

$$\leq \|P_V - P_{\widehat{V}}\|\|P_W\| + \|P_{\widehat{V}}\|\|P_W - P_{\widehat{W}}\| \tag{B.19}$$

$$= \rho\left(V, \widehat{V}\right) + \rho\left(W, \widehat{W}\right). \tag{B.20}$$

By Λ_1 and Λ_2 being $(\varepsilon/2)$-coverings, we have that for any V, W exist $\widehat{V} \in \Lambda_1$ and $\widehat{W} \in \Lambda_2$, such that $\rho\left((V,W), \left(\widehat{V}, \widehat{W}\right)\right) \leq \rho\left(V, \widehat{V}\right) + \rho\left(W, \widehat{W}\right) \leq \varepsilon$. Using the bounds on the cardinality of Λ_1, Λ_2 in (B.13) combined with $|\Lambda_3| = \binom{mn}{s}$ yields that the cardinality of Λ is bounded above by

$$\mathfrak{R}(\varepsilon) = |\Lambda_1|\,|\Lambda_2|\,|\Lambda_3| \leq \binom{mn}{s}\left(\frac{4\pi}{\varepsilon}\right)^{r(m+n-2r)}. \tag{B.21}$$

$\square$

Proof of Lemma 2.2 (Variation of $\widehat{\Delta}$ in RIC in respect to a perturbation of (V,W)), stated on page 10.

Proof. Recall the notation used in Lemma 2.2 that there are sets $\Sigma_1 := \Sigma_{m,n}(V_1, W_1, T, \tau)$ and $\Sigma_2 := \Sigma_{m,n}(V_2, W_2, T, \tau)$ which are subsets of $\mathrm{LS}_{m,n}(r,s,\tau)$ with a shared support T of the sparse component.

Let $Y \in \Sigma_2$, so we can write $Y = L + S$ such that $\mathrm{supp}(S) = T, \mathcal{C}(L) \subseteq V_2, \mathcal{C}(L^T) \subseteq W_2$ and $\|L\|_F \leq \tau$. By linearity of $\mathcal{A}$ assume without loss of generality that $\|Y\|_F = 1$. Denote $U_1 = (V_1, W_1)$ and $U_2 = (V_2, W_2)$ and let P_{U_i} be an orthogonal projection onto the space of matrices whose column and row space is defined by V_i, W_i such that left and right singular vectors of $P_{U_i}Y$ lie in V_i respectively W_i. Then

$$
\begin{aligned}
\|\mathcal{A}(Y)\| &= \|\mathcal{A}(L+S)\| = \|\mathcal{A}\left(P_{U_1}L + S - (P_{U_1}L - P_{U_2}L)\right)\| && \text{(B.22)}\\
&\leq \|\mathcal{A}\left(P_{U_1}L + S\right)\| + \|\mathcal{A}\left([P_{U_1} - P_{U_2}]\,L\right)\| && \text{(B.23)}\\
&\leq (1+\widehat{\Delta})\,\|P_{U_1}L + S\| + \|\mathcal{A}\|\rho\,(U_1,U_2)\,\|L\| && \text{(B.24)}\\
&= (1+\widehat{\Delta})\,\|P_{U_2}L + S + [P_{U_1} - P_{U_2}]\,L\| + \|\mathcal{A}\|\rho\,(U_1,U_2)\,\|L\| && \text{(B.25)}\\
&\leq (1+\widehat{\Delta})\,(\|Y\|_F + \rho(U_1,U_2)\|L\|) + \|\mathcal{A}\|\rho\,(U_1,U_2)\,\|L\| && \text{(B.26)}\\
&\leq \|Y\|_F\left(1+\widehat{\Delta} + \tau\rho(U_1,U_2)\left(1+\widehat{\Delta}+\|\mathcal{A}\|\right)\right), && \text{(B.27)}
\end{aligned}
$$

where in the first line (B.22) we use the fact that $P_{U_2}L = L$, the second line (B.23) follows by the triangle inequality and linearity of $\mathcal{A}$, and in the third inequality we bound the effect of $\mathcal{A}$ on $(P_{U_1}L + S)$ using the RICs of $\mathcal{A}$ combined with the definition of ρ in (20). We proceed in (B.25) and (B.26) by projecting L to space U_2 and again bounding the effect of $\mathcal{A}$ on $(P_{U_2}L + S)$. Finally, in (B.27) we use $\|L\|_F \leq \tau$. We obtain a similar lower bound using the reverse triangular inequality

$$
\begin{aligned}
\|\mathcal{A}(Y)\| &= \|\mathcal{A}\left(P_{U_1}L + S - (P_{U_1}L - P_{U_2}L)\right)\| && \text{(B.28)}\\
&\geq \|\mathcal{A}\left(P_{U_1}L + S\right)\| - \|\mathcal{A}\left([P_{U_1} - P_{U_2}]\,L\right)\| && \text{(B.29)}\\
&\geq \left(1-\widehat{\Delta}\right)\|P_{U_1}L + S\| - \|\mathcal{A}\|\rho(U_1,U_2)\|L\|_F && \text{(B.30)}\\
&= \left(1-\widehat{\Delta}\right)\|P_{U_2}L + S - [P_{U_2} - P_{U_1}]\,L\| - \|\mathcal{A}\|\rho(U_1,U_2)\|L\|_F && \text{(B.31)}\\
&\geq \left(1-\widehat{\Delta}\right)(\|Y\|_F - \rho(U_1,U_2)\|L\|_F) - \|\mathcal{A}\|\rho(U_1,U_2)\|L\|_F && \text{(B.32)}\\
&\geq \|Y\|_F\left(1-\widehat{\Delta} - \tau\rho(U_1,U_2)(1-\widehat{\Delta}+\|\mathcal{A}\|)\right). && \text{(B.33)}
\end{aligned}
$$

Combining (B.27) and (B.33) yields

$$
\forall Y \in \Sigma_2: \quad (1-\widehat{\Delta}')\|Y\|_F \leq \|\mathcal{A}(Y)\| \leq (1+\widehat{\Delta}')\|Y\|_F, \tag{B.34}
$$

with $\widehat{\Delta}' = \widehat{\Delta} + \tau\rho(U_1,U_2)\left(1+\widehat{\Delta}+\|\mathcal{A}\|\right)$. □

In the proof of Theorem 3 we make use of the following Lemma Appendix B.2 and Corollary Appendix B.1 from [11] which we restate here for completeness.

Lemma Appendix B.2 ([11, Lemma 3.4]). *Let A and B be matrices of the same dimensions. Then there exist matrices B_1 and B_2 such that*

(1) $B = B_1 + B_2$,

(2) $\operatorname{rank} B_1 \leq 2 \operatorname{rank} A$,

(3) $AB_2^T = 0$ *and* $A^T B_2 = 0$,

(4) $\langle B_1,\, B_2\rangle = 0$.

Corollary Appendix B.1 ([11, Lemma 2.3])**.** *Let A and B be matrices of the same dimensions. If $AB^T = 0$ and $A^T B = 0$, then $\|A+B\|_* = \|A\|_* + \|B\|_*$.*

Lemma Appendix B.3 (Decomposing $R^S = R_0^S + R_c^S$)**.** *Let $\operatorname{supp} S_0 = \Omega_0$ and construct a matrix R_0^S that has the entries of R^S at indices Ω_0*

$$(R_0^S)_{i,j} = \begin{cases} (R^S)_{i,j} & \textit{if} \quad (i,j) \in \Omega_0, \\ 0 & \textit{if} \quad (i,j) \notin \Omega_0, \end{cases} \tag{B.35}$$

and a matrix $R_c^S = R^S - R_0^S$ that has the entries of R^S at the indices of the complement of Ω_0. Then

(1) $\|R_0^S\|_0 \le \|S_0\|_0 = s$ *(by $|\Omega_0| = s$)*,

(2) $\|S_0 + R_c^S\|_1 = \|S_0\|_1 + \|R_c^S\|_1$ *(by $\operatorname{supp}(R_0^S) \cap \operatorname{supp}(R_c^S) = \emptyset$)*,

(3) $\langle R_0^S, R_c^S\rangle = 0$ *(by $\operatorname{supp}(R_0^S) \cap \operatorname{supp}(R_c^S) = \emptyset$)*.

Proof. It can be easily verified that R_0^S and R_c^S constructed as in (B.35) satisfy the conditions (1)-(3). □

Lemma Appendix B.4 (Upper bound on $\langle \mathcal{A}(\cdot), \mathcal{A}(\cdot)\rangle$)**.** *For an operator $\mathcal{A}(\cdot)$ whose RICs are upper bounded by $\Delta := \Delta_{2r,2s,2\tau}$ we have that*

$$|\langle \mathcal{A}(X), \mathcal{A}(Y)\rangle| \le \Delta \|X\|_F \|Y\|_F, \qquad \forall X, Y \in \mathrm{LS}_{m,n}(r,s,\tau). \tag{B.36}$$

Proof. By linearity of $\mathcal{A}$ we can assume without loss of generality that $\|X\|_F = \|Y\|_F = 1$. Applying the parallelogram law on $\|\mathcal{A}(X)\|_2$ and $\|\mathcal{A}(Y)\|_2$ gives

$$2\left(\|\mathcal{A}(X)\|_2^2 + \|\mathcal{A}(Y)\|_2^2\right) = \|\mathcal{A}(X) + \mathcal{A}(Y)\|_2^2 + \|\mathcal{A}(X) - \mathcal{A}(Y)\|_2^2. \tag{B.37}$$

Subtract $2\|\mathcal{A}(X) - \mathcal{A}(Y)\|_2^2$ from both sides of (B.37) gives

$$4\langle \mathcal{A}(X), \mathcal{A}(Y)\rangle = \|\mathcal{A}(X) + \mathcal{A}(Y)\|_2^2 - \|\mathcal{A}(X) - \mathcal{A}(Y)\|_2^2. \tag{B.38}$$

Using RICs of $\mathcal{A}(\cdot)$ and that $X+Y$ and $X-Y$ are in the set $\mathrm{LS}_{2r,2s,2\tau}$ we bound on the right hand side (B.38)

$$|\langle \mathcal{A}(X), \mathcal{A}(Y)\rangle| = \frac{1}{4}\left|\|\,\mathcal{A}(X+Y)\|_F^2 + \|\,\mathcal{A}(X-Y)\|_F^2\right| \tag{B.39}$$

$$\le \frac{1}{4}\left|(1+\Delta)\|X+Y\|_F - (1-\Delta)\|X-Y\|_F\right| \tag{B.40}$$

$$\le \frac{1}{2}\left|(1+\Delta) - (1-\Delta)\right| \le \Delta, \tag{B.41}$$

where in the third line we used that $\|X+Y\|_F \leq 2$ and $\|X-Y\|_F \leq 2$. Note that the bound can be lowered for specific matrices X, Y such that the matrices of their sums $X+Y$ and $X-Y$ are in LS
530 sets with smaller ranks, sparsities or Frobenius norm of their low-rank component. □

Remark Appendix B.1 (Bounding the residual of the sparse component)**.** *Herein we derive inequality in* (80) *as was done in* (70) *to* (73) *for the low-rank component of the error.*

Proof. Let $\Delta := \Delta_{4r,3s,2\tau}$ be an RIC with squared norms for $\mathrm{LS}_{m,n}(4r, 3s, 2\tau)$. Then let R_0^L, R_1^L and R_0^S, R_1^S be defined as above equation (56) and (60) respectively

$$(1-\Delta)\|R_0^S + R_1^S\|_F^2 \leq \|\mathcal{A}(R_0^S + R_1^S)\|_2^2 = \left|\left\langle \mathcal{A}(R_0^S + R_1^S),\, \mathcal{A}(R_0^S + R_1^S - R + R)\right\rangle\right| \tag{B.42}$$

$$= \left|\left\langle \mathcal{A}(R_0^S + R_1^S),\, \mathcal{A}(R_0^S + R_1^S - R)\right\rangle + \left\langle \mathcal{A}(R_0^S + R_1^S),\, \mathcal{A}(R)\right\rangle\right| \tag{B.43}$$

$$\leq \left|\left\langle \mathcal{A}\left(R_0^S + R_1^S\right), \mathcal{A}\left(-R_0^L - R_1^L - \sum_{j\geq 2} R_j\right)\right\rangle\right| + \left|\left\langle \mathcal{A}(R_0^S + R_1^S),\, \mathcal{A}(R)\right\rangle\right| \tag{B.44}$$

$$\leq \Delta\|R_0^S + R_1^S\|_F \left(\|R_0^L + R_1^L\|_F + \sum_{j\geq 2}\|R_j\|_F\right) + \left\|\mathcal{A}\left(R_0^S + R_1^S\right)\right\|_2 \|\mathcal{A}(R)\|_2, \tag{B.45}$$

where the inequality in the first line comes from $R_0^S + R_1^S \in \mathrm{LS}(0, 2s, 0) \subset \mathrm{LS}(4r, 3s, 2\tau)$ satisfying the RIC, the second line is the consequence of feasibility in (69), the third line comes from Lemma
535 Appendix B.4 and by sums of individual pairs in the inner product being in $\mathrm{LS}(4r, 3s, 2\tau)$.

The first term in (B.45) can be bounded as

$$\Delta\|R_0^S + R_1^S\|_F \left(\|R_0^L + R_1^L\|_F + \sum_{j\geq 2}\|R_j\|_F\right) \tag{B.46}$$

$$\leq \Delta\|R_0^S + R_1^S\|_F \left(\|R_0^L + R_1^L\|_F + \sqrt{2}\|R_0^L\|_F + \|R_0^S\|_F\right) \tag{B.47}$$

$$\leq \Delta\|R_0^S + R_1^S\|_F \left((1+\sqrt{2})\|R_0^L + R_1^L\|_F + \|R_0^S + R_1^S\|_F\right) \tag{B.48}$$

where the second line comes as a consequence of optimality in (68) with $M_r = r$ and $M_s = s$, and the third line comes from $\|R_0^L\|_F \leq \|R_0^L + R_1^L\|_F$ and $\|R_1^L\|_F \leq \|R_0^L + R_1^L\|_F$. Having bounded the first term in (B.45) we now move to upper bounding the second term in (B.45) which comes as a consequence of feasibility bound in (69) and of the RIC for $R_0^L + R_1^L \in \mathrm{LS}_{m,n}(4r, 3s, 2\tau)$

$$\left\|\mathcal{A}\left(R_0^S + R_1^S\right)\right\|_2 \|\mathcal{A}(R)\|_2 \leq \varepsilon_{\mathrm{b}}(1+\Delta)\|R_0^S + R_1^S\|_F. \tag{B.49}$$

We now bound (B.45) by combining the upper bounds of its constituents in (B.48) and (B.49)

$$(1-\Delta)\|R_0^S + R_1^S\|_F^2 \leq \Delta\|R_0^S + R_1^S\|_F \left((1+\sqrt{2})\|R_0^L + R_1^L\|_F + \|R_0^S + R_1^S\|_F + \varepsilon_{\mathrm{b}}\frac{1+\Delta}{\Delta}\right), \tag{B.50}$$

which after dividing both sides by $(1-\Delta)\|R_0^S + R_1^S\|_F$ yields the inequality in (80). □

Lemma Appendix B.5. *Let X^j, X^{j+1}, X_0 be any matrices in the set $\mathrm{LS}_{m,n}(r,s,\tau)$, $\alpha_j \geq 0$, and $\mathcal{A}(\cdot)$ be an operator whose RICs are sufficiently upper bounded, then the following two inequalities hold*

$$\begin{aligned}\langle X^j - X_0, X^{j+1} - X_0\rangle - \alpha_j \langle \mathcal{A}(X^j - X_0), \mathcal{A}(X^{j+1} - X_0)\rangle \\ \leq \|I - \alpha_j\, A_Q^T A_Q\|_2 \|X^j - X_0\|_F \|X^{j+1} - X_0\|_F,\end{aligned} \tag{B.51}$$

and

$$\|X^j - X_0 - \alpha_j\, \mathcal{A}^*\left(\mathcal{A}\left(X^j - X_0\right)\right)\|_F \leq \|I - \alpha_j\, A_Q^T A_Q\|_2 \|X^j - X_0\|_F, \tag{B.52}$$

where the spectrum of the matrix $\left(I - \alpha_j\, A_Q^T A_Q\right) \in \mathbb{R}^{mn \times mn}$ is bounded as

$$1 - \alpha_j\left(1 + \Delta_{3r,3s,2\tau}\right) \leq \lambda\left(I - A_Q^T A_Q\right) \leq 1 + \alpha_j\left(1 - \Delta_{3r,3s,2\tau}\right), \tag{B.53}$$

which gives an upper bound on the norm $\|I - \alpha_j\, A_Q^T A_Q\|_2 \leq |1 - \alpha_j\left(1 + \Delta_{3r,3s,2\tau}\right)|$ as the lower bound in (B.53) is larger then the upper bound.

Proof. We vectorize the matrices on the left hand side of (B.51) using a mapping $\mathrm{vec}(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ that stacks columns of a given matrix into a vector and a mapping $\mathrm{mat}(\cdot)$ from the space of linear operators $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ to the space of matrices of size $p \times mn$

$$\begin{aligned}x_0 = \mathrm{vec}\left(X_0\right), \; x^j = \mathrm{vec}\left(X^j\right), \; x^{j+1} = \mathrm{vec}\left(X^{j+1}\right) \in \mathbb{R}^{mn} \\ A = \mathrm{mat}\left(\mathcal{A}\right) = \begin{bmatrix} — & \mathrm{vec}\left(A_1\right)^T & — \\ & \vdots & \\ — & \mathrm{vec}\left(A_p\right)^T & — \end{bmatrix} \in \mathbb{R}^{p\times mn}.\end{aligned} \tag{B.54}$$

Let $X_0 = U^0\Sigma^0V^0 + S^0$, $X^j = U^j\Sigma^jV^j + S^j$, $X^{j+1} = U^{j+1}\Sigma^{j+1}V^{j+1} + S^{j+1}$ be the singular value decompositions where the matrices of the left singular vectors are $U^j \in \mathbb{R}^{n\times r}$ and their sparse components are supported at indices $\Omega^j = \mathrm{supp}\left(S^j\right)$. Consider the union of the index sets $\Omega := \left\{\Omega^0, \Omega^j, \Omega^{j+1}\right\}$ and construct the following orthogonal frame

$$Q = [I_n \otimes U \quad E] = \left[\begin{matrix} U & 0_{n,3r} & \dots & 0_{n,3r} \\ 0_{n,3r} & U & \dots & 0_{n,3r} \\ \vdots & & \ddots & \\ 0_{n,3r} & \dots & \dots & U \end{matrix} \;\middle|\; \begin{matrix} | & & | \\ | & & | \\ e_{\Omega_1} & \dots & e_{\Omega_{3s}} \\ | & & | \end{matrix}\right] \in \mathbb{R}^{n^2\times 3(nr+s)}, \tag{B.55}$$

where $U \in \mathbb{R}^{n\times 3r}$ is formed by concatenating U^0, U^j, U^{j+1} and e_{Ω_i} is a vector corresponding to a vectorized matrix with a single entry 1 at the index Ω_i. Note that $P_Q = Q\left(Q^TQ\right)^{-1}Q^T$ is an orthogonal projection matrix on the low-rank plus sparse subspace defined by the matrix U and the index set Ω. Note that by Q being formed by the low-rank plus sparse bases of X_0, X^j, X^{j+1} we have that the projection does not change the vectorized matrices

$$P_Q x_0 = x_0, \quad P_Q x^j = x^j, \quad P_Q x^{j+1} = x^{j+1}. \tag{B.56}$$

To establish the bound in (B.51) we write the left hand side in its vectorized form

$$\left(x^j - x_0\right)^T \left(x^{j+1} - x_0\right) - \alpha_j \left(A(x^j - x_0)\right)^T \left(A(x^{j+1} - x_0)\right), \tag{B.57}$$

and replacing A with $A_Q = AP_Q$ in (B.57) using the identities in (B.56) simplifies the term as

$$\begin{aligned}
&\left(x^j - x_0\right)^T \left(x^{j+1} - x_0\right) - \alpha_j \left(A_Q(x^j - x_0)\right)^T \left(A_Q(x^{j+1} - x_0)\right) && \text{(B.58)}\\
&= \left(x^j - x_0\right)^T \left((x^{j+1} - x_0) - \alpha_j\, A_Q^* A_Q (x^{j+1} - x_0)\right) && \text{(B.59)}\\
&= \left(x^j - x_0\right)^T \left((I - \alpha_j\, A_Q^* A_Q)(x^{j+1} - x_0)\right) && \text{(B.60)}\\
&\leq \|I - \alpha_j\, A_Q^* A_Q\|_2 \, \|x^j - x_0\|_2 \, \|x^{j+1} - x_0\|_2 && \text{(B.61)}\\
&= \|I - \alpha_j\, A_Q^* A_Q\|_2 \, \|X^j - X_0\|_F \, \|X^{j+1} - X_0\|_F, && \text{(B.62)}
\end{aligned}$$

where $\|I - \alpha_j\, A_Q^* A_Q\|_2$ is the ℓ_2 operator norm of an $mn \times mn$ matrix.

Similarly we now establish the bound in (B.52)

$$\begin{aligned}
\left\|X^j - X_0 - \alpha_j\, \mathcal{A}^* \left(\mathcal{A}\left(X^j - X_0\right)\right)\right\|_F &= \left\|x^j - x_0 + \alpha_j\, A^T A \left(x_0 - x^j\right)\right\|_2 && \text{(B.63)}\\
&= \left\|\left(I - \alpha_j\, A^T A\right)\left(x^j - x_0\right)\right\|_2 && \text{(B.64)}\\
&\leq \left\|I - \alpha_j\, A_Q^* A_Q\right\|_2 \left\|X^j - X_0\right\|_F, && \text{(B.65)}
\end{aligned}$$

where we just vectorized the matrices and the linear operator $\mathcal{A}(\cdot)$ and upper bounded the expression using ℓ_2 operator norm $\|I - \alpha_j\, A_Q^* A_Q\|_2$. Matrix A_Q acts on a subspace of $\mathrm{LS}_{m,n}(3r, 3s, 2\tau)$ and is self-adjoint, as such its eigenvalues can be bounded using the RICs as was done in [28, 40]

$$1 - \alpha_j \left(1 + \Delta_{3r,3s,2\tau}\right) \leq \lambda\left(I - A_Q^* A_Q\right) \leq 1 + \alpha_j \left(1 - \Delta_{3r,3s,2\tau}\right). \tag{B.66}$$

□