Spring 2018

Non-convex methods for spectrally sparse signal reconstruction via low-rank Hankel matrix completion

Tianming Wang
University of Iowa

NON-CONVEX METHODS FOR SPECTRALLY SPARSE SIGNAL RECONSTRUCTION VIA LOW-RANK HANKEL MATRIX COMPLETION

by

Tianming Wang

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy degree
in Applied Mathematical and Computational Sciences
in the Graduate College of
The University of Iowa

May 2018

Thesis Supervisors: Associate Professor Jian-Feng Cai
Assistant Professor Weiyu Xu

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Tianming Wang

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Applied Mathematical and Computational Sciences at the May 2018 graduation.

Thesis Committee: _____
Jian-Feng Cai, Thesis Supervisor

_____
Weiyu Xu, Thesis Supervisor

_____
Weimin Han

_____
David Stewart

_____
Xueyu Zhu

To my family

ACKNOWLEDGEMENTS

I sincerely thank my advisor, Professor Jian-Feng Cai. None of my achievements during these years would have been accomplished without his support. I am grateful to my co-advisor, Professor Weiyu Xu, for his guidance and encouragement.

I especially thank Professor Weimin Han for all the supports during my study at the University of Iowa. I also thank the rest of my Committee, Professor David Stewart and Professor Xueyu Zhu, who have made me feel supported one way or another.

ABSTRACT

Spectrally sparse signals arise in many applications of signal processing. A spectrally sparse signal is a mixture of a few undamped or damped complex sinusoids. An important problem from practice is to reconstruct such a signal from partial time domain samples. Previous convex methods have the drawback that the computation and storage costs do not scale well with respect to the signal length. This common drawback restricts their applicabilities to large and high-dimensional signals.

The reconstruction of a spectrally sparse signal from partial samples can be formulated as a low-rank Hankel matrix completion problem. We develop two fast and provable non-convex solvers, FIHT and PGD. FIHT is based on Riemannian optimization while PGD is based on Burer-Monteiro factorization with projected gradient descent. Suppose the underlying spectrally sparse signal is of model order r and length n. We prove that $O(r^2 \log^2(n))$ and $O(r^2 \log(n))$ random samples are sufficient for FIHT and PGD respectively to achieve exact recovery with overwhelming probability. Every iteration, the computation and storage costs of both methods are linear with respect to signal length n. Therefore they are suitable for handling spectrally sparse signals of large size, which may be prohibited for previous convex methods. Extensive numerical experiments verify their recovery abilities as well as computation efficiency, and also show that the algorithms are robust to noise and mis-specification of the model order. Comparing the two solvers, FIHT is faster for easier problems while PGD has a better recovery ability.

PUBLIC ABSTRACT

In many applications of signal processing, the underlying signals can be modeled well using relatively few parameters. The problem at hand is when we only have partial information of such a signal, can we completely determine it? The answer is affirmative, and it is due to the inherent simple structure of the signal. Previous methods are based on convex optimization. A common drawback is that they cannot handle signals of large size.

We convert the problem to determining a low-rank and structured matrix instead. Two non-convex optimization based methods, FIHT and PGD, are proposed. Compared to convex methods, non-convex methods are desirable for their computation efficiency and low storage cost. The proposed methods exploit the low rankness and the structure property from different perspectives. They not only are suitable for practical applications, but also have theoretically guaranteed performances. By comparison, FIHT is faster for relatively easy problems while PGD works better for relatively hard problems.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

1.1 Spectrally Sparse Signal Reconstruction

In this thesis, we are interested in the problem of reconstructing a spectrally sparse signal with or without damping from its incomplete time-domain samples. Let $x(t)$ be a one-dimensional signal. We say that $x(t)$ is spectrally sparse if it is superposition of a few complex sinusoids, namely

$$x(t) = \sum_{k=1}^{r} d_k e^{(2\pi\imath f_k - \tau_k)t}, \tag{1.1}$$

where $\imath = \sqrt{-1}$, r is the model order, f_k is the frequency of each sinusoid, d_k is the weight of each sinusoid, and $\tau_k \geq 0$ is a damping factor. Let $n > 0$ be a natural number. Without loss of generality, we assume $f_k \in [0, 1)$ and consider the samples of $x(t)$ at all the integer values from 0 to $n-1$, denoted $\boldsymbol{x}$. That is,

$$\boldsymbol{x} = \begin{bmatrix} x(0) & \cdots & x(n-1) \end{bmatrix}^T \in \mathbb{C}^n. \tag{1.2}$$

Spectrally sparse signals of the form (1.1) and the corresponding sampling model in (1.2) arise in many areas of science and engineering including magnetic resonance imaging [38], fluorescence microscopy [46], radar imaging [43], nuclear magnetic resonance spectroscopy [44], and analog-to-digital conversion [52]. However, in those real-world applications, full sampling at all the points on a uniform grid is ei-

ther time-consuming or technically prohibited. In addition, the signal may become too weak to be detected after a certain period of time when $\tau_k > 0$. Therefore, for the purpose of more efficient data acquisition, not all the discrete samples are obtained. When restricted to the sampling model in (1.2), this means that only partial entries of $\boldsymbol{x}$ are known and we need to estimate the missing ones. Let Ω be subset of $\{0, \cdots, n-1\}$ corresponding to the observed entries, and let $\mathcal{P}_\Omega$ be the associated sampling operator which acquires only the entries indexed by Ω. Then the task can be formally expressed as:

$$\text{Find} \quad \boldsymbol{x} \quad \text{subject to} \quad \mathcal{P}_\Omega(\boldsymbol{x}) = \sum_{a \in \Omega} x_a \boldsymbol{e}_a, \tag{1.3}$$

where $\{\boldsymbol{e}_a\}_{a=0}^{n-1}$ is a canonical basis of $\mathbb{C}^n$.

1.2 Overview of Related Works

It is clear that (1.3) is a task that cannot be achieved if $\boldsymbol{x}$ does not have any intrinsic simple structures. Fortunately, the signal of interest is spectrally sparse. Moreover, the number of degrees of freedom in $\boldsymbol{x}$ is completely determined by the number of Fourier modes in $x(t)$, which is proportional to r and independent of n. This key observation suggests the possibility of reconstructing $\boldsymbol{x}$ from its partial revealed entries, which can be further achieved by exploiting the simplicity of $\boldsymbol{x}$ in different ways.

Note that we are mainly interested in the scenario where $\boldsymbol{x}$ only has a few Fourier components (i.e., r is small). Thus, one can utilize the sparsity of $\boldsymbol{x}$ in

the frequency domain to design reconstruction algorithms. In particular, if there is no damping in $\boldsymbol{x}$, spectral compressed sensing can be recast as a conventional compressed sensing problem [11, 21] after discretization of the Fourier domain; so many existing algorithms for compressed sensing are available, such as Basis Pursuit [14], IHT [2–5, 24], CoSaMP [41] and SP [19]. However, the performance of the compressed sensing approach for spectrally sparse signal recovery suffers from the mismatch error between the true frequencies and the discrete frequencies [18, 30]. A grid-free approach was developed in [49] which exploited the frequency sparsity of $\boldsymbol{x}$ in a continuous manner via the atomic norm minimization (ANM). It was shown in [49] that ANM could achieve exact recovery from $O(r\log(r)\log(n))$ random time-domain samples under some mild conditions.

By the Vandermonde decomposition, one may easily see that the Hankel matrix computed from a spectrally sparse signal is low rank when r is small relative to n. Consequently, spectral compressed sensing can be reformulated as a low rank Hankel matrix completion problem, see the next section for details. Inspired by low rank matrix completion [10], another grid-fee method known as enhanced matrix completion (EMaC) was developed in [15] by reformulating the non-convex low rank Hankel matrix completion problem into a convex Hankel matrix nuclear norm minimization problem. EMaC was shown to be able to reconstruct a spectrally sparse signal with high probability provided the number of observed entries is $O(r\log^4(n))$. The same approach was studied in [7] under the Gaussian random sampling model, and various first-order methods were discussed in [22] for the regularized Hankel matrix nuclear

norm minimization problem. For multi-dimensional spectrally sparse signal recovery problems, we can also exploit the low rank tensor structure of the signal when developing recovery algorithms, see for example [60] and references therein.

1.3 Low-Rank Hankel Matrix Completion

For a spectrally sparse signal, we can exploit its simplicity via the low rank structure of the corresponding Hankel matrix. Recall that a Hankel matrix is a matrix in which each skew-diagonal from left to right is constant. We define $\mathcal{H}$ as a linear operator which maps a vector $\boldsymbol{z} \in \mathbb{C}^n$ to an $n_1 \times n_2$ $(n_1 + n_2 - 1 = n)$ Hankel matrix, denoted $\mathcal{H}\boldsymbol{z}$, whose i-th skew-diagonal is equal to the i-th entry of $\boldsymbol{z}$,

$$\mathcal{H}\boldsymbol{z} = \begin{bmatrix} z_0 & z_1 & z_2 & \cdots & \cdots & z_{n_2-1} \\ z_1 & z_2 & \cdots & \cdots & \cdots & z_{n_2} \\ z_2 & \cdots & \cdots & \cdots & \cdots & z_{n_2+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n_1-1} & z_{n_1} & \cdots & \cdots & \cdots & z_{n-1} \end{bmatrix}.$$

Thus, one has $[\mathcal{H}\boldsymbol{z}]^{(i,j)} = z_{i+j}$ for $i \in [n_1]$ and $j \in [n_2]$. In particular, the (i,j)-th entry of the Hankel matrix formed from the spectrally sparse signal $\boldsymbol{x}$ is given by

$$[\mathcal{H}\boldsymbol{x}]^{(i,j)} = x_{i+j} = \sum_{k=1}^{r} d_k e^{(2\pi\imath f_k - \tau_k)(i+j)} = \sum_{k=1}^{r} d_k e^{i(2\pi\imath f_k - \tau_k)} e^{j(2\pi\imath f_k - \tau_k)}.$$

If we let $w_k = e^{(2\pi\imath f_k - \tau_k)}$ for $k = 1, \cdots, r$, it follows immediately that $\mathcal{H}\boldsymbol{x}$ admits the following Vandermonde decomposition:

$$\mathcal{H}\boldsymbol{x} = \boldsymbol{E}_L \boldsymbol{D} \boldsymbol{E}_R^T,$$

where

$$\boldsymbol{E}_L = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ w_1 & w_2 & \cdots & w_r \\ \vdots & \vdots & \vdots & \vdots \\ w_1^{n_1-1} & w_2^{n_1-1} & \cdots & w_r^{n_1-1} \end{bmatrix}, \ \boldsymbol{E}_R = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ w_1 & w_2 & \cdots & w_r \\ \vdots & \vdots & \vdots & \vdots \\ w_1^{n_2-1} & w_2^{n_2-1} & \cdots & w_r^{n_2-1} \end{bmatrix}$$

and $\boldsymbol{D} = \mathrm{diag}(d_1, \cdots, d_r)$. Moreover, one has $\mathrm{rank}(\mathcal{H}\boldsymbol{x}) = r$ provided the frequencies $\{f_k\}_{k=1}^r$ are different with each other and the diagonal entries of $\boldsymbol{D}$ are all nonzeros. Since $\mathcal{H}$ is injective, the reconstruction of $\boldsymbol{x}$ from $\mathcal{P}_\Omega(\boldsymbol{x})$ is equivalent to the reconstruction of $\mathcal{H}\boldsymbol{x}$ from partial revealed anti-diagonals that corresponds to the known entries of $\boldsymbol{x}$. With a slight abuse of notation we also use $\mathcal{P}_\Omega$ to denote the projection of a matrix $\boldsymbol{Z} \in \mathbb{C}^{n_1 \times n_2}$ onto the subspace determined by a subset of an orthonormal basis of Hankel matrices; that is,

$$\mathcal{P}_\Omega(\boldsymbol{Z}) = \sum_{a \in \Omega} \langle \boldsymbol{Z}, \boldsymbol{H}_a \rangle \, \boldsymbol{H}_a,$$

where the set of matrices

$$\left\{ \boldsymbol{H}_a = \frac{1}{\sqrt{w_a}} \mathcal{H}\boldsymbol{e}_a \mid w_a = \#\left\{(i,j) \mid i+j=a\right\} \right\}_{a=0}^{n-1} \tag{1.4}$$

forms an orthonormal basis of $n_1 \times n_2$ Hankel matrices. To reconstruct $\mathcal{H}\boldsymbol{x}$, we seek a rank r Hankel matrix consistent with the revealed anti-diagonals by solving the following *low rank Hankel matrix completion* problem

$$\text{Find} \quad \mathcal{H}\boldsymbol{z} \quad \text{subject to} \quad \operatorname{rank}(\mathcal{H}\boldsymbol{z}) = r \text{ and } \mathcal{P}_\Omega(\mathcal{H}\boldsymbol{z}) = \mathcal{P}_\Omega(\mathcal{H}\boldsymbol{x}). \tag{1.5}$$

The key insight in matrix completion suggests that in order to achieve successful low rank Hankel matrix completion, it requires the singular vectors of the underlying Hankel matrix $\mathcal{H}\boldsymbol{x}$ are not aligned with the orthonormal basis $\{\boldsymbol{H}_a\}_{a=0}^{n-1}$. This can be guaranteed if the smallest singular values of the left matrix $\boldsymbol{E}_L$ and the right matrix $\boldsymbol{E}_R$ in the Vandermonde decomposition of $\mathcal{H}\boldsymbol{x}$ are bounded away from zero.

Definition 1.3.1. *The rank r Hankel matrix $\mathcal{H}\boldsymbol{x}$ with the Vandermonde decomposition $\mathcal{H}\boldsymbol{x} = \boldsymbol{E}_L \boldsymbol{D} \boldsymbol{E}_R^T$ is said to be μ_0-incoherent if there exists a numerical constant $\mu_0 > 0$ such that*

$$\sigma_{\min}(\boldsymbol{E}_L^* \boldsymbol{E}_L) \geq \frac{n_1}{\mu_0}, \ \sigma_{\min}(\boldsymbol{E}_R^* \boldsymbol{E}_R) \geq \frac{n_2}{\mu_0}.$$

This incoherence property was introduced in [15] and is crucial to derive the-

oretical guarantees. We know from [37, Thm. 2] that, in the undamped case, if the minimum wrap-around distance between the frequencies is greater than about $2/n$, this property can be satisfied. Let $\mathcal{H}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the reduced SVD of $\mathcal{H}\boldsymbol{x}$ and $\mathcal{P}_{\boldsymbol{U}}(\cdot)$ and $\mathcal{P}_{\boldsymbol{V}}(\cdot)$ respectively be the orthogonal projections onto the subspaces spanned by $\boldsymbol{U}$ and $\boldsymbol{V}$. The following lemma follows directly from Def. 1.3.1.

Lemma 1.3.1. *Let $\mathcal{H}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{E}_L\boldsymbol{D}\boldsymbol{E}_R^T$. Assume $\mathcal{H}\boldsymbol{x}$ is μ_0 incoherent and define $c_s = \max\left\{\frac{n}{n_1}, \frac{n}{n_2}\right\}$. Then*

$$\left\|\boldsymbol{U}^{(i,:)}\right\|^2 \leq \frac{\mu_0 c_s r}{n} \quad \textit{and} \quad \left\|\boldsymbol{V}^{(j,:)}\right\|^2 \leq \frac{\mu_0 c_s r}{n}, \tag{1.6}$$

$$\|\mathcal{P}_{\boldsymbol{U}}(\boldsymbol{H}_a)\|_F^2 \leq \frac{\mu_0 c_s r}{n} \quad \textit{and} \quad \|\mathcal{P}_{\boldsymbol{V}}(\boldsymbol{H}_a)\|_F^2 \leq \frac{\mu_0 c_s r}{n}. \tag{1.7}$$

Proof. The proof of (1.7) can be found in [15]. We include the proof here to be self-contained. We only prove the left inequalities of (1.6) and (1.7) as the right ones can be similarly established. Since $\boldsymbol{U} \in \mathbb{C}^{n_1 \times r}$ and $\boldsymbol{E}_l \in \mathbb{C}^{n_1 \times r}$ spans the same subspace and $\boldsymbol{U}$ is orthogonal, there exists an orthonormal matrix $\boldsymbol{Q} \in \mathbb{C}^{r \times r}$ such that $\boldsymbol{U} = \boldsymbol{E}_L(\boldsymbol{E}_L^*\boldsymbol{E}_L)^{-1/2}\boldsymbol{Q}$. So

$$\left\|\boldsymbol{U}^{(i,:)}\right\|^2 = \left\|\boldsymbol{e}_i^*\boldsymbol{E}_L(\boldsymbol{E}_L^*\boldsymbol{E}_L)^{-1/2}\right\|^2 \leq \|\boldsymbol{e}_i^*\boldsymbol{E}_L\|^2 \left\|(\boldsymbol{E}_L^*\boldsymbol{E}_L)^{-1}\right\| \leq \frac{\mu_0 r}{n_1} \leq \frac{\mu_0 c_s r}{n}.$$

$$\begin{aligned}\|\mathcal{P}_{\boldsymbol{U}}(\boldsymbol{H}_a)\|_F^2 &= \|\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{H}_a\|_F^2 \\ &= \left\|\boldsymbol{E}_L(\boldsymbol{E}_L^*\boldsymbol{E}_L)^{-1}\boldsymbol{E}_L^*\boldsymbol{H}_a\right\|_F^2 \leq \frac{\|\boldsymbol{E}_L^*\boldsymbol{H}_a\|_F^2}{\sigma_{\min}(\boldsymbol{E}_L^*\boldsymbol{E}_L)} \leq \frac{\mu_0 r}{n_1} \leq \frac{\mu_0 c_s r}{n},\end{aligned}$$

where we have used the fact that $\boldsymbol{H}_a$ only has w_a nonzero entries of magnitude $1/\sqrt{w_a}$ in its a-th anti-diagonal and the magnitudes of the entries of $\boldsymbol{E}_L$ is bounded above by one for both the damped and undamped case. □

1.3.1 Extension to Higher Dimension

For multi-dimensional spectrally sparse signals, we can formulate the reconstruction problem as a multi-level Hankel matrix completion problem. Without loss of generality, we discuss the two-dimensional setting but emphasize that the situation in general d-dimensions is similar.

Let $w_k = e^{(2\pi\imath f_{1k} - \tau_{1k})}$ and $z_k = e^{(2\pi\imath f_{2k} - \tau_{2k})}$ for r frequency pairs $(f_{1k}, f_{2k}) \in [0,1)^2$ and r damping factor pairs $(\tau_{1k}, \tau_{2k}) \in \mathbb{R}_+^2$. A two-dimensional spectrally sparse array $\boldsymbol{X} \in \mathbb{C}^{N_1 \times N_2}$ can be expressed as

$$\boldsymbol{X}^{(a,b)} = \sum_{k=1}^{r} d_k w_k^a z_k^b, \quad (a,b) \in [N_1] \times [N_2].$$

The two-level Hankel matrix of $\boldsymbol{X}$ is given by

$$\mathcal{H}\boldsymbol{X} = \begin{bmatrix} \mathcal{H}\boldsymbol{X}^{(:,0)} & \mathcal{H}\boldsymbol{X}^{(:,1)} & \mathcal{H}\boldsymbol{X}^{(:,2)} & \dots & \dots & \mathcal{H}\boldsymbol{X}^{(:,N_2-n_2)} \\ \mathcal{H}\boldsymbol{X}^{(:,1)} & \mathcal{H}\boldsymbol{X}^{(:,2)} & \dots & \dots & \dots & \mathcal{H}\boldsymbol{X}^{(:,N_2-n_2+1)} \\ \mathcal{H}\boldsymbol{X}^{(:,2)} & \dots & \dots & \dots & \dots & \mathcal{H}\boldsymbol{X}^{(:,N_2-n_2+2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathcal{H}\boldsymbol{X}^{(:,n_2-1)} & \mathcal{H}\boldsymbol{X}^{(:,n_2)} & \dots & \dots & \dots & \mathcal{H}\boldsymbol{X}^{(:,N_2-1)} \end{bmatrix},$$

where each block is an $n_1 \times (N_1 - n_1 + 1)$ Hankel matrix corresponding to a column

of $\boldsymbol{X}$,

$$\mathcal{H}\boldsymbol{X}^{(:,b)} = \begin{bmatrix} \mathcal{H}\boldsymbol{X}^{(0,b)} & \mathcal{H}\boldsymbol{X}^{(1,b)} & \mathcal{H}\boldsymbol{X}^{(2,b)} & \cdots & \cdots & \mathcal{H}\boldsymbol{X}^{(N_1-n_1,b)} \\ \mathcal{H}\boldsymbol{X}^{(1,b)} & \mathcal{H}\boldsymbol{X}^{(2,b)} & \cdots & \cdots & \cdots & \mathcal{H}\boldsymbol{X}^{(N_1-n_1+1,b)} \\ \mathcal{H}\boldsymbol{X}^{(2,b)} & \cdots & \cdots & \cdots & \cdots & \mathcal{H}\boldsymbol{X}^{(N_1-n_1+2,b)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathcal{H}\boldsymbol{X}^{(n_1-1,b)} & \mathcal{H}\boldsymbol{X}^{(n_1,b)} & \cdots & \cdots & \cdots & \mathcal{H}\boldsymbol{X}^{(N_1-1,b)} \end{bmatrix}.$$

Clearly, $\mathcal{H}\boldsymbol{X}$ is an $(n_1 n_2) \times (N_1 - n_1 + 1)(N_2 - n_2 + 1)$ matrix. Letting $i = i_1 + i_2 \cdot n_1$ and $j = j_1 + j_2 \cdot (N_1 - n_1 + 1)$, the (i,j)-th entry of $\mathcal{H}\boldsymbol{X}$ is given by

$$\mathcal{H}\boldsymbol{X}^{(i,j)} = \boldsymbol{X}^{(i_1+j_1, i_2+j_2)} = \sum_{k=1}^{r} d_k \left(w_k^{i_1} z_k^{i_2}\right) \left(w_k^{j_1} z_k^{j_2}\right). \tag{1.8}$$

For $k = 1, \cdots, r$, we define the four vectors $\boldsymbol{w}_k^{[n_1]}$, $\boldsymbol{w}_k^{[N_1-n_1+1]}$, $\boldsymbol{z}_k^{[n_2]}$, and $\boldsymbol{z}_k^{[N_2-n_2+1]}$ as

$$\boldsymbol{w}_k^{[n_1]} = \begin{bmatrix} 1 \\ w_k \\ \vdots \\ w_k^{n_1-1} \end{bmatrix}, \boldsymbol{w}_k^{[N_1-n_1+1]} = \begin{bmatrix} 1 \\ w_k \\ \vdots \\ w_k^{N_1-n_1} \end{bmatrix}, \boldsymbol{z}_k^{[n_2]} = \begin{bmatrix} 1 \\ z_k \\ \vdots \\ z_k^{n_2-1} \end{bmatrix}, \boldsymbol{z}_k^{[N_2-n_2+1]} = \begin{bmatrix} 1 \\ z_k \\ \vdots \\ z_k^{N_2-n_2} \end{bmatrix}.$$

Let $\boldsymbol{E}_L$ be an $(n_1 n_2) \times r$ matrix with the k-th column being given by $\boldsymbol{z}_k^{[n_2]} \otimes \boldsymbol{w}_k^{[n_1]}$, and let $\boldsymbol{E}_R$ be an $(N_1 - n_1 + 1)(N_2 - n_2 + 1) \times r$ matrix with the k-th column being given by $\boldsymbol{z}_k^{[N_2-n_2+1]} \otimes \boldsymbol{w}_k^{[N_1-n_1+1]}$. Then it follows from (1.8) that $\mathcal{H}\boldsymbol{X}$ admits the

Vandermonde decomposition

$$\mathcal{H}\boldsymbol{X} = \boldsymbol{E}_L \boldsymbol{D} \boldsymbol{E}_R^T,$$

where $\boldsymbol{D} = \mathrm{diag}(d_1, \cdots, d_r)$. Thus, it is self-evident that $\mathcal{H}\boldsymbol{X}$ is a rank r matrix.

As in the one-dimensional case, the goal in two-dimensional spectral sparse signal reconstruction is to reconstruct $\boldsymbol{X}$ from the partial revealed entries of $\boldsymbol{X}$, denoted $\mathcal{P}_\Omega(\boldsymbol{X})$, where Ω is a subset of $[N_1] \times [N_2]$. Let $\mathcal{H}\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the SVD of $\mathcal{H}\boldsymbol{X}$. We say $\mathcal{H}\boldsymbol{X}$ is μ_0-incoherent if there exists a numerical constant $\mu_0 > 0$ such that

$$\sigma_{\min}(\boldsymbol{E}_L^* \boldsymbol{E}_L) \geq \frac{n_1 n_2}{\mu_0}$$

and

$$\sigma_{\min}(\boldsymbol{E}_R^* \boldsymbol{E}_R) \geq \frac{(N_1 - n_1 + 1)(N_2 - n_2 + 1)}{\mu_0}.$$

Based on [36, Thm. 1], one can show that $= \mathcal{H}\boldsymbol{X}$ is μ_0-incoherent if there is no damping in $\boldsymbol{X}$ and the minimum wrap-around distance between the underlying frequencies $\{f_{ik}\}_{k=1}^r$ is greater than about $2/N_i$ for $i = 1, 2$.

We have shown that the reconstruction of multi-dimensional spectrally sparse signals can be formulated as a low-rank multi-level Hankel matrix completion problem. We will focus on the one-dimensional case in the rest of the thesis, and present two provable non-convex methods [8, 9] in Section 2 and 3 respectively to solve the corresponding low-rank Hankel matrix completion problem. Our readers are assured

that the implementation and the proofs of the two methods can be extended straightforwardly to the multi-dimensional cases, as shown in the supplementary materials of [8] and [9].

CHAPTER 2
FAST ITERATIVE HARD THRESHOLDING

2.1 Algorithms and Main Results

2.1.1 Algorithms

We present our first reconstruction algorithm in Alg. 2.1, which is an iterative hard thresholding algorithm for the following minimization problem,

$$\min_{\boldsymbol{z}} \langle \boldsymbol{z}-\boldsymbol{x}, \mathcal{P}_\Omega(\boldsymbol{z}-\boldsymbol{x})\rangle \quad \text{subject to} \quad \operatorname{rank}(\mathcal{H}\boldsymbol{z}) = r. \tag{2.1}$$

In each iteration of IHT, the current estimate $\boldsymbol{x}_l$ is first updated along the gradient descent direction under the Wirtinger calculus with the stepsize $p^{-1} = \frac{n}{m}$. Then the Hankel matrix corresponding to the update is formed via the application of the linear operator $\mathcal{H}$, followed by an SVD truncation to its nearest rank r approximation. The hard thresholding operator $\mathcal{T}_r(\cdot)$ in Step 3 of Alg. 2.1 is defined as

$$\mathcal{T}_r(\boldsymbol{Z}) = \sum_{k=1}^{r} \sigma_r \boldsymbol{u}_k \boldsymbol{v}_k^*,$$

where $\boldsymbol{Z} = \sum_{k=1}^{\min(n_1,n_2)} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^*$ is an SVD with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(n_1,n_2)}$. Finally the new estimate $\boldsymbol{x}_{l+1}$ is obtained via the application of $\mathcal{H}^\dagger$ on the low rank matrix $\boldsymbol{L}_{l+1}$.

IHT can achieve linear convergence rate as demonstrated in Sec. 2.2.2. However, it requires to compute the truncated SVD of an $n_1 \times n_2$ matrix in each iteration.

Algorithm 2.1 Iterative Hard Thresholding (IHT)

Initialize $\boldsymbol{L}_0$ and **Set** $\boldsymbol{x}_0 = \mathcal{H}^\dagger \boldsymbol{L}_0$
for $l = 0, 1, \cdots$ **do**
1. $\boldsymbol{g}_l = \mathcal{P}_\Omega(\boldsymbol{x} - \boldsymbol{x}_l)$
2. $\boldsymbol{W}_l = \mathcal{H}(\boldsymbol{x}_l + p^{-1}\boldsymbol{g}_l)$
3. $\boldsymbol{L}_{l+1} = \mathcal{T}_r(\boldsymbol{W}_l)$
4. $\boldsymbol{x}_{l+1} = \mathcal{H}^\dagger \boldsymbol{L}_{l+1}$

end for

Algorithm 2.2 Fast Iterative Hard Thresholding (FIHT)

Initialize $\boldsymbol{L}_0$ and **Set** $\boldsymbol{x}_0 = \mathcal{H}^\dagger \boldsymbol{L}_0$
for $l = 0, 1, \cdots$ **do**
1. $\boldsymbol{g}_l = \mathcal{P}_\Omega(\boldsymbol{x} - \boldsymbol{x}_l)$
2. $\boldsymbol{W}_l = \mathcal{P}_{\mathcal{S}_l}\mathcal{H}(\boldsymbol{x}_l + p^{-1}\boldsymbol{g}_l)$
3. $\boldsymbol{L}_{l+1} = \mathcal{T}_r(\boldsymbol{W}_l)$
4. $\boldsymbol{x}_{l+1} = \mathcal{H}^\dagger \boldsymbol{L}_{l+1}$

end for

Though there are fast SVD solvers [34, 57], it is still computationally expensive when n $(= n_1 + n_2 - 1)$ is large. To improve the computational efficiency we propose to project the Hankel matrix $\mathcal{H}(\boldsymbol{x}_l + p^{-1}\boldsymbol{g}_l)$ onto a low dimensional subspace $\mathcal{S}_l$ before truncating it to the best rank r approximation. The fast iterative hard thresholding algorithm equipped with an extra subspace projection step is presented in Alg. 2.2, where $\mathcal{P}_{\mathcal{S}_l}(\cdot)$ denotes the projection of $n_1 \times n_2$ matrices onto the subspace $\mathcal{S}_l$. Inspired by the Riemannian optimization algorithms for low rank matrix completion [54–56], $\mathcal{S}_l$ is selected to be the direct sum of the column and row spaces of $\boldsymbol{L}_l$,

$$\mathcal{S}_l = \{\boldsymbol{U}_l\boldsymbol{B} + \boldsymbol{C}\boldsymbol{V}_l^* \mid \boldsymbol{B} \in \mathbb{C}^{r\times n_2},\ \boldsymbol{C} \in \mathbb{C}^{n_1\times r}\}, \tag{2.2}$$

where $\boldsymbol{U}_l \in \mathbb{C}^{n_1 \times r}$ and $\boldsymbol{V}_l \in \mathbb{C}^{n_2 \times r}$ are the left and right singular vectors of $\boldsymbol{L}_l$. The subspace $\mathcal{S}_l$ defined in (2.2) can be geometrically interpreted as the tangent space of the embedded rank r matrix manifold at $\boldsymbol{L}_l$ [54]. For any matrix $\boldsymbol{Z} \in \mathbb{C}^{n_1 \times n_2}$, the projection of $\boldsymbol{Z}$ onto $\mathcal{S}_l$ is given by

$$\mathcal{P}_{\mathcal{S}_l}(\boldsymbol{Z}) = \boldsymbol{U}_l\boldsymbol{U}_l^*\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{V}_l\boldsymbol{V}_l^* - \boldsymbol{U}_l\boldsymbol{U}_l^*\boldsymbol{Z}\boldsymbol{V}_l\boldsymbol{V}_l^*.$$

Iterative hard thresholding is a family of simple yet efficient algorithms for compressed sensing [2,4,5,24] and matrix completion [28,31,50], where in compressed sensing signals of interest are sparse and in matrix completion signals of interest are low rank. However, the signal of interest is neither sparse nor low rank itself, but instead the Hankel matrix corresponding to the signal is low rank. Therefore Algs. 2.1 and 2.2 alternate between the vector space and the matrix space and this alternating structure does not exist in typical iterative hard thresholding algorithms for compressed sensing and matrix completion.

2.1.2 Computational Complexity

We focus on the implementation details of FIHT and show that the SVD of $\boldsymbol{W}_l$ in the third step of Alg. 2.2 can be computed using $O(r^3)$ floating point operations (flops) owing to the low rank structure of the matrices in $\mathcal{S}_l$.

Assume the rank r matrix $\boldsymbol{L}_l$ is stored by its SVD $\boldsymbol{L}_l = \boldsymbol{U}_l\boldsymbol{\Sigma}_l\boldsymbol{V}_l^*$ in each

iteration. Then,

$$\boldsymbol{x}_l = \mathcal{H}^\dagger \boldsymbol{L}_l = \mathcal{D}^{-2}\mathcal{H}^* \boldsymbol{L}_l = \mathcal{D}^{-2} \sum_{k=1}^{r} \boldsymbol{\Sigma}_l^{(k,k)} \mathcal{H}^* \left(\boldsymbol{U}_l^{(:,k)} \left(\boldsymbol{V}_l^{(:,k)} \right)^* \right),$$

where $\mathcal{H}^* \left(\boldsymbol{U}_l^{(:,k)} \left(\boldsymbol{V}_l^{(:,k)} \right)^* \right)$ can be computed via fast convolution by noting that

$$\left[\mathcal{H}^* \left(\boldsymbol{U}_l^{(:,k)} \left(\boldsymbol{V}_l^{(:,k)} \right)^* \right) \right]_a = \sum_{i+j=a} \boldsymbol{U}_l^{(i,k)} \overline{\boldsymbol{V}}_l^{(j,k)}, \quad a = 0, \cdots, n-1.$$

Therefore computing the last step of Alg. 2.2 costs $O(rn\log(n))$ flops.

We distinguish two cases regarding to the computations of $\boldsymbol{W}_l$ and its SVD.

Case 1: $n_1 \neq n_2$. Let $\boldsymbol{H}_l = \mathcal{H}\left(\boldsymbol{x}_l + p^{-1}\boldsymbol{g}_l\right)$. The intermediate matrix $\boldsymbol{W}_l$ is stored by the following decomposition

$$\begin{aligned}
\boldsymbol{W}_l = \mathcal{P}_{\mathcal{S}_l}\boldsymbol{H}_l &= \boldsymbol{U}_l\boldsymbol{U}_l^*\boldsymbol{H}_l + \boldsymbol{H}_l\boldsymbol{V}_l\boldsymbol{V}_l^* - \boldsymbol{U}_l\boldsymbol{U}_l^*\boldsymbol{H}_l\boldsymbol{V}_l\boldsymbol{V}_l^* \\
&= \boldsymbol{U}_l \underbrace{\boldsymbol{U}_l^*\boldsymbol{H}_l\boldsymbol{V}_l}_{\boldsymbol{C}\in\mathbb{C}^{r\times r}} \boldsymbol{V}_l^* + \boldsymbol{U}_l \underbrace{\boldsymbol{U}_l^*\boldsymbol{H}_l(\boldsymbol{I} - \boldsymbol{V}_l\boldsymbol{V}_l^*)}_{\boldsymbol{X}^*\in\mathbb{C}^{r\times n_2}} + \underbrace{(\boldsymbol{I} - \boldsymbol{U}_l\boldsymbol{U}_l^*)\boldsymbol{H}_l\boldsymbol{V}_l}_{\boldsymbol{Y}\in\mathbb{C}^{n_1\times r}} \boldsymbol{V}_l^* \\
&= \boldsymbol{U}_l\boldsymbol{C}\boldsymbol{V}_l^* + \boldsymbol{U}_l\boldsymbol{X}^* + \boldsymbol{Y}\boldsymbol{V}_l^*.
\end{aligned}$$

Note that $\boldsymbol{H}_l^*\boldsymbol{U}_l$ and $\boldsymbol{H}_l\boldsymbol{V}_l$ in $\boldsymbol{C}$, $\boldsymbol{X}$ and $\boldsymbol{M}$ can be computed using r fast Hankel matrix-vector multiplications without forming $\boldsymbol{H}_l$, which requires $O(rn\log(n))$ flops. Therefore the total computational cost for computing $\boldsymbol{C}$, $\boldsymbol{X}$ and $\boldsymbol{M}$ is $O(r^2n + rn\log(n))$ flops.

Let $\boldsymbol{X} = \boldsymbol{Q}_1\boldsymbol{R}_1$ and $\boldsymbol{Y} = \boldsymbol{Q}_2\boldsymbol{R}_2$ respectively be the QR factorizations of $\boldsymbol{X}$

and $\boldsymbol{M}$. Then $\boldsymbol{Q}_1 \perp \boldsymbol{V}_l$, $\boldsymbol{Q}_2 \perp \boldsymbol{U}_l$ and $\boldsymbol{W}_l$ can be rewritten as

$$\boldsymbol{W}_l = \boldsymbol{U}_l \boldsymbol{C} \boldsymbol{V}_l^* + \boldsymbol{U}_l \boldsymbol{R}_1^* \boldsymbol{Q}_1^* + \boldsymbol{Q}_2 \boldsymbol{R}_2 \boldsymbol{V}_l^* = \begin{bmatrix} \boldsymbol{U}_l & \boldsymbol{Q}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{C} & \boldsymbol{Q}_1^* \\ \boldsymbol{Q}_2 & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_l & \boldsymbol{Q}_1 \end{bmatrix}^*.$$

Suppose the SVD of the middle $2r \times 2r$ matrix is given by

$$\begin{bmatrix} \boldsymbol{C} & \boldsymbol{Q}_1^* \\ \boldsymbol{Q}_2 & \boldsymbol{0} \end{bmatrix} = \boldsymbol{U}_c \boldsymbol{\Sigma}_c \boldsymbol{V}_c^*.$$

Then SVD of $\boldsymbol{W}_l$ can be computed as

$$\boldsymbol{W}_l = \left(\begin{bmatrix} \boldsymbol{U}_l & \boldsymbol{Q}_2 \end{bmatrix} \boldsymbol{U}_c \right) \boldsymbol{\Sigma}_c \left(\begin{bmatrix} \boldsymbol{V}_l & \boldsymbol{Q}_1 \end{bmatrix} \boldsymbol{V}_c \right)^*.$$

Thus computing the SVD of $\boldsymbol{W}_l$ requires $O(r^2 n + r^3)$ flops.

Case 2: $n_1 = n_2$. In this case, $\boldsymbol{H}_l$ is a square and symmetric matrix (but not Hermitian). Assume $\boldsymbol{L}_l$ is also symmetric which can be achieved when $l = 0$. Then $\boldsymbol{L}_l$ admits a Takagi factorization $\boldsymbol{L}_l = \boldsymbol{U}_l \boldsymbol{\Sigma}_l \boldsymbol{U}_l^T$, which is also the SVD of $\boldsymbol{L}_l$ [57]. So

$$\begin{aligned} \boldsymbol{W}_l = \mathcal{P}_{\mathcal{S}_l}(\boldsymbol{H}_l) &= \boldsymbol{U}_l \boldsymbol{U}_l^* \boldsymbol{H}_l + \boldsymbol{H}_l \overline{\boldsymbol{U}_l} \boldsymbol{U}_l^T - \boldsymbol{U}_l \boldsymbol{U}_l^* \boldsymbol{H}_l \overline{\boldsymbol{U}_l} \boldsymbol{U}_l^T \\ &= \boldsymbol{U}_l \underbrace{\boldsymbol{U}_l^* \boldsymbol{H}_l \overline{\boldsymbol{U}_l}}_{\boldsymbol{C} \in \mathbb{C}^{r \times r}} \boldsymbol{U}_l^T + \boldsymbol{U}_l \underbrace{\boldsymbol{U}_l^* \boldsymbol{H}_l (\boldsymbol{I} - \overline{\boldsymbol{U}_l} \boldsymbol{U}_l^T)}_{\boldsymbol{X}^T \in \mathbb{C}^{r \times n_1}} + \underbrace{(\boldsymbol{I} - \boldsymbol{U}_l \boldsymbol{U}_l^*) \boldsymbol{H}_l \overline{\boldsymbol{U}_l}}_{\boldsymbol{X} \in \mathbb{C}^{n_1 \times r}} \boldsymbol{U}_l^T \\ &= \boldsymbol{U}_l \boldsymbol{C} \boldsymbol{U}_l^T + \boldsymbol{U}_l \boldsymbol{X}^T + \boldsymbol{X} \boldsymbol{U}_l^T \end{aligned}$$

is also a symmetric matrix and nearly half of the computational costs will be saved

compared with the non-square case.

Let $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ be the QR factorization of $\boldsymbol{X}$. Then $\boldsymbol{Q} \perp \boldsymbol{U}$ and

$$\boldsymbol{W}_l = \boldsymbol{U}_l \boldsymbol{C} \boldsymbol{U}_l^T + \boldsymbol{U}_l \boldsymbol{R}^T \boldsymbol{Q}^T + \boldsymbol{Q}\boldsymbol{R}\boldsymbol{U}_l^T = \begin{bmatrix} \boldsymbol{U}_l & \boldsymbol{Q} \end{bmatrix} \begin{bmatrix} \boldsymbol{C} & \boldsymbol{Q}^T \\ \boldsymbol{Q} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_l & \boldsymbol{Q} \end{bmatrix}^T.$$

This, together with the Takagi factorization (also the SVD) of the middle $2r \times 2r$ matrix

$$\begin{bmatrix} \boldsymbol{C} & \boldsymbol{R}^T \\ \boldsymbol{R} & \boldsymbol{0} \end{bmatrix} = \boldsymbol{U}_c \boldsymbol{\Sigma}_c \boldsymbol{U}_c^T,$$

gives the Takagi factorization (also the SVD) of $\boldsymbol{W}_l$

$$\boldsymbol{W}_l = \left(\begin{bmatrix} \boldsymbol{U}_l & \boldsymbol{Q} \end{bmatrix} \boldsymbol{U}_c \right) \boldsymbol{\Sigma}_c \left(\begin{bmatrix} \boldsymbol{U}_l & \boldsymbol{Q} \end{bmatrix} \boldsymbol{U}_c \right)^T.$$

Moreover, $\boldsymbol{L}_{l+1}$ remains symmetric and admits a Takagi factorization as the best rank r approximation of $\boldsymbol{W}_l$.

In summary, the leading order per iteration computational cost of FIHT is $O(r^2 n + rn\log(n) + r^3)$ flops, which can be further reduced by exploring the symmetric structure of matrices when $n_1 = n_2$. In addition, the largest matrices that need to be stored are the singular vector matrices of $\boldsymbol{W}_l$. Therefore, FIHT requires only $O(rn)$ memory.

2.1.3 Main Results

In this section, we present theoretical recovery guarantees for FIHT (Alg. 2.2). The guarantee analysis relies on restricted isometry properties of $\mathcal{P}_\Omega$ which cannot be established for IHT (Alg. 2.1). Moreover, numerical simulations in Sec. 2.2 suggest that while FIHT and IHT both have linear convergence rate, FIHT can be sufficiently faster due to the low per iteration computational cost.

Let $\Omega = \{a_k \mid k = 1, \ldots, m\}$. We consider the *sampling with replacement model* for Ω; that is each index a_k is drawn independently and uniformly from $\{0, \cdots, n-1\}$. Recall that we use $\mathcal{P}_\Omega(\cdot)$ to represent the projection of vectors onto a subset of the canonical basis of $\mathbb{C}^n$, i.e.,

$$\mathcal{P}_\Omega(\boldsymbol{z}) = \sum_{k=1}^{m} \langle \boldsymbol{z}, \boldsymbol{e}_{a_k} \rangle \, \boldsymbol{e}_{a_k}, \quad \forall \boldsymbol{z} \in \mathbb{C}^n$$

as well as the projection of matrices onto a subset of an orthonormal basis of Hankel matrices, i.e.,

$$\mathcal{P}_\Omega(\boldsymbol{Z}) = \sum_{k=1}^{m} \langle \boldsymbol{Z}, \boldsymbol{H}_{a_k} \rangle \, \boldsymbol{H}_{a_k}, \quad \forall \boldsymbol{Z} \in \mathbb{C}^{n_1 \times n_2}$$

since they are corresponding to each other and the context will make their distinction clear.

As is typical in non-convex optimization, the theoretical recovery guarantees of FIHT are closely related to the initial guess. We will discuss two initialization strategies and the corresponding recovery guarantees for FIHT. The proofs of the

lemmas and theorems in Secs. 2.1.3.1 and 2.1.3.2 will be provided in Sec. 2.3.

2.1.3.1 Initialization via One Step Hard Thresholding

Our first initial guess is $\boldsymbol{L}_0 = p^{-1}\mathcal{T}_r(\mathcal{H}\mathcal{P}_\Omega(\boldsymbol{x}))$, which is obtained by truncating the Hankel matrix constructed from the observed entries of $\boldsymbol{x}$. The following lemma which is of independent interest bounds the deviation of $\boldsymbol{L}_0$ from $\mathcal{H}\boldsymbol{x}$.

Lemma 2.1.1. *Assume $\mathcal{H}\boldsymbol{x}$ is μ_0-incoherent. Then there exists a universal constant $C > 0$ such that*

$$\|\boldsymbol{L}_0 - \mathcal{H}\boldsymbol{x}\| \leq C\sqrt{\frac{\mu_0 c_s r \log(n)}{m}}\|\mathcal{H}\boldsymbol{x}\|$$

with probability at least $1 - n^{-2}$.

It follows from Lem. 2.1.1 that, if m is sufficiently large and in the order of $r\log(n)$, the spectral norm distance between $\boldsymbol{L}_0$ and $\mathcal{H}\boldsymbol{x}$ can be less than any arbitrarily small constant. The following theoretical recovery guarantee can be established for FIHT based on this lemma.

Theorem 2.1.1 (Guarantee I)**.** *Assume $\mathcal{H}\boldsymbol{x}$ is μ_0-incoherent. Let $0 < \varepsilon_0 < \frac{1}{10}$ be a numerical constant and $\nu = 10\varepsilon_0 < 1$. Then with probability at least $1 - 3n^{-2}$, the iterates generated by FIHT (Alg. 2.2) with the initial guess $\boldsymbol{L}_0 = p^{-1}\mathcal{T}_r(\mathcal{H}\mathcal{P}_\Omega(\boldsymbol{x}))$ satisfy*

$$\|\boldsymbol{x}_l - \boldsymbol{x}\| \leq \nu^l \|\boldsymbol{L}_0 - \mathcal{H}\boldsymbol{x}\|_F,$$

provided

$$m \geq C \max\left\{\varepsilon_0^{-2}\mu_0 c_s, (1+\varepsilon_0)\varepsilon_0^{-1}\mu_0^{1/2}c_s^{1/2}\right\}\kappa r n^{1/2}\log^{3/2}(n)$$

for some universal constant $C > 0$, where $\kappa = \frac{\sigma_{\max}(\mathcal{H}\boldsymbol{x})}{\sigma_{\min}(\mathcal{H}\boldsymbol{x})}$ denotes the condition number of $\mathcal{H}\boldsymbol{x}$.

Remark. Since $\mathcal{H}\boldsymbol{x} = \boldsymbol{E}_L \boldsymbol{D} \boldsymbol{E}_R^T$, we have

$$\kappa \leq \frac{\sigma_{\max}(\boldsymbol{E}_L)}{\sigma_{\min}(\boldsymbol{E}_L)} \cdot \frac{\max_k |d_k|}{\min_k |d_k|} \cdot \frac{\sigma_{\max}(\boldsymbol{E}_R)}{\sigma_{\min}(\boldsymbol{E}_R)}.$$

It follows from [37, Thm. 2] that $\sigma_{\max}(\boldsymbol{E}_L)$ (resp. $\sigma_{\max}(\boldsymbol{E}_R)$) and $\sigma_{\min}(\boldsymbol{E}_L)$ (resp. $\sigma_{\min}(\boldsymbol{E}_R)$) are both proportional to $\sqrt{n_1}$ (resp. $\sqrt{n_2}$) when the frequencies of $\boldsymbol{x}$ are well separated. Thus the condition number of $\mathcal{H}\boldsymbol{x}$ is essentially proportional to the dynamical range $\max_k |d_k| / \min_k |d_k|$.

Since the number of measurements required in Thm. 2.1.1 is proportional to $c_s = \max\left\{\frac{n}{n_1}, \frac{n}{n_2}\right\}$ and $n_1 + n_2 - 1 = n$, it makes sense to construct a nearly square Hankel matrix to recover spectrally sparse signals via low rank Hankel matrix completion.

2.1.3.2 Initialization via Resampling and Trimming

The sampling complexity in Thm. 2.1.1 depends on $\sqrt{n}$ which is not desirable since the degrees of freedom in a spectrally sparse signal is only proportional to r. To eliminate the dependence on $\sqrt{n}$, we investigate another initialization procedure which is described in Alg. 2.3.

Alg. 2.3 begins with partitioning the sampling set Ω into $L+1$ disjoint subsets. In each iteration, the new estimate is obtained via an application of FIHT on the new sampling set followed by the trimming procedure. The use of a fresh sampling set in

Algorithm 2.3 Initialization via Resampled FIHT and Trimming

Partition Ω into $L+1$ disjoint sets $\Omega_0, \cdots, \Omega_L$ of equal size $\widehat{m}$, let $\widehat{p} = \frac{\widehat{m}}{n}$.
Set $\widetilde{\boldsymbol{L}}_0 = \mathcal{T}_r\left(\widehat{p}^{-1}\mathcal{H}\mathcal{P}_{\Omega_0}\left(\boldsymbol{x}\right)\right)$,
for $l = 0, \cdots, L-1$ **do**
1. $\widehat{\boldsymbol{L}}_l = \mathrm{Trim}_{\mu_0}(\widetilde{\boldsymbol{L}}_l)$
2. $\widehat{\boldsymbol{x}}_l = \mathcal{H}^{\dagger}\widehat{\boldsymbol{L}}_l$
3. $\widetilde{\boldsymbol{L}}_{l+1} = \mathcal{T}_r\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{H}\left(\widehat{\boldsymbol{x}}_l + \widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}}\left(\boldsymbol{x} - \widehat{\boldsymbol{x}}_l\right)\right)$

end for

Algorithm 2.4 Trim_{μ}

Input: $\widetilde{\boldsymbol{L}}_{l+1} = \widetilde{\boldsymbol{U}}_{l+1}\widetilde{\boldsymbol{\Sigma}}_{l+1}\widetilde{\boldsymbol{V}}_{l+1}^*$
Output: $\widehat{\boldsymbol{L}}_{l+1} = \widehat{\boldsymbol{A}}_{l+1}\widetilde{\boldsymbol{\Sigma}}_{l+1}\widehat{\boldsymbol{B}}_{l+1}^*$, where

$$\widehat{\boldsymbol{A}}_{l+1}^{(i,:)} = \frac{\widetilde{\boldsymbol{U}}_{l+1}^{(i,:)}}{\left\|\widetilde{\boldsymbol{U}}_{l+1}^{(i,:)}\right\|}\min\left\{\left\|\widetilde{\boldsymbol{U}}_{l+1}^{(i,:)}\right\|, \sqrt{\frac{\mu c_s r}{n}}\right\},$$

$$\widehat{\boldsymbol{B}}_{l+1}^{(i,:)} = \frac{\widetilde{\boldsymbol{V}}_{l+1}^{(i,:)}}{\left\|\widetilde{\boldsymbol{V}}_{l+1}^{(i,:)}\right\|}\min\left\{\left\|\widetilde{\boldsymbol{V}}_{l+1}^{(i,:)}\right\|, \sqrt{\frac{\mu c_s r}{n}}\right\}.$$

each iteration breaks the dependence between the last estimate and the sampling set, while the trimming procedure ensures that the estimate remains an incoherent matrix after each iteration. The following lemma provides an estimation of the approximation accuracy of the initial guess returned by Alg. 2.3.

Lemma 2.1.2. *Assume $\mathcal{H}\boldsymbol{x}$ is μ_0-incoherent. Then with probability at least $1-(2L+1)n^{-2}$, the output of Alg. 2.3 satisfies*

$$\|\widetilde{\boldsymbol{L}}_L - \mathcal{H}\boldsymbol{x}\|_F \leq \left(\frac{5}{6}\right)^L \frac{\sigma_{\min}(\mathcal{H}\boldsymbol{x})}{256\kappa^2}$$

provided $\widehat{m} \geq C\mu_0 c_s \kappa^6 r^2 \log(n)$ *for some universal constant* $C > 0$.

We can obtain the following recovery guarantee for FIHT with $\boldsymbol{L}_0$ being the output of Alg. 2.3.

Theorem 2.1.2 (Guarantee II)**.** *Assume* $\mathcal{H}\boldsymbol{x}$ *is* μ_0*-incoherent. Let* $0 < \varepsilon_0 < \frac{1}{10}$ *and* $L = \left\lceil 6 \log\left(\frac{\sqrt{n}\log(n)}{16\varepsilon_0}\right)\right\rceil$. *Define* $\nu = 10\varepsilon_0 < 1$. *Then with probability at least* $1 - (2L+3)\, n^{-2}$*, the iterates generated by FIHT (Alg. 2.2) with* $\boldsymbol{L}_0 = \widetilde{\boldsymbol{L}}_L$ *(the output of Alg. 2.3) satisfies*

$$\|\boldsymbol{x}_l - \boldsymbol{x}\| \leq \nu^l \|\boldsymbol{L}_0 - \mathcal{H}\boldsymbol{x}\|_F,$$

provided

$$m \geq C\mu_0 c_s \kappa^6 r^2 \log(n) \log\left(\frac{\sqrt{n}\log(n)}{16\varepsilon_0}\right)$$

for some universal constant $C > 0$.

2.2 Numerical Experiments

In this section, we conduct numerical experiments to evaluate the performance of IHT and FIHT. The experiments are executed from Matlab 2014a on a MacBook Pro with a 2.7GHz dual-core Intel i5 CPU and 8 GB memory, and the algorithms are evaluated against successful recovery rates, computational efficiency, robustness to additive noise, sensitivity to mis-specification of model order and capability of handling high-dimensional data. We initialize IHT and FIHT using one step hard thresholding computed via the PROPACK package [34] rather than the resampled FIHT (Alg. 2.3), as the former one has already shown very good performance and

preliminary numerical results didn't present dramatic difference between those two initialization procedures for our simulations.

2.2.1 Empirical Phase Transition

We investigate the recovery rates of IHT and FIHT in the framework of phase transition and compare them with EMaC [15] and ANM [49]. ANM and EMaC are implemented using CVX [29] with default parameters. The spectrally sparse signals of length n with r frequency components are formed in the following way: each frequency f_k is uniformly sampled from $[0, 1)$, and the argument of each complex coefficient d_k is uniformly sampled from $[0, 2\pi)$ while the amplitude is selected to be $1 + 10^{0.5c_k}$ with c_k being uniformly distributed on $[0, 1]$. Then m entries of the test signals are sampled uniformly at random. For a given triple (n, r, m), 50 random tests are conducted. We consider an algorithm to have successfully reconstructed a test signal if $\|\boldsymbol{x}_{rec} - \boldsymbol{x}\|_2/\|\boldsymbol{x}\| \leq 10^{-3}$. The tests are conducted with $n = 127$ and $p = m/n$ taking 18 equispaced values from 0.1 to 0.95. For a fixed pair of (n, m), we start with $r = 1$ and then increase it by one until it reaches a value such that the tested algorithm fails all the 50 random tests.

The empirical phase transitions for the four tested algorithms ANM, EMaC, IHT and FIHT are presented in Fig. 2.1, where white color indicates that the algorithm can recover all of the 50 random test signals and on the other hand black color indicates the algorithm fails to recover each of the randomly generated signals. The top four plots of the figure present the recovery phase transitions where no separation

of the frequencies is imposed, while the bottom four plots presents the recovery phase transitions where the wrap-around distances between the randomly drawn frequencies are greater than $1.5/n$.

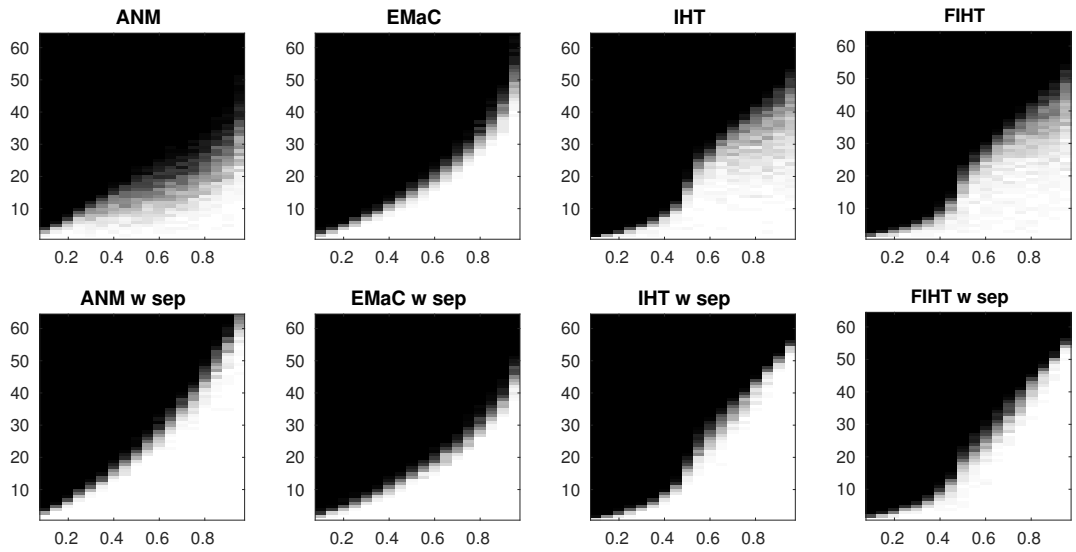

Figure 2.1: Phase transition comparisons: x-axis is $p = m/n$ and y-axis is r. Top: no restriction on frequencies of test signals; Bottom: wrap-around distances between frequencies is at least $1.5/n$.

First the figure shows that IHT and FIHT have similar empirical phase transitions for signals both with and without frequency separation. When the frequencies of test signals are separated, the phase transitions of IHT and FIHT are slightly lower than that of ANM, but higher than that of EMaC. The performance of ANM degrades severely when the frequencies of test signals are not sufficiently separated, while IHT and FIHT can still achieve good performance. The phase transitions of EMaC seem to be irrelevant to the separation of frequencies.

2.2.2 Computational Efficiency

In this section, we compare IHT and FIHT with PWGD on computational time. PWGD is an alternating projection algorithm which has been reported to be superior to ANM and EMaC in terms of computational efficiency [6]. In particular, we compare IHT and FIHT with an accelerated variant of PWGD based on Nesterov's memory technique. In our experiments, PWGD is also initialized via one step hard thresholding and the parameters are tuned as suggested in [6]. The algorithms are tested with $n \in \{3999, 7999\}$, $r \in \{15, 30\}$ and $m \in \{800, 1200\}$ and they are terminated whenever $\|\boldsymbol{x}_{l+1} - \boldsymbol{x}_l\|_2 / \|\boldsymbol{x}_l\|_2$ is less than 10^{-10}. For each triple (n, r, m), we run the algorithms on 10 randomly generated problem instances where the signals are formed in the same way as in Sec. 2.2.1. The average computational time and average number of iterations for each tested algorithm are presented in Tab. 2.1.

Table 2.1: Average computational time (seconds) and average number of iterations of PWGD, IHT and FIHT over 10 random problem instances per (n, r, m) for $n \in \{3999, 7999\}$, $r \in \{15, 30\}$ and $m \in \{800, 1200\}$.

r	15						30					
m	800			1200			800			1200		
	rel.err	iter	time	rel.err	iter	time	rel.err	iter	time	rel.err	iter	time
	n=3999											
PWGD	8.7e-11	178	10.4	7.9e-11	109	6.26	8.7e-11	223	34.1	6.0e-11	127	19.5
IHT	6.3e-11	23.4	1.14	6.4e-11	17.5	0.87	7.6e-11	38.7	4.65	7.0e-11	25.7	3.07
FIHT	6.1e-11	23.8	0.39	6.9e-11	17.6	0.29	7.5e-11	38.8	1.29	6.5e-11	25.8	0.86
	n=7999											
PWGD	8.3e-11	344	36.2	8.4e-11	212	22.3	7.5e-11	358	132	8.9e-11	259	91.6
IHT	7.6e-11	24.4	2.18	5.4e-11	20.0	1.74	8.3e-11	46.9	12.54	7.0e-11	30.3	7.72
FIHT	6.7e-11	24.6	0.64	5.8e-11	20.1	0.53	8.1e-11	47.7	2.72	6.9e-11	30.3	1.66

The table shows that it takes almost the same number of iterations for IHT and FIHT to converge below the given tolerance, but FIHT requires less computational time due to low per iteration computational complexity. Moreover, both IHT and FIHT are significantly faster than PWGD.

2.2.3 Robustness to Additive Noise

We demonstrate the performance of IHT and FIHT under additive noise by conducting tests with measurements corrupted by the vector

$$e = \sigma \cdot \|\mathcal{P}_\Omega(\boldsymbol{x})\|_2 \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2},$$

where $\boldsymbol{x}$ is the random signal to be reconstructed, the entries of $\boldsymbol{w}$ are i.i.d. standard Gaussian random variables and σ is referred to as the noise level.

Tests are conducted with 9 different values of σ from 10^{-4} to 1, corresponding to 9 equispaced signal-to-noise ratios (SNR) from 80 to 0 dB. For each σ, 10 random instances are tested for $(n, r, m) = (3999, 15, 800)$ and $(n, r, m) = (3999, 15, 1200)$, respectively. The algorithms are terminated when $\|\boldsymbol{x}_{l+1} - \boldsymbol{x}_l\|_2 / \|\boldsymbol{x}_l\|_2 < 10^{-10}$. The average relative reconstruction error in dB plotted against the SNR is presented in Fig. 2.2 for IHT and FIHT. The figure clearly shows the desirable linear scaling between the noise levels and the relative reconstruction errors for both IHT and FIHT. It can be further observed that the reconstruction error decreases as the number of measurements increases for both algorithms.

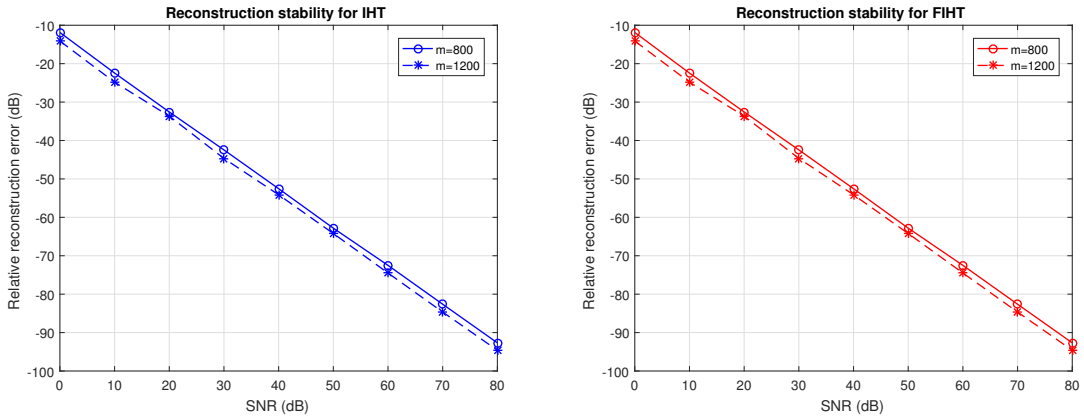

Figure 2.2: Reconstruction stability of IHT (Left) and FIHT (Right) under different SNR values.

2.2.4 Sensitivity to Model Order

In practical applications, we may not know the exact model order of the spectrally sparse signal but only have an estimation of it. We now examine the sensitivity of the proposed algorithms to under- and over-estimated model orders under both the noiseless and noisy settings. The experiments are conducted for $(n, r, m) = (3999, 15, 1200)$ following the noise model in Sec. 2.2.3. Three noise levels are tested: SNR=∞ (noise-free), SNR= 20 (light noise) and SNR= 0 (heavy noise). For each fixed noise level, the input rank varies from from $0.2r$ to $2r$ with increment $0.2r$, and then 10 random problem instances are tested. The algorithms are terminated if either $\|\boldsymbol{x}_{l+1}-\boldsymbol{x}_l\|_2/\|\boldsymbol{x}_l\|_2 < 10^{-10}$ or $\|\boldsymbol{x}_{l+1}-\boldsymbol{x}_l\|_2/\|\boldsymbol{x}_l\|_2 > 1$. We say an algorithm converges and has successfully returned an solution if $\|\boldsymbol{x}_{l+1} - \boldsymbol{x}_l\|_2/\|\boldsymbol{x}_l\|_2 < 10^{-10}$, otherwise we say the algorithm diverges and fails to return a solution. The solution rates out of 10 problem instances and the average SNR over all the solutions are reported in Tab. 2.2. When the algorithm diverges in all the 10 random trials, the average SNR

is marked as NA in the table.

Table 2.2: Solution rates and average SNR computed from outputs of IHT and FIHT over 10 random problem instances under different noise levels and test ranks for $(n, r, m) = (3999, 15, 1200)$.

Test Rank		3	6	9	12	15	18	21	24	27	30
		SNR= ∞									
IHT	Solution Rate	0.7	0.8	0.8	0.8	1	0.8	0.4	0.5	0.1	0.1
	Average SNR	1.719	3.352	5.793	9.810	215.9	215.9	216.0	219.6	221.9	223.4
FIHT	Solution Rate	0.9	0.8	0.9	1	1	0.9	0.6	0.3	0.2	0.2
	Average SNR	1.690	3.586	5.788	9.709	214.6	219.0	220.7	216.2	220.2	214.8
		SNR= 20									
IHT	Solution Rate	0.8	0.8	0.8	0.9	1	0.4	0.2	0.1	0	0.1
	Average SNR	1.725	3.523	5.791	9.768	34.25	31.71	29.83	29.35	NA	27.98
FIHT	Solution Rate	0.9	0.9	0.9	0.9	1	0.8	0.8	1	0.9	1
	Average SNR	1.677	3.496	5.786	9.768	34.25	31.72	30.26	29.28	28.49	27.82
		SNR= 0									
IHT	Solution Rate	0.9	0.8	0.8	1	1	0.5	0.2	0.2	0	0.1
	Average SNR	1.600	3.246	5.259	8.394	14.32	11.42	9.737	9.249	NA	7.750
FIHT	Solution Rate	0.9	1	1	1	1	0.9	0.9	0.9	1	0.8
	Average SNR	1.600	3.181	5.328	8.347	14.32	11.67	10.26	9.275	8.482	7.757

When the input rank is under-estimated, Tab. 2.2 shows that both IHT and FIHT can converge to a local minima with high probability under all the test noise levels. Moreover, as the discrepancy between the estimated rank and the true rank increases, the average SNR decreases. When the rank is over-estimated, it becomes difficult for IHT to converge under all the noise levels. Though with high probability FIHT fails to return a solution when the rank is highly over-estimated in the noiseless tests, what makes it prominent is that it can always return a solution with an acceptable SNR when there exists noise. The study of this observation will be left for future work. A potential explanation is that random noise can help the algorithms

avoid the worst case while the projection onto the tangent space in FIHT has the denoising ability.

Next, we suggest a rank increasing heuristic for the algorithms when the underlying model order is not known a priori in practice. For conciseness, we restrict our focus to FIHT since it is superior to IHT. Starting from a sufficiently small rank, we execute FIHT until convergence and compute the relative mean square error (RMSE) on the observed entries; we then increase the rank until the RMSE does not improve significantly when the rank is increased. To validate the effectiveness of this heuristic, we test FIHT under two noisy settings, SNR= 20 and SNR= 0, with the input rank increasing from 1 to 29. The RMSE for each rank and the difference between two consecutive ranks are presented in Fig. 2.3. It is evident that when the input rank is greater than the true rank $r = 15$, the improvement of the RMSE becomes very marginal.

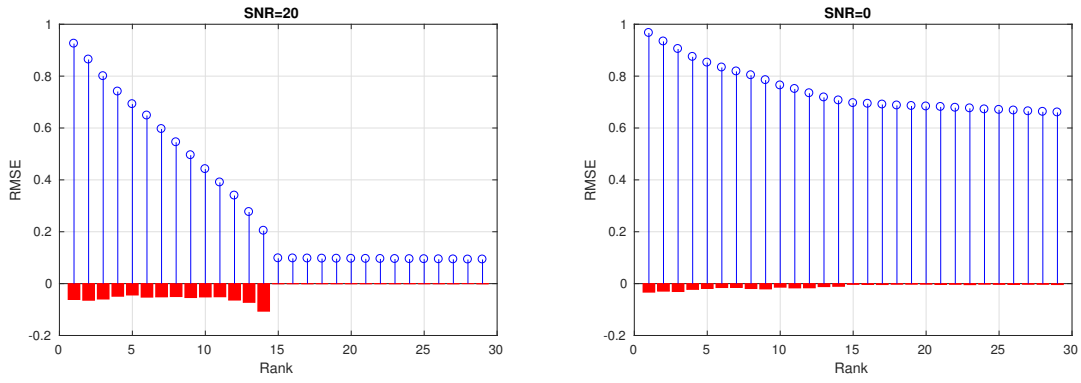

Figure 2.3: Demonstration of rank increasing heuristic for problem instances with SNR 20 (Left) and 0 (Right).

2.2.5 A 3D Example

To explore the capability of FIHT on handling large data, we conduct tests on a 3D damped signal with $n = N_1 \times N_2 \times N_3 = 128 \times 128 \times 1024 = 16,777,216$, $r = 20$ and $m = 167,772$ (about 1% of n). The signal is constructed to simulate real data from Nuclear Magnetic Resonance (NMR) spectroscopy. In this experiment, FIHT is terminated when $\|\boldsymbol{x}_{l+1} - \boldsymbol{x}_l\|_2 / \|\boldsymbol{x}_l\|_2 < 10^{-5}$. It takes FIHT **11** iterations and **2985** seconds to converge below the tolerance with the relative reconstruction error being 7.52×10^{-6}.

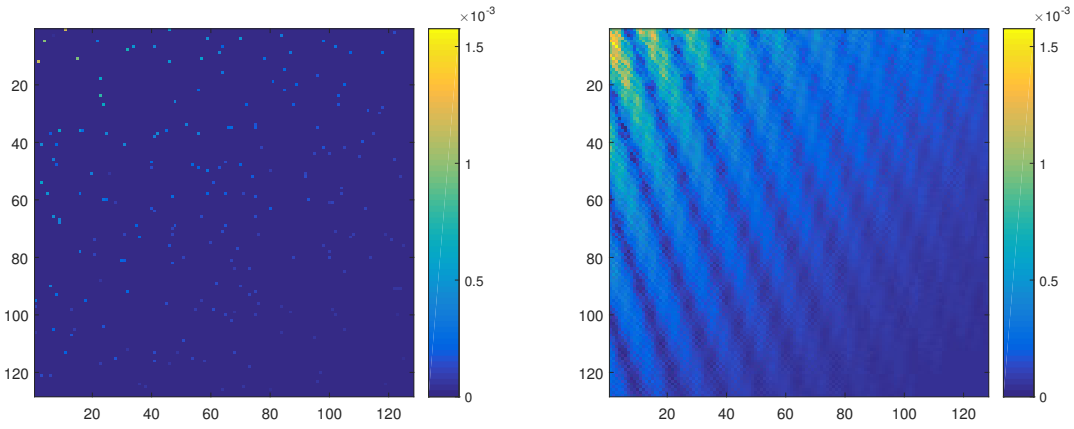

Figure 2.4: Samples (Left) on the slice with $N_3 = 812$ and its reconstruction result (Right) by FIHT.

To visualize the reconstruction result, we randomly pick a slice of the 3D signal and plot the amplitudes of sampled and reconstructed entries on this slice in Fig. 2.4. The differences between each entry of the original and reconstructed signals on the same slice is plotted in Fig. 2.5, which shows that the reconstruction is very accurate.

Furthermore, the plots in Fig. 2.6 compare the projection spectra of the original signal and the reconstructed one, which is obtained by first taking the Fourier transform of the 3D signal and then sum the spectrum along the third dimension.

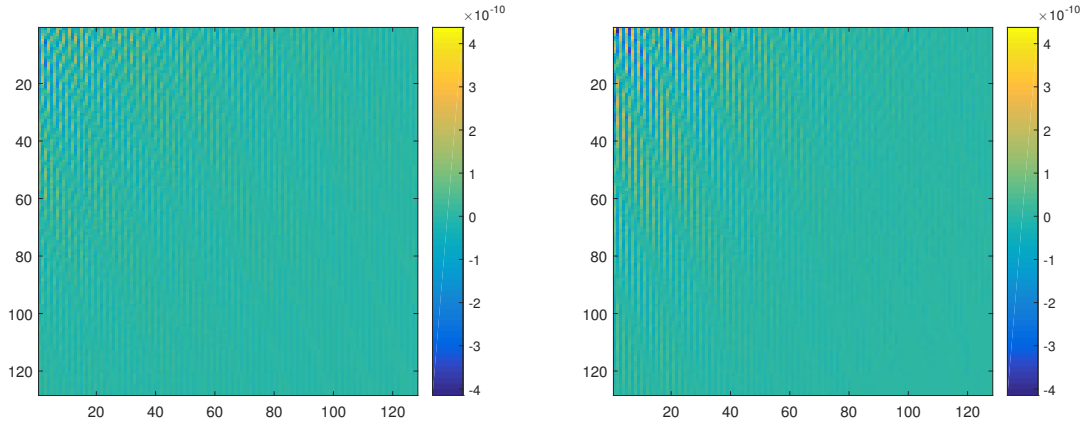

Figure 2.5: Real (Left) and imaginary (Right) parts of reconstruction errors for each entry on the slice with $N_3 = 812$.

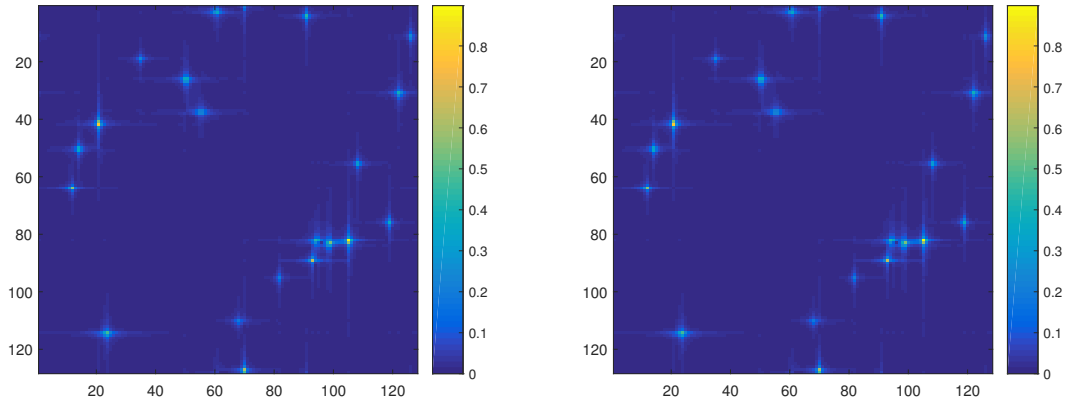

Figure 2.6: Projection spectra of the original signal (Left) and its reconstruction result (Right) by FIHT.

2.3 Proofs

This section presents the proofs for the theoretical results in Sec. 2.1.3. We first introduce several new variables and notation. Recall that $\mathcal{H}$ is a Hankel operator which maps a vector to a Hankel matrix and $\mathcal{H}^*$ is the adjoint of $\mathcal{H}$. Moreover, $\mathcal{D}^2 = \mathcal{H}^*\mathcal{H} = \mathrm{diag}(w_0, \cdots, w_{n-1})$ is a diagonal operator which multiply the a-th entry of a vector by the number of elements in the a-th anti-diagonal of the corresponding Hankel matrix. Recall that $\{\boldsymbol{H}_a\}_{a=0}^{n-1} \subset \mathbb{C}^{n_1 \times n_2}$ forms an orthonormal basis for all the $n_1 \times n_2$ Hankel matrices with $n_1 + n_2 = n + 1$.

Define $\mathcal{G} = \mathcal{H}\mathcal{D}^{-1}$. Then the adjoint of $\mathcal{G}$ is given by $\mathcal{G}^* = \mathcal{D}^{-1}\mathcal{H}^*$. It can be easily verified that $\mathcal{G}$ and $\mathcal{G}^*$ have the following properties:

- $\mathcal{G}^*\mathcal{G} = \mathcal{I}$, $\|\mathcal{G}\| \leq 1$, and $\|\mathcal{G}^*\| \leq 1$;
- $\mathcal{G}\boldsymbol{z} = \sum_{a=0}^{n-1} z_a \boldsymbol{H}_a, \ \forall \boldsymbol{z} \in \mathbb{C}^n$;
- $\mathcal{G}^*\boldsymbol{Z} = \{\langle \boldsymbol{Z}, \boldsymbol{H}_a \rangle\}_{a=0}^{n-1}, \ \forall \boldsymbol{Z} \in \mathbb{C}^{n_1 \times n_2}$.

Notice that the iteration of FIHT (Alg. 2.2) can be written in a compact form

$$\boldsymbol{x}_{l+1} = \mathcal{H}^\dagger \mathcal{T}_r \mathcal{P}_{\mathcal{S}_l} \mathcal{H}(\boldsymbol{x}_l + p^{-1}\mathcal{P}_\Omega(\boldsymbol{x} - \boldsymbol{x}_l)). \tag{2.3}$$

If we define $\boldsymbol{y} = \mathcal{D}\boldsymbol{x}$ and $\boldsymbol{y}_l = \mathcal{D}\boldsymbol{x}_l$, the following iteration can be established for $\boldsymbol{y}_l$

$$\boldsymbol{y}_{l+1} = \mathcal{G}^* \mathcal{T}_r \mathcal{P}_{\mathcal{S}_l} \mathcal{G}(\boldsymbol{y}_l + p^{-1}\mathcal{P}_\Omega(\boldsymbol{y} - \boldsymbol{y}_l)) \tag{2.4}$$

since $\mathcal{P}_\Omega$ and $\mathcal{D}^{-1}$ commute with each other. *For ease of exposition, we will prove the*

lemmas and theorems in Sec. 2.1.3 in terms of $\boldsymbol{y}_l$ *and* $\boldsymbol{y}$ *but note that the results in terms of* $\boldsymbol{x}_l$ *and* $\boldsymbol{x}$ *follow immediately since* $\mathcal{H}\boldsymbol{x} = \mathcal{G}\boldsymbol{y}$ *and*

$$\|\boldsymbol{x}_l - \boldsymbol{x}\| = \left\|\mathcal{D}^{-1}(\boldsymbol{y}_l - \boldsymbol{y})\right\| \leq \|\boldsymbol{y}_l - \boldsymbol{y}\|. \tag{2.5}$$

The following supplementary results from the literature but using our notation will be used repeatedly in the proofs of the main results.

Lemma 2.3.1 ([45, Prop. 3.3])**.** *Under the sampling with replacement model, the maximum number of repetitions of any entry in* Ω *is less than* $8\log(n)$ *with probability at least* $1 - n^{-2}$ *provided* $n \geq 9$*.*

Lemma 2.3.2 ([15, Lem. 3])**.** *Let* $\boldsymbol{U} \in \mathbb{C}^{n_1 \times r}$ *and* $\boldsymbol{V} \in \mathbb{C}^{n_2 \times r}$ *be two orthogonal matrices which satisfy*

$$\|\mathcal{P}_{\boldsymbol{U}}(\boldsymbol{H}_a)\|_F^2 \leq \frac{\mu c_s r}{n} \quad \textit{and} \quad \|\mathcal{P}_{\boldsymbol{V}}(\boldsymbol{H}_a)\|_F^2 \leq \frac{\mu c_s r}{n}.$$

Let $\mathcal{S}$ *be the subspace defined in* (2.2)*. Then*

$$\|\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathcal{S}} - p^{-1}\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}\| \leq \sqrt{\frac{32\mu c_s r\log(n)}{m}} \tag{2.6}$$

holds with probability at least $1 - n^{-2}$ *provided that*

$$m \geq 32\mu c_s r\log(n).$$

Lemma 2.3.3 ([55, Lem. 4.1]). *Let $\boldsymbol{L}_l = \boldsymbol{U}_l \boldsymbol{\Sigma}_l \boldsymbol{V}_l^*$ be another rank r matrix and $\mathcal{S}_l$ be the tangent space of the rank r matrix manifold at $\boldsymbol{L}_l$ as defined in* (2.2)*. Then*

$$\|(\mathcal{I} - \mathcal{P}_{\mathcal{S}_l})(\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y})\|_F \leq \frac{\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F^2}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})}, \quad \|\mathcal{P}_{\mathcal{S}_l} - \mathcal{P}_{\mathcal{S}}\| \leq \frac{2\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})}.$$

Lemma 2.3.4 ([51, Thm. 1.6]). *Consider a finite sequence $\{\boldsymbol{Z}_k\}$ of independent, random matrices with dimensions $d_1 \times d_2$. Assume that each random matrix satisfies*

$$\mathbb{E}(\boldsymbol{Z}_k) = 0 \quad \textit{and} \quad \|\boldsymbol{Z}_k\| \leq R \quad \textit{almost surely.}$$

Define

$$\sigma^2 := \max\left\{ \left\| \sum_k \mathbb{E}(\boldsymbol{Z}_k \boldsymbol{Z}_k^*) \right\|, \left\| \sum_k \mathbb{E}(\boldsymbol{Z}_k^* \boldsymbol{Z}_k) \right\| \right\}.$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{ \left\| \sum_k \boldsymbol{Z}_k \right\| \geq t \right\} \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right).$$

2.3.1 Local Convergence

We begin with a deterministic convergence result which characterizes the *basin of attraction* for FIHT. If the initial guess is located in this attraction region, FIHT will converge linearly to the underlying true solution.

Theorem 2.3.1. *Assume* $0 < \varepsilon_0 < \frac{1}{10}$ *and the following conditions*

$$\|\mathcal{P}_\Omega\| \leq 8\log(n), \tag{2.7}$$

$$\|\mathcal{P}_\mathcal{S}\mathcal{G}\mathcal{G}^*\mathcal{P}_\mathcal{S} - p^{-1}\mathcal{P}_\mathcal{S}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*\mathcal{P}_\mathcal{S}\| \leq \varepsilon_0, \tag{2.8}$$

$$\frac{\|\boldsymbol{L}_0 - \mathcal{G}\boldsymbol{y}\|_F}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})} \leq \frac{p^{1/2}\varepsilon_0}{16\log(n)(1+\varepsilon_0)} \tag{2.9}$$

are satisfied. Then the iterate $\boldsymbol{y}_l$ *in* (2.4) *satisfies* $\|\boldsymbol{y}_l - \boldsymbol{y}\| \leq \nu^l \|\boldsymbol{L}_0 - \mathcal{G}\boldsymbol{y}\|_F$ *with* $\nu = 10\varepsilon_0 < 1$.

The proof of Thm. 2.3.1 makes use of the restricted isometry property of $\mathcal{P}_\Omega(\cdot)$ on $\mathcal{S}_l$ when $\boldsymbol{L}_l$ is in a small neighborhood of $\mathcal{G}\boldsymbol{y}$.

Lemma 2.3.5. *Suppose* (2.7), (2.8) *hold and*

$$\frac{\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})} \leq \frac{p^{1/2}\varepsilon_0}{16\log(n)(1+\varepsilon_0)}. \tag{2.10}$$

Then we have

$$\|\mathcal{P}_\Omega\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l}\| \leq 8\log(n)(1+\varepsilon_0)p^{1/2} \tag{2.11}$$

and

$$\|\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l} - p^{-1}\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l}\| \leq 4\varepsilon_0. \tag{2.12}$$

Proof. Since $\|\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{P}_{\Omega}\| = \|(\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{P}_{\Omega})^*\| = \|\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}\|$, for any $\boldsymbol{Z} \in \mathbb{C}^{n_1\times n_2}$,

$$\begin{aligned}\|\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}(\boldsymbol{Z})\|^2 &= \langle \mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}(\boldsymbol{Z}), \mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}(\boldsymbol{Z})\rangle \\ &\leq 8\log(n)\langle \mathcal{G}^*\mathcal{P}_{\mathcal{S}}(\boldsymbol{Z}), \mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}(\boldsymbol{Z})\rangle \\ &= 8\log(n)\langle \boldsymbol{Z}, \mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}(\boldsymbol{Z})\rangle \\ &\leq 8\log(n)(1+\varepsilon_0)p\|\boldsymbol{Z}\|_F^2\end{aligned}$$

where the first inequality follows from (2.7) and the second inequality follows from (2.8). So it follows that $\|\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{P}_{\Omega}\| = \|\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}\| \leq \sqrt{8\log(n)(1+\varepsilon_0)p}$ and

$$\begin{aligned}\|\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l}\| &\leq \|\mathcal{P}_{\Omega}\mathcal{G}^*(\mathcal{P}_{\mathcal{S}_l} - \mathcal{P}_{\mathcal{S}})\| + \|\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}\| \\ &\leq 8\log(n)\frac{2\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})} + \|\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}\| \\ &\leq 8\log(n)\frac{p^{1/2}\varepsilon_0}{8\log(n)(1+\varepsilon_0)} + \sqrt{8\log(n)(1+\varepsilon_0)p} \\ &\leq 8\log(n)(1+\varepsilon_0)p^{1/2},\end{aligned}$$

where the second inequality follows from (2.7) and Lem. 2.3.3, the third inequality follows from (2.10). Finally,

$$\begin{aligned}&\|\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l} - p^{-1}\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l}\| \\ \leq&\|\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathcal{S}} - p^{-1}\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}}\| + \|(\mathcal{P}_{\mathcal{S}} - \mathcal{P}_{\mathcal{S}_l})\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l}\| + \|\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{G}^*(\mathcal{P}_{\mathcal{S}} - \mathcal{P}_{\mathcal{S}_l})\| \\ &+ \|p^{-1}(\mathcal{P}_{\mathcal{S}} - \mathcal{P}_{\mathcal{S}_l})\mathcal{G}\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l}\| + \|p^{-1}\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{P}_{\Omega}\mathcal{G}^*(\mathcal{P}_{\mathcal{S}} - \mathcal{P}_{\mathcal{S}_l})\| \\ \leq&\varepsilon_0 + \frac{4\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})} + p^{-1}\cdot\frac{2\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})}\cdot(\|\mathcal{P}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l}\| + \|\mathcal{P}_{\mathcal{S}}\mathcal{G}\mathcal{P}_{\Omega}\|) \leq 4\varepsilon_0,\end{aligned}$$

which completes the proof of (2.12). □

Proof of Thm. 2.3.1. First note that $\boldsymbol{L}_{l+1} = \mathcal{T}_r(\boldsymbol{W}_l)$, where

$$\begin{aligned}\boldsymbol{W}_l &= \mathcal{P}_{\mathcal{S}_l}\mathcal{H}(\boldsymbol{x}_l + p^{-1}\mathcal{P}_\Omega(\boldsymbol{x} - \boldsymbol{x}_l)) \\ &= \mathcal{P}_{\mathcal{S}_l}\mathcal{G}(\boldsymbol{y}_l + p^{-1}\mathcal{P}_\Omega(\boldsymbol{y} - \boldsymbol{y}_l)).\end{aligned}$$

So we have

$$\begin{aligned}&\|\boldsymbol{L}_{l+1} - \mathcal{G}\boldsymbol{y}\|_F \\ \leq&\|\boldsymbol{W}_l - \boldsymbol{L}_{l+1}\|_F + \|\boldsymbol{W}_l - \mathcal{G}\boldsymbol{y}\|_F \leq 2\|\boldsymbol{W}_l - \mathcal{G}\boldsymbol{y}\|_F \\ =&2\|\mathcal{P}_{\mathcal{S}_l}\mathcal{G}(\boldsymbol{y}_l + p^{-1}\mathcal{P}_\Omega(\boldsymbol{y} - \boldsymbol{y}_l)) - \mathcal{G}\boldsymbol{y}\|_F \\ \leq&2\|\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\boldsymbol{y} - \mathcal{G}\boldsymbol{y}\|_F + 2\|(\mathcal{P}_{\mathcal{S}_l}\mathcal{G} - p^{-1}\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{P}_\Omega)(\boldsymbol{y}_l - \boldsymbol{y})\|_F \\ =&2\|(\mathcal{I} - \mathcal{P}_{\mathcal{S}_l})(\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y})\|_F + 2\|(\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{G}^* - p^{-1}\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*)(\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y})\|_F \\ \leq&2\|(\mathcal{I} - \mathcal{P}_{\mathcal{S}_l})(\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y})\|_F + 2\|(\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l} - p^{-1}\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*\mathcal{P}_{\mathcal{S}_l})(\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y})\|_F \\ &+ 2\|\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{G}^*(\mathcal{I} - \mathcal{P}_{\mathcal{S}_l})(\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y})\|_F + 2p^{-1}\|\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\mathcal{I} - \mathcal{P}_{\mathcal{S}_l})(\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y})\|_F \\ :=&I_1 + I_2 + I_3 + I_4,\end{aligned}$$

where the second inequality comes from the fact that $\boldsymbol{L}_{l+1}$ is the best rank r approximation to $\boldsymbol{W}_l$, the second equality follows from $(\mathcal{I} - \mathcal{P}_{\mathcal{S}_l})\boldsymbol{L}_l = 0$, $\boldsymbol{y}_l = \mathcal{G}^*\boldsymbol{L}_l$ and $\mathcal{G}^*\mathcal{G} = \mathcal{I}$.

Let us first assume (2.10) holds. Then the application of Lem. 2.3.3 gives

$$\begin{aligned} I_1 + I_3 + I_4 &\leq \left(\frac{4\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})} + 2p^{-1}\|\mathcal{P}_\Omega \mathcal{G}^* \mathcal{P}_{\mathcal{S}_l}\| \frac{\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})} \right) \|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F \\ &\leq 2\varepsilon_0 \|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F, \end{aligned}$$

where the last inequality follows from (2.8), (2.11) and $\|\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\mathcal{P}_\Omega\| = \|\mathcal{P}_\Omega \mathcal{G}^* \mathcal{P}_{\mathcal{S}_l}\|$. Moreover, (2.12) implies

$$I_2 \leq 8\varepsilon_0 \|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F.$$

Therefore putting the bounds for $I_1,\ I_2,\ I_3,$ and I_4 together gives

$$\|\boldsymbol{L}_{l+1} - \mathcal{G}\boldsymbol{y}\|_F \leq \nu \|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F,$$

where $\nu = 10\varepsilon_0 < 1$. Since (2.10) holds for $l = 0$ by the assumption of Thm. 2.3.1 and $\|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F$ is a contractive sequence, (2.10) holds for all $l \geq 0$. Thus

$$\|\boldsymbol{y}_l - \boldsymbol{y}\| = \|\mathcal{G}^*(\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y})\| \leq \|\boldsymbol{L}_l - \mathcal{G}\boldsymbol{y}\|_F \leq \nu^l \|\boldsymbol{L}_0 - \mathcal{G}\boldsymbol{y}\|_F,$$

where we have utilized the facts $\boldsymbol{y}_l = \mathcal{G}^*\boldsymbol{L}_l$, $\mathcal{G}^*\mathcal{G} = \mathcal{I}$ and $\|\mathcal{G}^*\| \leq 1$. □

2.3.2 Proofs of Lemma 2.1.1 and Theorem 2.1.1

Proof of Lem. 2.1.1. Recall that $\boldsymbol{L}_0 = \mathcal{T}_r(p^{-1}\mathcal{H}\mathcal{P}_\Omega(\boldsymbol{x})) = \mathcal{T}_r(p^{-1}\mathcal{G}\mathcal{P}_\Omega(\boldsymbol{y}))$ and $\mathcal{H}\boldsymbol{x} = \mathcal{G}\boldsymbol{y}$. Let us first bound $\|p^{-1}\mathcal{G}\mathcal{P}_\Omega(\boldsymbol{y}) - \mathcal{G}\boldsymbol{y}\|$. Since $p = \frac{m}{n}$, we have

$$p^{-1}\mathcal{G}\mathcal{P}_\Omega(\boldsymbol{y}) - \mathcal{G}\boldsymbol{y} = \sum_{k=1}^{m}\left(\frac{n}{m}y_{a_k}\boldsymbol{H}_{a_k} - \frac{1}{m}\mathcal{G}\boldsymbol{y}\right) := \sum_{k=1}^{m}\boldsymbol{Z}_{a_k}.$$

Because each a_k is drawn uniformly from $\{0, \cdots, n-1\}$, it is trivial that $\mathbb{E}\left(\boldsymbol{Z}_{a_k}\right) = 0$. Moreover, we have

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{Z}_{a_k}\boldsymbol{Z}_{a_k}^*\right) &= \mathbb{E}\left(\frac{n^2}{m^2}|y_{a_k}|^2\boldsymbol{H}_{a_k}\boldsymbol{H}_{a_k}^*\right) - \frac{1}{m^2}(\mathcal{G}\boldsymbol{y})(\mathcal{G}\boldsymbol{y})^* \\
&= \frac{n}{m^2}\sum_{a=0}^{n-1}|y_a|^2\boldsymbol{H}_a\boldsymbol{H}_a^* - \frac{1}{m^2}(\mathcal{G}\boldsymbol{y})(\mathcal{G}\boldsymbol{y})^* \\
&= \frac{n}{m^2}\boldsymbol{C} - \frac{1}{m^2}(\mathcal{G}\boldsymbol{y})(\mathcal{G}\boldsymbol{y})^*,
\end{aligned}$$

where $\boldsymbol{C}$ is a diagonal matrix which corresponds to the diagonal part of $(\mathcal{G}\boldsymbol{y})(\mathcal{G}\boldsymbol{y})^*$. Therefore

$$\begin{aligned}
\left\|\mathbb{E}\left(\sum_{k=1}^{m}\boldsymbol{Z}_{a_k}\boldsymbol{Z}_{a_k}^*\right)\right\| &\le \max\left\{\frac{n}{m}\|\boldsymbol{C}\|, \frac{1}{m}\|(\mathcal{G}\boldsymbol{y})(\mathcal{G}\boldsymbol{y})^*\|\right\} \\
&\le \frac{n}{m}\|\mathcal{G}\boldsymbol{y}\|_{2\to\infty}^2,
\end{aligned}$$

where $\|\mathcal{G}\boldsymbol{y}\|_{2\to\infty}$ denotes the maximum row ℓ_2 norm of $\mathcal{G}\boldsymbol{y}$. Similarly we can get

$$\left\|\mathbb{E}\left(\sum_{k=1}^{m}\boldsymbol{Z}_{a_k}^*\boldsymbol{Z}_{a_k}\right)\right\| \le \frac{n}{m}\|(\mathcal{G}\boldsymbol{y})^*\|_{2\to\infty}^2.$$

The definition of $\boldsymbol{H}_a$ in (1.4) implies $\|\boldsymbol{H}_a\| \leq \frac{1}{\sqrt{w_a}}$. So

$$\|\boldsymbol{Z}_{a_k}\| \leq \frac{n}{m}|y_{a_k}|\,\|\boldsymbol{H}_{a_k}\| + \frac{1}{m}\sum_{a=0}^{n-1}|y_a|\,\|\boldsymbol{H}_a\| \leq \frac{2n}{m}\left\|\mathcal{D}^{-1}\boldsymbol{y}\right\|_\infty.$$

By matrix Bernstein inequality in Lem. 2.3.4, one can show that there exists a universal constant $C > 0$ such that

$$\left\|\sum_{k=1}^{m}\boldsymbol{Z}_{a_k}\right\| \leq C\left(\sqrt{\frac{n\log(n)}{m}}\max\left\{\|\mathcal{G}\boldsymbol{y}\|_{2\to\infty}, \|(\mathcal{G}\boldsymbol{y})^*\|_{2\to\infty}\right\} + \frac{n\log(n)}{m}\left\|\mathcal{D}^{-1}\boldsymbol{y}\right\|_\infty\right)$$

with probability at least $1 - n^{-2}$. Consequently on the same event we have

$$\begin{aligned}
&\|\boldsymbol{L}_0 - \mathcal{G}\boldsymbol{y}\| \\
\leq&\left\|\boldsymbol{L}_0 - p^{-1}\mathcal{G}\mathcal{P}_\Omega(\boldsymbol{y})\right\| + \left\|p^{-1}\mathcal{G}\mathcal{P}_\Omega(\boldsymbol{y}) - \mathcal{G}\boldsymbol{y}\right\| \leq 2\left\|p^{-1}\mathcal{G}\mathcal{P}_\Omega(\boldsymbol{y}) - \mathcal{G}\boldsymbol{y}\right\| \\
\leq& C\left(\sqrt{\frac{n\log(n)}{m}}\max\left\{\|\mathcal{G}\boldsymbol{y}\|_{2\to\infty}, \|(\mathcal{G}\boldsymbol{y})^*\|_{2\to\infty}\right\} + \frac{n\log(n)}{m}\left\|\mathcal{D}^{-1}\boldsymbol{y}\right\|_\infty\right). \qquad (2.13)
\end{aligned}$$

Thus it only remains to bound $\max\left\{\|\mathcal{G}\boldsymbol{y}\|_{2\to\infty}, \|(\mathcal{G}\boldsymbol{y})^*\|_{2\to\infty}\right\}$ and $\|\mathcal{D}^{-1}\boldsymbol{y}\|_\infty$ in terms of $\|\mathcal{G}\boldsymbol{y}\|$. From $\mathcal{G}\boldsymbol{y} = \mathcal{H}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{E}_L\boldsymbol{D}\boldsymbol{E}_R^T$, we get

$$\begin{aligned}
\|\mathcal{G}\boldsymbol{y}\|_{2\to\infty}^2 &= \max_i \|\boldsymbol{e}_i^*(\mathcal{G}\boldsymbol{y})\|^2 = \max_i \|\boldsymbol{e}_i^*\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\|^2 \leq \max_i \|\boldsymbol{e}_i^*\boldsymbol{U}\|^2\|\boldsymbol{\Sigma}\|^2 \\
&= \max_i \left\|\boldsymbol{U}^{(i,:)}\right\|^2 \|\mathcal{G}\boldsymbol{y}\|_2^2 \leq \frac{\mu_0 c_s r}{n}\|\mathcal{G}\boldsymbol{y}\|_2^2, \qquad (2.14)
\end{aligned}$$

where the last inequality follows from Lem. 1.3.1. Similarly we also have

$$\|(\mathcal{G}\boldsymbol{y})^*\|_{2\to\infty}^2 \leq \frac{\mu_0 c_s r}{n}\|\mathcal{G}\boldsymbol{y}\|_2^2. \tag{2.15}$$

The infinity norm of $\mathcal{D}^{-1}\boldsymbol{y}$ can be bounded as follows

$$\begin{aligned}\left\|\mathcal{D}^{-1}\boldsymbol{y}\right\|_\infty = \|\mathcal{G}\boldsymbol{y}\|_\infty = \max_{i,j}|\boldsymbol{e}_i^*(\mathcal{G}\boldsymbol{y})\boldsymbol{e}_j| &\leq \max_{i,j}\|\boldsymbol{e}_i^*\boldsymbol{E}_L\|\,\|\boldsymbol{D}\|\,\left\|\boldsymbol{E}_R^T\boldsymbol{e}_j\right\| \\ &\leq r\,\|\boldsymbol{D}\| \leq r\left\|\boldsymbol{E}_L^\dagger\right\|\|\mathcal{G}\boldsymbol{y}\|\,\left\|(\boldsymbol{E}_R^T)^\dagger\right\| \leq \frac{\mu_0 c_s r}{n}\,\|\mathcal{G}\boldsymbol{y}\|\,,\end{aligned} \tag{2.16}$$

where the last inequality follows from the μ_0-incoherence of $\mathcal{G}\boldsymbol{y}$.

Finally inserting (2.14), (2.15) and (2.16) into (2.13) gives

$$\|\boldsymbol{L}_0 - \mathcal{G}\boldsymbol{y}\| \leq C\sqrt{\frac{\mu_0 c_s r\log(n)}{m}}\,\|\mathcal{G}\boldsymbol{y}\|$$

provided $m \geq \mu_0 c_s r\log(n)$. □

Proof of Thm. 2.1.1. Following from (2.5), we only need to verify when the three conditions in Thm. 2.3.1 are satisfied. Lem. 2.3.1 implies (2.7) holds with probability at least $1-n^{-2}$. Lems. 1.3.1 and 2.3.2 guarantees (2.8) is true with probability at least $1-n^{-2}$ if $m \geq C\varepsilon_0^{-2}\mu_0 c_s r\log(n)$ for a sufficiently large numerical constant $C>0$. Similarly (2.9) can be satisfied with probability at least $1-n^2$ if $m \geq C(1+\varepsilon_0)\varepsilon_0^{-1}\mu_0^{1/2}c_s^{1/2}\kappa r n^{1/2}\log^{3/2}(n)$ following Lem. 2.1.1 and the fact $\|\boldsymbol{L}_0-\mathcal{G}\boldsymbol{y}\|_F \leq \sqrt{2r}\,\|\boldsymbol{L}_0-\mathcal{G}\boldsymbol{y}\|$, where κ denotes the condition number of $\mathcal{G}\boldsymbol{y}$. Taking an upper bound on the number of measurements completes the proof of Thm. 2.1.1. □

2.3.3 Proofs of Lemma 2.1.2 and Theorem 2.1.2

The proof of Lem. 2.1.2 relies on the following estimation of

$$\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\left(\widehat{p}^{-1}\mathcal{P}_{\widehat{\Omega}_{l+1}}-\mathcal{I}\right)\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)\right\|,$$

which is a generalization of the asymmetric restricted isometry property [55] from matrix completion to low rank Hankel matrix completion.

Lemma 2.3.6. *Assume there exists a numerical constant μ such that*

$$\|\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\boldsymbol{H}_a\|_F^2 \le \frac{\mu c_s r}{n}, \quad \|\mathcal{P}_{\widehat{\boldsymbol{V}}_l}\boldsymbol{H}_a\|_F^2 \le \frac{\mu c_s r}{n}, \tag{2.17}$$

and

$$\|\mathcal{P}_{\boldsymbol{U}}\boldsymbol{H}_a\|_F^2 \le \frac{\mu c_s r}{n}, \quad \|\mathcal{P}_{\boldsymbol{V}}\boldsymbol{H}_a\|_F^2 \le \frac{\mu c_s r}{n}. \tag{2.18}$$

for all $0 \le a \le n-1$. Let $\widehat{\Omega}_{l+1} = \{a_k \mid k = 1, \cdots, \widehat{m}\}$ be a set of indices sampled with replacement. If $\mathcal{P}_{\widehat{\Omega}_{l+1}}$ is independent of $\boldsymbol{U}$, $\boldsymbol{V}$, $\widehat{\boldsymbol{U}}_l$ and $\widehat{\boldsymbol{V}}_l$, then

$$\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\left(\mathcal{I}-\widehat{p}^{-1}\mathcal{P}_{\widehat{\Omega}_{l+1}}\right)\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)\right\| \le \sqrt{\frac{160\mu c_s r\log(n)}{\widehat{m}}}$$

with probability at least $1-n^{-2}$ provided

$$\widehat{m} \ge \frac{125}{18}\mu c_s r\log(n).$$

Proof. Since for any $\boldsymbol{Z} \in \mathbb{C}^{n_1 \times n_2}$

$$\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{P}_{\widehat{\Omega}_{l+1}}\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)(\boldsymbol{Z}) = \sum_{k=1}^{\widehat{m}}\left\langle \boldsymbol{Z}, \left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)(\boldsymbol{H}_{a_k})\right\rangle \mathcal{P}_{\mathcal{S}_l}(\boldsymbol{H}_{a_k}),$$

we can rewrite $\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{P}_{\widehat{\Omega}_{l+1}}\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)$ as

$$\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{P}_{\widehat{\Omega}_{l+1}}\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right) = \sum_{k=1}^{\widehat{m}}\mathcal{P}_{\mathcal{S}_l}(\boldsymbol{H}_{a_k}) \otimes \left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)(\boldsymbol{H}_{a_k}).$$

Define the random operator

$$\mathbb{R}_{a_k} = \mathcal{P}_{\widehat{\mathcal{S}}_l}(\boldsymbol{H}_{a_k}) \otimes \left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)(\boldsymbol{H}_{a_k}) - \frac{1}{n}\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right).$$

Then it is easy to see that $\mathbb{E}\left(\mathbb{R}_{a_k}\right) = 0$. By assumption, for any $0 \le a \le n-1$,

$$\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\left(\boldsymbol{H}_a\right)\|_F^2 \le \|\mathcal{P}_{\widehat{U}_l}\left(\boldsymbol{H}_a\right)\|_F^2 + \|\mathcal{P}_{\widehat{V}_l}\left(\boldsymbol{H}_a\right)\|_F^2 \le \frac{2\mu c_s r}{n}.$$

So

$$\begin{aligned}
&\|\mathbb{R}_{a_k}\| \\
\le &\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\left(\boldsymbol{H}_{a_k}\right)\right\|_F \left\|\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)(\boldsymbol{H}_{a_k})\right\|_F + \frac{1}{n}\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)\right\| \\
\le &\frac{5\mu c_s r}{n}.
\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}(\mathcal{R}_{a_k}\mathcal{R}_{a_k}^*) =&\mathbb{E}\left(\left\|\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)(\boldsymbol{H}_{a_k})\right\|_F^2\mathcal{P}_{\widehat{\mathcal{S}}_l}(\boldsymbol{H}_{a_k})\otimes\mathcal{P}_{\widehat{\mathcal{S}}_l}(\boldsymbol{H}_{a_k})\right)\\&-\frac{1}{n^2}\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)^2\mathcal{G}\mathcal{G}^*\mathcal{P}_{\widehat{\mathcal{S}}_l}.\end{aligned}$$

$\left\|\mathbb{E}(\mathcal{R}_{a_k}\mathcal{R}_{a_k}^*)\right\|$ can be bounded as follows

$$\begin{aligned}\left\|\mathbb{E}(\mathcal{R}_{a_k}\mathcal{R}_{a_k}^*)\right\| &\le \left\|\mathbb{E}\left(\left\|\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)(\boldsymbol{H}_{a_k})\right\|_F^2\mathcal{P}_{\widehat{\mathcal{S}}_l}(\boldsymbol{H}_{a_k})\otimes\mathcal{P}_{\widehat{\mathcal{S}}_l}(\boldsymbol{H}_{a_k})\right)\right\|+\frac{4}{n^2}\\&\le\frac{4\mu c_s r}{n}\left\|\mathbb{E}\left(\mathcal{P}_{\widehat{\mathcal{S}}_l}(\boldsymbol{H}_{a_k})\otimes\mathcal{P}_{\widehat{\mathcal{S}}_l}(\boldsymbol{H}_{a_k})\right)\right\|+\frac{4}{n^2}\\&=\frac{4\mu c_s r}{n^2}\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\widehat{\mathcal{S}}_l}\right\|+\frac{4}{n^2}\\&\le\frac{8\mu c_s r}{n^2}.\end{aligned}$$

This implies

$$\left\|\mathbb{E}\left(\sum_{k=1}^{\widehat{m}}\mathbb{R}_{a_k}\mathbb{R}_{a_k}^*\right)\right\|\le\sum_{k=1}^{\widehat{m}}\left\|\mathbb{E}(\mathcal{R}_{a_k}\mathcal{R}_{a_k}^*)\right\|\le\frac{8\mu c_s r\widehat{m}}{n^2}.$$

We can similarly obtain

$$\left\|\mathbb{E}\left(\sum_{k=1}^{\widehat{m}}\mathbb{R}_{a_k}^*\mathbb{R}_{a_k}\right)\right\|\le\frac{12\mu c_s r\widehat{m}}{n^2}.$$

So the application of the matrix Bernstein inequality in Lem. 2.3.4 gives

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{\widehat{m}}\mathcal{R}_{a_k}\right\|\ge t\right\}\le 2n_1n_2\exp\left(\frac{-t^2/2}{\frac{12\mu c_s\widehat{m}r}{n^2}+\frac{5\mu c_s r}{n}t/3}\right).$$

If $t \leq \frac{24\widehat{m}}{5n}$, then

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{\widehat{m}} \mathcal{R}_{a_k}\right\| \geq t\right\} \leq 2n_1n_2\text{exp}\left(\frac{-t^2/2}{\frac{20\mu c_s\widehat{m}r}{n^2}}\right) \leq n^2\text{exp}\left(\frac{-t^2/2}{\frac{20\mu c_s\widehat{m}r}{n^2}}\right).$$

Setting $t = \sqrt{\frac{160\mu c_s\widehat{m}r\log(n)}{n^2}}$ gives

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{\widehat{m}} \mathcal{R}_{a_k}\right\| \geq t\right\} \leq n^{-2}.$$

The condition $t \leq \frac{24\widehat{m}}{5n}$ implies $\widehat{m} \geq \frac{125}{18}\mu c_s r\log(n)$. The proof is complete because

$$\frac{n}{\widehat{m}}\sum_{k=1}^{\widehat{m}} \mathcal{R}_{a_k} = \mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\left(\widehat{p}^{-1}\mathcal{P}_{\widehat{\Omega}_{l+1}} - \mathcal{I}\right)\mathcal{G}^*\left(\mathcal{P}_{\boldsymbol{U}} - \mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right).$$

□

The following lemma from [55] will also be used in the proof of Lem. 2.1.2.

Lemma 2.3.7. *Let $\widetilde{\boldsymbol{L}}_l = \widetilde{\boldsymbol{U}}_l\widetilde{\boldsymbol{\Sigma}}_l\widetilde{\boldsymbol{V}}_l^*$ and $\mathcal{G}\boldsymbol{y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be two rank r matrices which satisfy*

$$\left\|\widetilde{\boldsymbol{L}}_l - \mathcal{G}\boldsymbol{y}\right\|_F \leq \frac{\sigma_{\min}(\mathcal{G}\boldsymbol{y})}{10\sqrt{2}}.$$

Assume $\left\|\boldsymbol{U}^{(i,:)}\right\|^2 \leq \frac{\mu_0 c_s r}{n}$ and $\left\|\boldsymbol{V}^{(j,:)}\right\|^2 \leq \frac{\mu_0 c_s r}{n}$. Then the matrix $\widehat{\boldsymbol{L}}_l = \text{Trim}_{\mu_0}(\widetilde{\boldsymbol{L}}_l) = \widehat{\boldsymbol{U}}_l\widehat{\boldsymbol{\Sigma}}_l\widehat{\boldsymbol{V}}_l^*$ returned by Alg. 2.4 satisfies

$$\left\|\widehat{\boldsymbol{L}}_l - \mathcal{G}\boldsymbol{y}\right\|_F \leq 8\kappa\left\|\widetilde{\boldsymbol{L}}_l - \mathcal{G}\boldsymbol{y}\right\|_F \quad \textit{and} \quad \max\left\{\left\|\widehat{\boldsymbol{U}}^{(i,:)}\right\|^2, \left\|\widehat{\boldsymbol{V}}^{(j,:)}\right\|^2\right\} \leq \frac{100\mu_0 c_s r}{81n},$$

where κ denotes the condition number of $\mathcal{G}\boldsymbol{y}$.

Proof of Lem. 2.1.2. Let us first assume that

$$\left\|\widetilde{\boldsymbol{L}}_l - \mathcal{G}\boldsymbol{y}\right\|_F \leq \frac{\sigma_{\min}(\mathcal{G}\boldsymbol{y})}{256\kappa^2}. \tag{2.19}$$

Then the application of Lem. 2.3.7 implies that

$$\left\|\widehat{\boldsymbol{L}}_l - \mathcal{G}\boldsymbol{y}\right\|_F \leq 8\kappa \left\|\widetilde{\boldsymbol{L}}_l - \mathcal{G}\boldsymbol{y}\right\|_F \tag{2.20a}$$

$$\max\left\{\left\|\widehat{\boldsymbol{U}}^{(i,:)}\right\|^2, \left\|\widehat{\boldsymbol{V}}^{(j,:)}\right\|^2\right\} \leq \frac{100\mu_0 c_s r}{81n} \tag{2.20b}$$

by noting that $\left\|\boldsymbol{U}^{(i,:)}\right\|^2 \leq \frac{\mu_0 c_s r}{n}$ and $\left\|\boldsymbol{V}^{(j,:)}\right\|^2 \leq \frac{\mu_0 c_s r}{n}$ following from Lem. 1.3.1. Moreover, direct calculation gives

$$\left\|\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\boldsymbol{H}_a\right\|_F^2 = \left\|\widehat{\boldsymbol{U}}_l^*\boldsymbol{H}_a\right\|_F^2 = \frac{1}{|\Gamma_a|}\sum_{i\in\Gamma_a}\left\|\left(\widehat{\boldsymbol{U}}_l\right)^{(i,:)}\right\|_2^2 \leq \frac{100\mu_0 c_s r}{81n}, \tag{2.21}$$

where Γ_a is the set of row indices for non-zero entries in $\boldsymbol{H}_a$ with cardinality $|\Gamma_a| = w_a$. Similarly,

$$\left\|\mathcal{P}_{\widehat{\boldsymbol{V}}_l}\boldsymbol{H}_a\right\|_F^2 \leq \frac{100\mu_0 c_s r}{81n}. \tag{2.22}$$

Recall that $\boldsymbol{y} = \mathcal{D}\boldsymbol{x}$ and $\mathcal{G}\boldsymbol{y} = \mathcal{H}\boldsymbol{x}$. Define $\widehat{\boldsymbol{y}}_l = \mathcal{D}\widehat{\boldsymbol{x}}_l$. Then $\widehat{\boldsymbol{y}}_l = \mathcal{G}^*\widehat{\boldsymbol{L}}_l$ and

$$\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{H}\left(\widehat{\boldsymbol{x}}_l + \widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}}\left(\boldsymbol{x} - \widehat{\boldsymbol{x}}_l\right)\right) = \mathcal{P}_{\mathcal{S}_l}\mathcal{G}\left(\widehat{\boldsymbol{y}}_l + \widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}}\left(\boldsymbol{y} - \widehat{\boldsymbol{y}}_l\right)\right).$$

Consequently,

$$
\begin{aligned}
&\|\widetilde{\boldsymbol{L}}_{l+1}-\mathcal{G}\boldsymbol{y}\|_F\\
\leq&2\left\|\mathcal{P}_{\mathcal{S}_l}\mathcal{G}\left(\widehat{\boldsymbol{y}}_l+\widehat{p}^{-1}\mathcal{P}_{\widehat{\Omega}_{l+1}}\left(\boldsymbol{y}-\widehat{\boldsymbol{y}}_l\right)\right)-\mathcal{G}\boldsymbol{y}\right\|_F\\
\leq&2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\boldsymbol{y}-\mathcal{G}\boldsymbol{y}\right\|_F+2\left\|\left(\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}-\widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{P}_{\widehat{\Omega}_{l+1}}\right)\left(\widehat{\boldsymbol{y}}_l-\boldsymbol{y}\right)\right\|_F\\
=&2\left\|\left(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l}\right)\mathcal{G}\boldsymbol{y}\right\|_F+2\left\|\left(\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{G}^*-\widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{P}_{\widehat{\Omega}_{l+1}}\mathcal{G}^*\right)\left(\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right)\right\|_F\\
\leq&2\left\|\left(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l}\right)\left(\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right)\right\|_F\\
&+2\left\|\left(\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\widehat{\mathcal{S}}_l}-\widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\mathcal{P}_{\widehat{\Omega}_{l+1}}\mathcal{G}^*\mathcal{P}_{\widehat{\mathcal{S}}_l}\right)\left(\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right)\right\|_F\\
&+2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\left(\mathcal{I}-\widehat{p}^{-1}\mathcal{P}_{\widehat{\Omega}_{l+1}}\right)\mathcal{G}^*\left(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l}\right)\left(\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right)\right\|_F\\
:=&I_5+I_6+I_7.
\end{aligned}
$$

The first item I_5 can be bounded as

$$
I_5\leq\frac{2\left\|\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right\|_F^2}{\sigma_{\min}(\mathcal{G}\boldsymbol{y})}\leq\frac{1}{2}\left\|\widetilde{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right\|_F,
$$

which follows from Lem. 2.3.3, inequality (2.20a) and the assumption (2.19). The application of Lem. 2.3.2 together with (2.21) and (2.22) implies

$$
I_6\leq2\sqrt{\frac{3200\mu_0c_sr\log(n)}{81\widehat{m}}}\left\|\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right\|_F\leq16\kappa\sqrt{\frac{3200\mu_0c_sr\log(n)}{81\widehat{m}}}\left\|\widetilde{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right\|_F
$$

with probability at least $1-n^2$. To bound I_7, first note that

$$\begin{aligned}\left(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l}\right)\left(\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right) &= \left(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l}\right)\left(-\mathcal{G}\boldsymbol{y}\right) = \left(\boldsymbol{I}-\widehat{\boldsymbol{U}}_l\widehat{\boldsymbol{U}}_l^*\right)\left(-\mathcal{G}\boldsymbol{y}\right)\left(\boldsymbol{I}-\widehat{\boldsymbol{V}}_l\widehat{\boldsymbol{V}}_l^*\right)\\ &= \left(\boldsymbol{U}\boldsymbol{U}^*-\widehat{\boldsymbol{U}}_l\widehat{\boldsymbol{U}}_l^*\right)\left(\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right)\left(\boldsymbol{I}-\widehat{\boldsymbol{V}}_l\widehat{\boldsymbol{V}}_l^*\right)\\ &= \left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)\left(\mathcal{I}-\mathcal{P}_{\widehat{\boldsymbol{V}}_l}\right)\left(\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right).\end{aligned}$$

Therefore

$$\begin{aligned}I_7 &= 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\left(\mathcal{I}-\widehat{p}^{-1}\mathcal{P}_{\widehat{\Omega}_{l+1}}\right)\mathcal{G}^*\left(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l}\right)\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)\left(\mathcal{I}-\mathcal{P}_{\widehat{\boldsymbol{V}}_l}\right)\left(\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right)\right\|_F\\ &\le 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\mathcal{G}\left(\mathcal{I}-\widehat{p}^{-1}\mathcal{P}_{\widehat{\Omega}_{l+1}}\right)\mathcal{G}^*\left(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l}\right)\left(\mathcal{P}_{\boldsymbol{U}}-\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\right)\right\|\left\|\widehat{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right\|_F\\ &\le 16\kappa\sqrt{\frac{16000\mu_0 c_s r\log(n)}{81\widehat{m}}}\left\|\widetilde{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right\|_F\end{aligned}$$

with probability at least $1-n^2$, where the last inequality follows from Lem. 2.3.6 and the inequality (2.20a). Putting the bounds for I_5, I_6 and I_7 together gives

$$\|\widetilde{\boldsymbol{L}}_{l+1}-\mathcal{G}\boldsymbol{y}\|_F \le \left(\frac{1}{2}+326\kappa\sqrt{\frac{\mu_0 c_s r\log(n)}{\widehat{m}}}\right)\left\|\widetilde{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right\|_F \le \frac{5}{6}\left\|\widetilde{\boldsymbol{L}}_l-\mathcal{G}\boldsymbol{y}\right\|_F$$

with probability at least $1-2n^{-2}$ provided $\widehat{m}\ge C\mu_0 c_s\kappa^2 r\log(n)$ for a sufficiently large universal constant C. Clearly on the same event, (2.19) also holds for the $(l+1)$-th iteration.

Since $\widetilde{\boldsymbol{L}}_0=\mathcal{T}_r\left(\widehat{p}^{-1}\mathcal{H}\mathcal{P}_{\Omega_0}\left(\boldsymbol{x}\right)\right)$, (2.19) is valid for $l=0$ with probability at least $1-n^2$ provides

$$\widehat{m}\ge C\mu_0 c_s\kappa^6 r^2\log(n)$$

for some numerical constant $C > 0$. Taking the upper bound on the number of measurements completes the proof of Lem. 2.1.2 by noting $\mathcal{H}\boldsymbol{x} = \mathcal{G}\boldsymbol{y}$. □

Proof of Thm. 2.1.2. The third condition (2.9) in Thm. 2.3.1 can be satisfied with probability at least $1-(2L+1)n^{-2}$ if we take $L = \left\lceil 6 \log\left(\frac{\sqrt{n}\log(n)}{16\varepsilon_0}\right)\right\rceil$. So the theorem can be proved by combining this result together with Lems. 2.3.1 and 2.3.2. □

CHAPTER 3
PROJECTED GRADIENT DESCENT

3.1 Algorithm and Main Result

3.1.1 Algorithm

3.1.1.1 Which Objective Function?

Obviously, each observed entry of $\boldsymbol{x}$ corresponds to a revealed skew-diagonal of $\mathcal{H}\boldsymbol{x}$. With a slight abuse of notation, denote by Ω the subset of the revealed skew-diagonals of $\mathcal{H}\boldsymbol{x}$. Given a vector $\boldsymbol{z} \in \mathbb{C}^n$, a simple calculation shows

$$\begin{aligned}\langle \mathcal{P}_\Omega(\mathcal{H}\boldsymbol{z} - \mathcal{H}\boldsymbol{x}), \mathcal{H}\boldsymbol{z} - \mathcal{H}\boldsymbol{x}\rangle &= \sum_{a\in\Omega}\sum_{i+j=a} \left([\mathcal{H}\boldsymbol{z}]^{(i,j)} - [\mathcal{H}\boldsymbol{x}]^{(i,j)}\right)^2 \\ &= \sum_{a\in\Omega} w_a (z_a - x_a)^2 \\ &= \langle \mathcal{P}_\Omega\left(\mathcal{D}(\boldsymbol{z}-\boldsymbol{x})\right), \mathcal{D}(\boldsymbol{z}-\boldsymbol{x})\rangle\,,\end{aligned}$$

where w_a in the second line is the number of entries in the a-th skew-diagonal of an $n_1 \times n_2$ matrix, and $\mathcal{D}$ in the last line is a linear map which scales the a-th entry of a vector by a factor of $\sqrt{w_a}$ for all $a = 0, \cdots, n-1$. We have seen that $\mathcal{H}\boldsymbol{x}$ is a rank r matrix. Thus, to reconstruct $\boldsymbol{x}$, we may seek a signal $\boldsymbol{z}$ such that $\operatorname{rank}(\mathcal{H}\boldsymbol{z}) = r$ and $\mathcal{H}\boldsymbol{z}$ fits the revealed skew-diagonals of $\mathcal{H}\boldsymbol{x}$ as well as possible by solving a rank constraint *weighted least square* problem:

$$\min_{\boldsymbol{z}\in\mathbb{C}^n} \langle \mathcal{P}_\Omega\left(\mathcal{D}(\boldsymbol{z}-\boldsymbol{x})\right), \mathcal{D}(\boldsymbol{z}-\boldsymbol{x})\rangle \quad \text{subject to} \quad \operatorname{rank}(\mathcal{H}\boldsymbol{z}) = r. \tag{3.1}$$

For ease of exposition, we will make a change of variables and rewrite (3.1) using the new variable $\boldsymbol{y} = \mathcal{D}\boldsymbol{x}$. Denote by $\mathcal{H}^*$ the adjoint of $\mathcal{H}$, which maps a matrix $\boldsymbol{Z} \in \mathbb{C}^{n_1 \times n_2}$ to a vector $\mathcal{H}^*\boldsymbol{Z} = \left\{\sum_{i+j=a} Z^{(i,j)}\right\}_{a=0}^{n-1}$. It is easy to show that $\mathcal{H}^*\mathcal{H} = \mathcal{D}^2$. Letting $\mathcal{G} = \mathcal{H}\mathcal{D}^{-1}$, we find that $\mathcal{G}$ has the desirable orthogonal property $\mathcal{G}^*\mathcal{G} = \mathcal{I}$, where $\mathcal{I}$ denotes the identity operator. After the substitution of $\mathcal{D}\boldsymbol{x}$ by $\boldsymbol{y}$ and the substitution of $\mathcal{D}\boldsymbol{z}$ by $\boldsymbol{z}$, we can rewrite (3.1) as

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \langle \mathcal{P}_\Omega(\boldsymbol{z} - \boldsymbol{y}), \boldsymbol{z} - \boldsymbol{y} \rangle \quad \text{subject to} \quad \operatorname{rank}(\mathcal{G}\boldsymbol{z}) = r, \tag{3.2}$$

which will be our primary focus. A more direct interpretation of (3.2) is as follows. Since $\boldsymbol{y} = \mathcal{D}\boldsymbol{x}$, $\mathcal{P}_\Omega(\boldsymbol{y}) = \mathcal{P}_\Omega(\mathcal{D}\boldsymbol{x}) = \mathcal{D}\mathcal{P}_\Omega(\boldsymbol{x})$, $\operatorname{rank}(\mathcal{G}\boldsymbol{y}) = \operatorname{rank}(\mathcal{H}\boldsymbol{x}) = r$, and $\mathcal{D}$ is invertible, one can instead attempt to reconstruct $\boldsymbol{y}$ from $\mathcal{P}_\Omega(\boldsymbol{y})$ by seeking a signal that corresponds to a low rank Hankel matrix and fits the observations as well as possible.

In order to eliminate the rank constraint in (3.2), we parameterize $\mathcal{G}\boldsymbol{z}$ by a product of two rank r matrices and write $\mathcal{G}\boldsymbol{z}$ as $\mathcal{G}\boldsymbol{z} = \boldsymbol{Z}_U\boldsymbol{Z}_V^*$, where $\boldsymbol{Z}_U \in \mathbb{C}^{n_1 \times r}$ and $\boldsymbol{Z}_V \in \mathbb{C}^{n_2 \times r}$. We note that $\boldsymbol{Z}_U\boldsymbol{Z}_V^*$ is a Hankel matrix if and only if

$$(\mathcal{I} - \mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U\boldsymbol{Z}_V^*) = \boldsymbol{0}.$$

Thus, by further noting that $\boldsymbol{z} = \mathcal{G}^*(\mathcal{G}\boldsymbol{z}) = \mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*)$, we can rewrite (3.2) using

$\boldsymbol{Z}_U$ and $\boldsymbol{Z}_V$ as

$$\begin{aligned} &\min_{\boldsymbol{Z}_U, \boldsymbol{Z}_V} \left\langle \mathcal{P}_\Omega \left(\mathcal{G}^* \left(\boldsymbol{Z}_U \boldsymbol{Z}_V^* \right) - \boldsymbol{y} \right), \mathcal{G}^* \left(\boldsymbol{Z}_U \boldsymbol{Z}_V^* \right) - \boldsymbol{y} \right\rangle \\ &\text{subject to} \quad (\mathcal{I} - \mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U \boldsymbol{Z}_V^*) = \boldsymbol{0}, \end{aligned} \tag{3.3}$$

which is an equality constraint minimization problem. Alternatively, (3.3) can be interpreted as follows: we estimate the rank r matrix $\mathcal{G}\boldsymbol{y}$ by a Hankel matrix of the form $\boldsymbol{Z}_U \boldsymbol{Z}_V^*$ that minimizes the mismatch in the measurement domain. Once $\mathcal{G}\boldsymbol{y}$ is reconstructed, one can recover $\boldsymbol{y}$ via $\boldsymbol{y} = \mathcal{G}^*(\mathcal{G}\boldsymbol{y})$.

Putting the constraint and the objective function in (3.3) together allows us to consider an optimization problem without the equality constraint by minimizing

$$f(\boldsymbol{Z}) = \left\| (\mathcal{I} - \mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U \boldsymbol{Z}_V^*) \right\|_F^2 + p^{-1} \left\langle \mathcal{P}_\Omega(\mathcal{G}^*(\boldsymbol{Z}_U \boldsymbol{Z}_V^*) - \boldsymbol{y}), \mathcal{G}^*(\boldsymbol{Z}_U \boldsymbol{Z}_V^*) - \boldsymbol{y} \right\rangle,$$

where

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_U \\ \\ \boldsymbol{Z}_V \end{bmatrix} \in \mathbb{C}^{(n+1)\times r}$$

denotes the concatenation of $\boldsymbol{Z}_U$ and $\boldsymbol{Z}_V$, and the weight $p = m/n$ is the sampling ratio. Let $\mathcal{G}\boldsymbol{y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the reduced singular value decomposition (SVD) of $\mathcal{G}\boldsymbol{y}$. Define

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_U \\ \\ \boldsymbol{M}_V \end{bmatrix} \in \mathbb{C}^{(n+1)\times r}, \tag{3.4}$$

where $\boldsymbol{M}_U = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{M}_V = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$. It is easily shown that $f(\boldsymbol{Z}) = 0$ and thus achieves its minimum for the set of matrices

$$\left\{ \begin{bmatrix} \boldsymbol{M}_U\boldsymbol{X} \\ \boldsymbol{M}_V(\boldsymbol{X}^{-1})^* \end{bmatrix}, \ \boldsymbol{X} \in \mathbb{C}^{r\times r} \text{ is invertible} \right\}. \tag{3.5}$$

Note that (3.5) is also a set of solutions for the equality constrained problem (3.3). Among this set of solutions, there are ones which are highly unbalanced, i.e., these having $\|\boldsymbol{Z}_U\|_F \to 0$ and $\|\boldsymbol{Z}_V\|_F \to \infty$, or vice versa. For example, let $\boldsymbol{Z}_U = \alpha\boldsymbol{M}_U$ and $\boldsymbol{Z}_V = \alpha^{-1}\boldsymbol{M}_V$ for α being a real number that approaches either zero or infinity. Those solutions are unfavorable for the purpose of both computation and analysis. In order to reduce the solution space and avoid the occurrence of the pathological solutions, we add the regularizer function

$$g(\boldsymbol{Z}) = \frac{1}{2}\left\|\boldsymbol{Z}_U^*\boldsymbol{Z}_U - \boldsymbol{Z}_V^*\boldsymbol{Z}_V\right\|_F^2$$

to $f(\boldsymbol{Z})$ and instead consider the minimization problem with respect to

$$F(\boldsymbol{Z}) = f(\boldsymbol{Z}) + \lambda \cdot g(\boldsymbol{Z}), \tag{3.6}$$

where $\lambda > 0$ is to be determined. Here, $g(\boldsymbol{Z})$ in some sense penalizes the mismatch between the sizes of $\boldsymbol{Z}_U$ and $\boldsymbol{Z}_V$, and it was also used in rectangular low rank matrix recovery, see [53, 62].

Now, the set of solutions that minimizes $F(\boldsymbol{Z})$ or at which $F(\boldsymbol{Z}) = 0$ is given

by

$$\mathcal{O} = \left\{ \begin{bmatrix} \boldsymbol{M}_U \boldsymbol{Q} \\ \boldsymbol{M}_V \boldsymbol{Q} \end{bmatrix}, \ \boldsymbol{Q} \in \mathbb{C}^{r \times r} \text{ is unitary} \right\}. \tag{3.7}$$

The distance of a matrix $\boldsymbol{Z} \in \mathbb{C}^{(n+1)\times r}$ to the solution set, denoted $\text{dist}(\boldsymbol{Z}, \boldsymbol{M})$, is defined as

$$\text{dist}(\boldsymbol{Z}, \boldsymbol{M}) = \min_{\boldsymbol{Q}\boldsymbol{Q}^* = \boldsymbol{Q}^*\boldsymbol{Q} = \boldsymbol{I}} \|\boldsymbol{Z} - \boldsymbol{M}\boldsymbol{Q}\|_F .$$

Let $\boldsymbol{M}^*\boldsymbol{Z} = \boldsymbol{Q}_1 \boldsymbol{\Lambda} \boldsymbol{Q}_2^*$ be the SVD of $\boldsymbol{M}^*\boldsymbol{Z}$. By the Von Neumann's trace inequality [40], the above minimum is achieved at the unitary matrix $\boldsymbol{Q}_{\boldsymbol{Z}}$ given by

$$\boldsymbol{Q}_{\boldsymbol{Z}} = \boldsymbol{Q}_1 \boldsymbol{Q}_2^*. \tag{3.8}$$

3.1.1.2 Which Feasible Set?

Let μ and σ be two numerical constants such that $\mu \geq \mu_0$ and $\sigma \geq \sigma_1(\mathcal{G}\boldsymbol{y})$. When $\mathcal{G}\boldsymbol{y}$ is μ_0-incoherent, the matrix $\boldsymbol{M}$ constructed in (3.4) satisfies $\|\boldsymbol{M}\|_{2,\infty} \leq \sqrt{\mu c_s r \sigma / n}$. Moreover, letting $\mathcal{C}$ be a convex set defined as

$$\mathcal{C} = \left\{ \boldsymbol{Z} \in \mathbb{C}^{(n+1)\times r} \mid \|\boldsymbol{Z}\|_{2,\infty} \leq \sqrt{\frac{\mu c_s r \sigma}{n}} \right\}, \tag{3.9}$$

it is evident that $\mathcal{O} \subset \mathcal{C}$. Therefore, we can restrict our search on the feasible set $\mathcal{C}$ when computing the minimum or zero value of $F(\boldsymbol{Z})$.

3.1.1.3 Algorithm

The discussion above tells us that we can reconstruct the low rank factors $\boldsymbol{M}_U$ and $\boldsymbol{M}_V$ of the ground truth matrix $\mathcal{G}\boldsymbol{y}$ by minimizing the function $F(\boldsymbol{Z})$ on the feasible set $\mathcal{C}$, namely

$$\min_{\boldsymbol{Z}\in\mathcal{C}} F(\boldsymbol{Z}), \tag{3.10}$$

where $F(\boldsymbol{Z})$ is defined in (3.6) and $\mathcal{C}$ is defined in (3.9). We present a simple projected gradient descent algorithm for this problem, see Alg. 3.1. The algorithm consists

Algorithm 3.1 Projected Gradient Descent (PGD)

Initialize $\boldsymbol{L}^0 = p^{-1}\mathcal{T}_r(\mathcal{G}\mathcal{P}_\Omega(\boldsymbol{y})) = \boldsymbol{U}^0\boldsymbol{\Sigma}^0(\boldsymbol{V}^0)^*$, $\tilde{\boldsymbol{Z}}^0 = \begin{bmatrix}\boldsymbol{U}^0(\boldsymbol{\Sigma}^0)^{\frac{1}{2}} \\ \boldsymbol{V}^0(\boldsymbol{\Sigma}^0)^{\frac{1}{2}}\end{bmatrix}$, $\boldsymbol{Z}^0 = \mathcal{P}_{\mathcal{C}}(\tilde{\boldsymbol{Z}}^0)$.

for $k = 0, 1, \cdots$ **do**

1. $\tilde{\boldsymbol{Z}}^{k+1} = \boldsymbol{Z}^k - \eta\nabla F(\boldsymbol{Z}^k)$
2. $\boldsymbol{Z}^{k+1} = \mathcal{P}_{\mathcal{C}}(\tilde{\boldsymbol{Z}}^{k+1})$

end for

Output: $\boldsymbol{Z}^k$ in the last iteration, $\boldsymbol{y}^k = \mathcal{G}^*(\boldsymbol{Z}_U^k(\boldsymbol{Z}_V^k)^*)$ and $\boldsymbol{x}^k = \mathcal{D}^{-1}\boldsymbol{y}^k$.

of two phases: Initialization and gradient descent with a constant stepsize. The initial guess is computed via one-step hard thresholding, followed by projection onto the convex set $\mathcal{C}$. The hard thresholding operator $\mathcal{T}_r(\cdot)$ returns the best rank r approximation of a matrix, which can be computed via the partial SVD. Given a

matrix $\boldsymbol{Z} \in \mathbb{C}^{(n+1)\times r}$, the projection $\mathcal{P}_{\mathcal{C}}(\boldsymbol{Z})$ can be computed by row-wise trimming,

$$[\mathcal{P}_{\mathcal{C}}(\boldsymbol{Z})]^{(i,:)} = \begin{cases} \boldsymbol{Z}^{(i,:)} & \text{if } \left\|\boldsymbol{Z}^{(i,:)}\right\|_2 \leq \sqrt{\frac{\mu c_s r \sigma}{n}}, \\ \frac{\boldsymbol{Z}^{(i,:)}}{\left\|\boldsymbol{Z}^{(i,:)}\right\|_2}\sqrt{\frac{\mu c_s r \sigma}{n}} & \text{otherwise.} \end{cases}$$

In each iteration of the algorithm, the current estimate $\boldsymbol{Z}^k$ is updated along the negative gradient descent direction $-\nabla F(\boldsymbol{Z}^k)$, using a stepsize η, followed by projection onto the convex set $\mathcal{C}$. Since we are working with complex matrices, the gradient $F(\boldsymbol{Z})$ of a matrix $\boldsymbol{Z}$ is calculated under the Wirtinger calculus, given by

$$\nabla F(\boldsymbol{Z}) = \begin{bmatrix} \nabla F_{\boldsymbol{U}}(\boldsymbol{Z}) \\ \nabla F_{\boldsymbol{V}}(\boldsymbol{Z}) \end{bmatrix} = \begin{bmatrix} \nabla f_{\boldsymbol{U}}(\boldsymbol{Z}) + \lambda \cdot \nabla g_{\boldsymbol{U}}(\boldsymbol{Z}) \\ \nabla f_{\boldsymbol{V}}(\boldsymbol{Z}) + \lambda \cdot \nabla g_{\boldsymbol{V}}(\boldsymbol{Z}) \end{bmatrix},$$

where

$$\nabla f_{\boldsymbol{U}}(\boldsymbol{Z}) = \left((\mathcal{I} - \mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*)\right)\boldsymbol{Z}_{\boldsymbol{V}} + p^{-1}\left(\mathcal{G}\mathcal{P}_{\Omega}(\mathcal{G}^*(\boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*) - \boldsymbol{y})\right)\boldsymbol{Z}_{\boldsymbol{V}},$$

$$\nabla f_{\boldsymbol{V}}(\boldsymbol{Z}) = \left((\mathcal{I} - \mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*)\right)^*\boldsymbol{Z}_{\boldsymbol{U}} + p^{-1}\left(\mathcal{G}\mathcal{P}_{\Omega}(\mathcal{G}^*(\boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*) - \boldsymbol{y})\right)^*\boldsymbol{Z}_{\boldsymbol{U}},$$

$$\nabla g_{\boldsymbol{U}}(\boldsymbol{Z}) = \boldsymbol{Z}_{\boldsymbol{U}}(\boldsymbol{Z}_{\boldsymbol{U}}^*\boldsymbol{Z}_{\boldsymbol{U}} - \boldsymbol{Z}_{\boldsymbol{V}}^*\boldsymbol{Z}_{\boldsymbol{V}}),$$

$$\nabla g_{\boldsymbol{V}}(\boldsymbol{Z}) = \boldsymbol{Z}_{\boldsymbol{V}}(\boldsymbol{Z}_{\boldsymbol{V}}^*\boldsymbol{Z}_{\boldsymbol{V}} - \boldsymbol{Z}_{\boldsymbol{U}}^*\boldsymbol{Z}_{\boldsymbol{U}}).$$

PGD can be implemented very efficiently and the main computational cost per iteration is $O(r^2 n + rn\log(n))$ flops, which lies in the computation of $\nabla F(\boldsymbol{Z})$ in

each iteration. Taking the computation of $\nabla F_U(\boldsymbol{Z})$ as an example, we note that

$$\begin{aligned}&\nabla F_U(\boldsymbol{Z})\\ =&\mathcal{G}\left(p^{-1}\mathcal{P}_\Omega(\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*)-\boldsymbol{y})-\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*)\right)\boldsymbol{Z}_V+\boldsymbol{Z}_U\left(\lambda\boldsymbol{Z}_U^*\boldsymbol{Z}_U+(1-\lambda)\boldsymbol{Z}_V^*\boldsymbol{Z}_V\right).\end{aligned}$$

Clearly, the second term can be computed using $O(r^2n)$ flops. Let

$$\boldsymbol{w}=p^{-1}\mathcal{P}_\Omega(\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*)-\boldsymbol{y})-\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*).$$

Since we can compute $\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*)$ by r fast convolutions, $\boldsymbol{w}$ can be obtained using $O(rn\log(n))$ flops. Moreover, $(\mathcal{G}\boldsymbol{w})\boldsymbol{Z}_V$ can be computed via r fast Hankel matrix-vector multiplications that also cost $O(rn\log(n))$ flops.

Before proceeding, it is worth noting that non-convex (projected) gradient decent methods have received intensive investigations for other low rank matrix recovery problems, such as unstructured low rank matrix recovery and matrix completion [53, 61, 62], phase retrieval [13, 17], robust principle component analysis [16, 59], and blind deconvolution [35]. In those papers, lower bounds on the sampling complexity have been established under different random measurement models, showing that the number of measurements needed for the successful recovery of the target matrices is essentially determined by the number of degrees of freedom in the matrices. In particular, a projected gradient descent algorithm was studied in [62] for unstructured rectangular low rank matrix completion. The convergence analysis of PGD is directly inspired by [62], though the technical details are substantially different.

3.1.2 Main Result

Let $\Omega = \{a_k \mid k = 1, \cdots, m\}$. We consider the sampling with replacement model, where each index a_k is drawn independently and uniformly from $\{0, \cdots, n-1\}$. Under this sampling model, for a vector $\boldsymbol{z} \in \mathbb{C}^n$, the projection $\mathcal{P}_\Omega(\boldsymbol{z})$ is given by

$$\mathcal{P}_\Omega(\boldsymbol{z}) = \sum_{k=1}^{m} z_{a_k} \boldsymbol{e}_{a_k}, \tag{3.11}$$

and for two vectors $\boldsymbol{z},\ \boldsymbol{w} \in \mathbb{C}^n$, the inner product $\langle \mathcal{P}_\Omega(\boldsymbol{z}), \boldsymbol{w} \rangle$ is given by

$$\langle \mathcal{P}_\Omega(\boldsymbol{z}), \boldsymbol{w} \rangle = \sum_{k=1}^{m} \bar{z}_{a_k} w_{a_k}. \tag{3.12}$$

In the guarantee analysis of PGD, we assume μ and σ in (3.9) are two tuning parameters obeying $\mu \geq \mu_0$ and $\sigma \geq \sigma_1(\mathcal{G}\boldsymbol{y})$ so that $\boldsymbol{M} \in \mathcal{C}$. For conciseness, we take $\sigma = \sigma_1(\boldsymbol{L}_0)/(1-\varepsilon_0)$ for some $0 < \varepsilon_0 < 1$ and will later show that $\sigma \geq \sigma_1(\mathcal{G}\boldsymbol{y})$ with high probability.

Theorem 3.1.1 (Exact Recovery)**.** *Assume $\mathcal{G}\boldsymbol{y}$ is μ_0-incoherent. Let ε_0 be a absolute constant obeying $0 < \varepsilon_0 \leq 1/11$. Let $\mu \geq \mu_0$ and $\sigma = \sigma_1(\boldsymbol{L}_0)/(1-\varepsilon_0)$. If we take $\lambda = 1/4$ in (3.6), then with probability at least $1 - c_1 \cdot n^{-2}$, the sequence $\left\{\boldsymbol{Z}^k\right\}_{k \geq 1}$ returned by Alg. 3.1 obeys*

$$\mathrm{dist}^2(\boldsymbol{Z}^k, \boldsymbol{M}) \leq (1 - \eta\nu)^k \mathrm{dist}^2(\boldsymbol{Z}^0, \boldsymbol{M})$$

for

$$\eta \leq \frac{\sigma_r(\mathcal{G}\boldsymbol{y})}{600(\mu c_s r)^2 \sigma_1^2(\mathcal{G}\boldsymbol{y})} \quad \textit{and} \quad \nu = \frac{1}{10}\sigma_r(\mathcal{G}\boldsymbol{y})$$

provided $m \geq c_2 \varepsilon_0^{-2} \mu^2 c_s^2 \kappa^2 r^2 \log(n)$, *where* $\kappa = \sigma_1(\mathcal{G}\boldsymbol{y})/\sigma_r(\mathcal{G}\boldsymbol{y})$.

Remark. 1). After an approximation of $\mathcal{G}\boldsymbol{y}$, given by $\boldsymbol{Z}_U^k(\boldsymbol{Z}_V^k)^*$, is obtained from PGD, we can estimate $\boldsymbol{y}$ by $\boldsymbol{y}^k = \mathcal{G}^*(\boldsymbol{Z}_U^k(\boldsymbol{Z}_V^k)^*)$, and in turn estimate $\boldsymbol{x}$ by $\mathcal{D}^{-1}\boldsymbol{y}^k$. Let $\boldsymbol{Q}_{\boldsymbol{Z}^k}$ be the unitary matrix that satisfies $\operatorname{dist}(\boldsymbol{Z}^k, \boldsymbol{M}) = \left\|\boldsymbol{Z}^k - \boldsymbol{M}\boldsymbol{Q}_{\boldsymbol{Z}^k}\right\|_F$. A simple calculation yields

$$\begin{aligned}
\left\|\boldsymbol{x}^k - \boldsymbol{x}\right\|_2 \leq \left\|\boldsymbol{y}^k - \boldsymbol{y}\right\|_2 &= \left\|\mathcal{G}^*(\boldsymbol{Z}_U^k(\boldsymbol{Z}_V^k)^*) - \mathcal{G}^*(\mathcal{G}\boldsymbol{y})\right\|_2 \\
&\leq \left\|\boldsymbol{Z}_U^k(\boldsymbol{Z}_V^k)^* - \boldsymbol{M}_U\boldsymbol{M}_V^*\right\|_F \leq \frac{1}{\sqrt{2}}\left\|\boldsymbol{Z}^k(\boldsymbol{Z}^k)^* - \boldsymbol{M}\boldsymbol{M}^*\right\|_F \\
&= \frac{1}{\sqrt{2}}\left\|\boldsymbol{Z}^k(\boldsymbol{Z}^k - \boldsymbol{M}\boldsymbol{Q}_{\boldsymbol{Z}^k})^* + (\boldsymbol{Z}^k - \boldsymbol{M}\boldsymbol{Q}_{\boldsymbol{Z}^k})(\boldsymbol{M}\boldsymbol{Q}_{\boldsymbol{Z}^k})^*\right\|_F \\
&\leq \frac{1}{\sqrt{2}}\left(\|\boldsymbol{Z}^k\|_2 + \|\boldsymbol{M}\|_2\right)\operatorname{dist}(\boldsymbol{Z}^k, \boldsymbol{M}) \to 0, \quad \text{as } \operatorname{dist}(\boldsymbol{Z}^k, \boldsymbol{M}) \to 0.
\end{aligned}$$

2). After each iteration, Thm. 3.1.1 implies that the distance between the estimate given by PGD and $\boldsymbol{M}$ is reduced by at least of a factor of $1 - O(1/(\mu c_s r\kappa)^2)$. Thus, after $k \approx O((\mu c_s r\kappa)^2 \log(1/\epsilon))$ iterations, $\operatorname{dist}^2(\boldsymbol{Z}^k, \boldsymbol{M}) \leq \epsilon \cdot \operatorname{dist}^2(\boldsymbol{Z}^0, \boldsymbol{M})$.

3). FIHT can achieve exact recovery when the number of revealed entries is of order $O(\kappa^6 r^2 \log^2(n))$. In contrast, the sampling complexity of PGD is only a quadratic function of κ and a linear function of $\log(n)$. Moreover, the exact recovery guarantee of FIHT relies on a more complicated initialization scheme which

requires a partition of the observed entries into $O(\log(n))$ groups, while the initial guess constructed for the exact recovery guarantee of PGD can be computed much more easily.

3.2 Numerical Experiments

In this section, we conduct numerical experiments to evaluate the performance of PGD[1]. The experiments are executed from MATLAB R2017a on a 64-bit Linux machine with multi-core Intel Xeon CPU E5-2667 v3 at 3.20GHz and 64GB of RAM. In Sec. 3.2.1, we investigate the largest number of Fourier components that can be successfully recovered by PGD. The tests are conducted on one-dimensional signals in large part due to the high computational cost of this type of simulations. Then we evaluate PGD against computational efficiency, robustness to additive noise, and sensitivity to mis-specification of model order on three-dimensional signals in Secs. 3.2.2, 3.2.3, and 3.2.4, respectively. The initial guess of PGD is computed using the PROPACK package [34], and the parameters μ and σ used in the projection are estimated from the initialization. Instead of using the constant stepsize suggested in the main result which appears to be conservative, we choose the stepsize via a backtracking line search in the implementation.

[1]In our random simulations, we didn't find much difference between the performance of PGD and the performance of the gradient descent algorithm applied to $f(\boldsymbol{Z})$ directly. However, since the extra cost incurred by computing the gradient of $g(\boldsymbol{Z})$ and the projection $\mathcal{P}_{\mathcal{C}}(\boldsymbol{Z})$ is marginal, it is appealing to run PGD for its recovery guarantee.

3.2.1 Empirical Phase Transition

We evaluate the recovery ability of PGD in the framework of phase transition and compare it with ANM [49], EMaC [15] and FIHT [8]. ANM and EMaC are implemented using CVX [29] with default parameters. The test spectrally sparse signals of length n with r frequency components are formed in the following way: each frequency f_k is randomly generated from $[0, 1)$, and the argument of each complex coefficient d_k is uniformly sampled from $[0, 2\pi)$ while the amplitude is selected to be $1 + 10^{0.5c_k}$ with c_k being uniformly distributed on $[0, 1]$. We test two different settings for the frequencies: a) no separation condition is imposed on $\{f_k\}_{k=1}^r$, and b) the wrap-around distances between each pair of the randomly drawn frequencies are guaranteed to be greater than $1.5/n$. After a signal is formed, m of its entries are sampled uniformly at random. For a given triple (n, r, m), 50 random tests are conducted. We consider an algorithm to have successfully reconstructed a test signal if the root mean squared error (RMSE) is less than 10^{-3},

$$\|\boldsymbol{x}_{rec} - \boldsymbol{x}\|_2 / \|\boldsymbol{x}\| \leq 10^{-3}.$$

The tests are conducted with $n = 127$ and $p = m/n$ taking 18 equispaced values from 0.1 to 0.95. For a fixed pair of (n, m), we start with $r = 1$ and then increase the value of r by one until it reaches a value such that the tested algorithm fails all the 50 random tests. FIHT is terminated when $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 / \|\boldsymbol{x}^k\|_2 \leq 10^{-6}$ or a maximum number of iteration is reached. PGD is terminated when one of the following three

conditions is met: $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 / \|\boldsymbol{x}^k\|_2 \leq 10^{-7}$, $|F(\tilde{\boldsymbol{Z}}^{k+1}) - F(\boldsymbol{Z}^k)|/F(\boldsymbol{Z}^k) \leq 10^{-5}$, or a maximum number of iteration is reached.

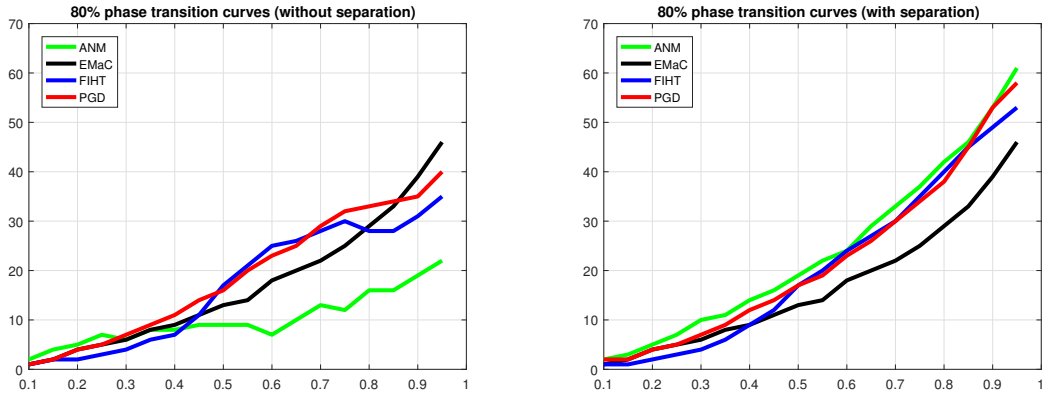

Figure 3.1: 80% phase transition curves: x-axis is $p = m/n$ and y-axis is r. Left: signals are formed by random frequencies without separation; Right: signals are formed by random frequencies separated by at least $1.5/n$.

We plot in Fig. 3.1 the empirical recovery phase transition curves that identify the 80% success rate for each tested algorithm under the two different frequency settings. When the frequencies are separated by at least $1.5/n$, the right plot shows that ANM has the highest phase transition curve, and the phase transition curve of PGD closely tracks that of ANM. The performance of ANM degrades severely when there is no frequency separation requirement. In both of the frequency settings, the recovery phase transition curves of PGD are overall higher than that of EMaC. In the region of greatest interest where $p \leq 0.5$, the recovery phase transition curves of PGD are substantially higher than that of FIHT.

Provided that there are about $4/n$ separations between the frequencies in the

undamped case, ANM [49] is guaranteed to achieve successful recovery with high probability. The separation condition is not only sufficient but also necessary for ANM to work, as shown recently by [58]. EMaC [15] minimizes the nuclear norm of the Hankel matrix with partial revealed anti-diagonals. However, [58] shows that the frequencies can be arbitrarily close for EMaC to have successful recovery. These two findings by [58] are also reflected in Fig. 3.1.

FIHT and PGD are guaranteed to work if the incoherence assumption is satisfied. We only know this assumption holds in the undamped case if the frequencies are sufficiently separated. From Fig. 3.1, these two methods seem to depend on separations, but less than ANM, to achieve successful recoveries. It could be an area of future research to derive some theories when the frequencies are less separated and/or there is damping in the signal.

3.2.2 Computational Efficiency

PGD has the same leading-order computational complexity as FIHT, and both of them are able to handle large and high-dimensional signals. We compare the computational performance of these two algorithms on undamped and damped three-dimensional spectrally sparse signals of size $n = 64 \times 128 \times 512$. Tests are conducted with $r \in \{20, 30\}$ and $m \approx 130 \log(n)$ in the undamped setting while $m \approx 0.03n$ in the damped setting, and we test signals which obey the frequency separation condition as well as signals which are fully random. As to the damping factors, for $1 \leq k \leq r$, $1/\tau_{1k}$ is uniformly sampled from [8 16], $1/\tau_{2k}$ is uni-

formly sampled from [16 32], and $1/\tau_{3k}$ is uniformly sampled from [64 128]. For each triple of (r, undamped/damped, with/without separation), 10 random problem instances are tested. FIHT is terminated when $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2/\|\boldsymbol{x}^k\|_2 \leq 10^{-3}$ or $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2/\|\boldsymbol{x}^k\|_2 \geq 2$ which usually implies divergence. PGD is terminated when $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2/\|\boldsymbol{x}^k\|_2 \leq 2 \times 10^{-4}$. The average computational time (referred to as TIME) and average number of iterations (referred to as ITER) of FIHT and PGD over tests of successful recovery are summarized in Tabs. 3.1 and 3.2 for the undamped and damped signals, respectively. For the sake of completeness, we also include the ratio of successful recovery out of the 10 random tests (referred to as SR) for each algorithm in the tables.

Table 3.1: Average SR, RMSE, ITER and TIME values of FIHT and PGD over 10 random problem instances in the undamped case.

r	20				30			
	SR	RMSE	ITER	TIME (s)	SR	RMSE	ITER	TIME (s)
	with separation							
FIHT	1	3.6e-4	18.7	256	0.5	4.8e-4	123	2278
PGD	1	1.4e-4	33.6	490	1	2.7e-4	48.3	1049
	without separation							
FIHT	1	3.5e-4	18.6	250	0.2	4.8e-4	66.5	1275
PGD	1	1.7e-4	33.6	492	1	3.0e-4	54.6	1186

First it is worth noting that PGD succeeded in all the 10 random tests under each test setting when $r = 30$, whereas FIHT only succeeded in a small fraction of the tests. Thus, Tabs. 3.1 and 3.2 show that PGD is able to more reliably recover

Table 3.2: Average SR, RMSE, ITER and TIME values of FIHT and PGD over 10 random problem instances in the damped case.

r	20				30			
	SR	RMSE	ITER	TIME (s)	SR	RMSE	ITER	TIME (s)
	with separation							
FIHT	1	2.9e-4	12.7	170	0.2	3.2e-4	16.5	321
PGD	1	3.3e-4	21.8	321	1	4.8e-4	41.5	1028
	without separation							
FIHT	1	2.4e-4	10.9	152	0.1	4.1e-4	16	325
PGD	1	2.6e-4	17.4	258	1	4.5e-4	37.4	863

signals that consist of a larger number of Fourier components, which coincides with our observations on one-dimensional signals in Sec. 3.2.1. The tables also show that FIHT requires fewer number of iterations and less computational time than PGD to achieve convergence for easier problem instances when $r = 20$, while PGD is faster when $r = 30$ and the test signals are undamped.

3.2.3 Robustness to Additive Noise

We demonstrate the performance of PGD under additive noise by conducting tests on 3D signals of the same size as in Sec. 3.2.2 but with measurements corrupted by the vector

$$\boldsymbol{e} = \theta \cdot \|\mathcal{P}_\Omega(\boldsymbol{x})\|_2 \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2},$$

where $\boldsymbol{x}$ is a reshaped three-dimensional spectrally sparse signal to be reconstructed, the entries of $\boldsymbol{w}$ are i.i.d. standard complex Gaussian random variables, and θ is referred to as the noise level.

Tests are conducted with 7 different values of θ from 10^{-3} to 1, corresponding

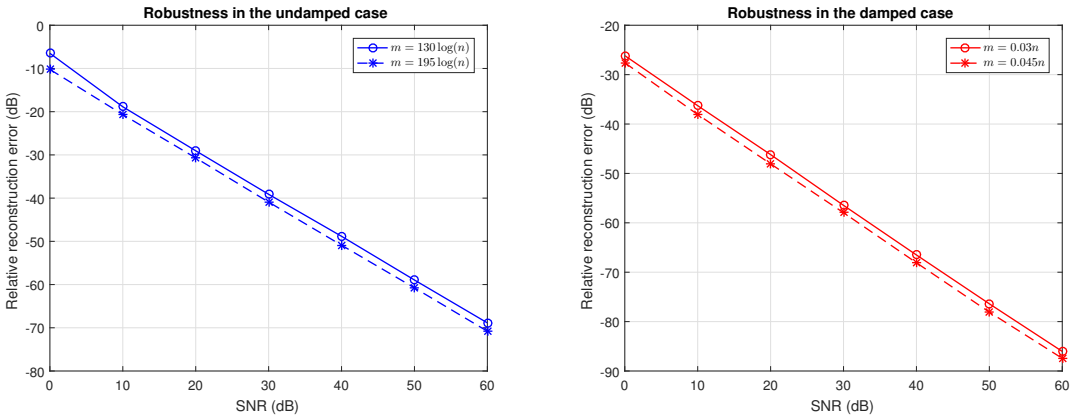

Figure 3.2: Performance of PGD under additive noise. Left: no damping in the test signals; Right: signals are generated with damping.

to 7 equispaced signal-to-noise ratios (SNR) from 60 to 0 dB. For each value of θ, 10 random instances are tested. PGD is terminated when $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 / \|\boldsymbol{x}^k\|_2 \leq 10^{-5}$. In our simulations, we fix $r = 20$ and choose $m \in \{130\log(n), 195\log(n)\}$ in the undamped setting while $m \in \{0.03n, 0.045n\}$ in the damped setting. The frequencies of the test signals are randomly generated from $[0, 1)$ without the separation requirement and the damping factors are generated in the same fashion as in Sec. 3.2.2. The average RMSE of the reconstructed signals (measured in negative dB) plotted against the input SNR values of the samples is presented in Fig. 3.2. The plots display a desirable linear scaling between the relative reconstruction error and the noise level for both the undamped and damped signals. Moreover, the relative reconstruction error decreases linearly on a log-log scale as the number of measurements increases.

3.2.4 Sensitivity to Model Order

In practice, we may not know the exact model order of a spectrally sparse signal but only have an estimation of it. Thus, it is of great interest to examine the performance of PGD when the model order is under- or over- estimated. The experiments are conducted for three-dimensional signals of the same size as in Sec. 3.2.2. Here the true model order is $r = 20$, and we observe $m = 130\log(n)$ entries for undamped signals while $m = 0.03n$ entries for damped signals. The frequencies are generated randomly and the damping factors are generated in the same way as in Sec. 3.2.2. Three noise levels are investigated: SNR= ∞ (noise-free), SNR= 20 (light noise) and SNR= 0 (heavy noise), and tests are conducted under the same additive noise model as in Sec. 3.2.3. For a fixed noise level, we test PGD starting from $r = 5$ and then increase the value of r by 5 each time until the maximum value 40 is reached. For each pair of (SNR, r), 10 random problem instances are tested, and PGD is terminated when $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2/\|\boldsymbol{x}^k\|_2 \leq 10^{-5}$. The median values of ITER and SNR when convergence is attained are reported in Tabs. 3.3 and 3.4 for undamped and damped signals, respectively.

As expected, PGD achieves the best SNR when the input value of r is equal to 20 (the true model order). The SNR of the estimation is usually very low when r is smaller than 20 due to the systematic truncation error. On the other hand, even when r is twice as large as the true model order, the SNR of the estimation is still desirable though it requires dramatically more number of iterations for PGD to converge.

Next, we suggest a rank increasing heuristic for PGD when the underlying

Table 3.3: Median values of ITER and SNR over 10 random problem instances with $5 \leq r \leq 40$ and SNR $\in \{\infty, 20, 0\}$ for undamped signals. The true model order is $r = 20$.

Test Rank	5	10	15	20	25	30	35	40
	SNR= ∞							
ITER	18.5	18.5	23.5	45	798	1047	1209	1343
SNR	2.093	4.844	8.293	99.63	69.43	67.08	65.00	63.72
	SNR= 20							
ITER	17.5	23	28.5	40.5	1524	1969	1964	2514
SNR	2.040	4.848	8.277	29.05	26.70	25.58	24.55	23.75
	SNR= 0							
ITER	19.5	25	218.5	427.5	589	569.5	638	787.5
SNR	1.812	3.952	5.807	6.407	5.464	4.438	3.773	3.234

Table 3.4: Median values of ITER and SNR over 10 random problem instances with $5 \leq r \leq 40$ and SNR $\in \{\infty, 20, 0\}$ for damped signals. The true model order is $r = 20$.

Test Rank	5	10	15	20	25	30	35	40
	SNR= ∞							
ITER	43.5	40.5	48.5	24	679.5	942.5	1014	1130
SNR	2.224	4.873	9.000	96.62	64.54	61.20	59.57	59.00
	SNR= 20							
ITER	46	40.5	52.5	26	3852	4213	6048	5608
SNR	2.223	4.872	8.999	46.23	44.55	43.36	42.46	41.65
	SNR= 0							
ITER	57.5	74	52.5	36.5	2025	1566	2431	3281
SNR	2.217	4.857	8.904	26.26	24.40	23.18	22.29	21.52

model order is not known a priori. Starting from a sufficiently small r, we run PGD until convergence is reached (i.e., when $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 / \|\boldsymbol{x}^k\|_2 \leq 10^{-5}$). Then we compute and compare the relative residuals over the observed entries for the two successive testing values of r. If the relative residual is improved significantly, we increase the value of r; otherwise the algorithm is terminated. To validate the potential effectiveness of this heuristic, we test PGD for problem instances with SNR=

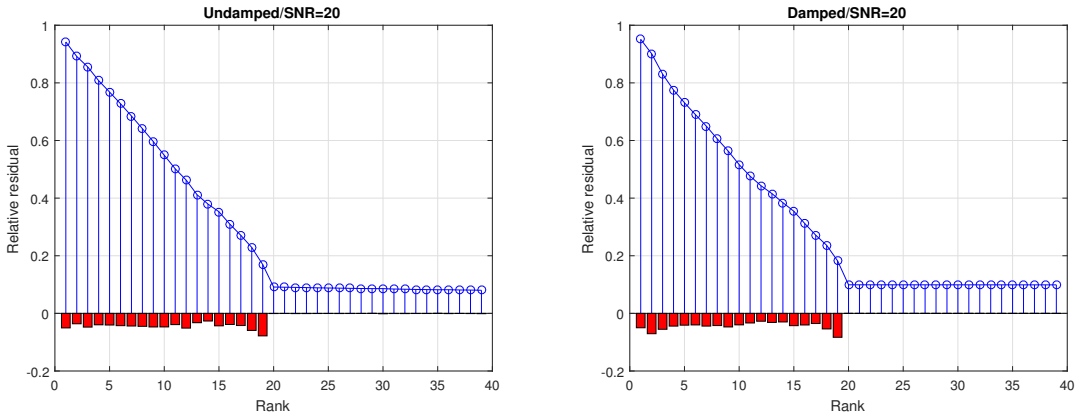

Figure 3.3: Demonstration of rank increasing heuristic for problem instances with SNR = 20 for undamped (Left) and damped (Right) signals.

20 for both undamped and damped signals, and with the values of r increasing from 1 to 40. The computational results are presented in Fig. 3.3, where we show the relative residual plotted against the values of r, as well as the change of the relative residual when r is increased by one. The figure shows that when r is greater than 20, the improvement of the relative residuals becomes very marginal for both undamped and damped signals.

3.3 Proof of Theorem 3.1.1

The structure of the proof for Thm. 3.1.1 follows the typical two-step strategy in the convergence analysis of non-convex optimization algorithms: a *basin of attraction* is firstly established, in which the algorithm converges linearly to the true solution; and then it can be shown that the initial guess constructed in the algorithm lies inside the basin of attraction. We begin our presentation of the proof with a proposition about the initialization.

Proposition 3.3.1 (Initialization Error)**.** *Suppose $\mathcal{G}\boldsymbol{y}$ is μ_0-incoherent. If $m \geq c\varepsilon_0^{-2}\mu c_s\kappa^2 r^2\log(n)$, then one has $\boldsymbol{M} \in \mathcal{C}$ and*

$$\operatorname{dist}^2(\boldsymbol{Z}^0, \boldsymbol{M}) \leq 3\varepsilon_0^2\sigma_r(\mathcal{G}\boldsymbol{y}) \tag{3.13}$$

with probability at least $1 - n^{-2}$.

Proof. By Lem. 2.1.1, one has

$$\|\boldsymbol{L}_0 - \mathcal{G}\boldsymbol{y}\|_2 \lesssim \sqrt{\frac{\mu_0 c_s r\log(n)}{m}}\,\|\mathcal{G}\boldsymbol{y}\|_2 \leq \sqrt{\frac{\mu c_s r\log(n)}{m}}\,\|\mathcal{G}\boldsymbol{y}\|_2 \tag{3.14}$$

with probability at least $1-n^{-2}$, where in the second inequality we use the assumption $\mu_0 \leq \mu$. Together with the assumption on m, it follows immediately that

$$\sigma_1(\mathcal{G}\boldsymbol{y}) \leq \frac{\sigma_1(\boldsymbol{L}^0)}{1-\varepsilon_0}.$$

Consequently, one has $\boldsymbol{M} \in \mathcal{C}$ since $\|\boldsymbol{M}\|_{2,\infty} \leq \sqrt{\sigma_1(\mathcal{G}\boldsymbol{y})}\max\{\|\boldsymbol{U}\|_{2,\infty}, \|\boldsymbol{V}\|_{2,\infty}\}$. Moreover, one can easily see that $\boldsymbol{M}\boldsymbol{Q} \in \mathcal{C}$ for all r by r unitary matrices $\boldsymbol{Q}$.

Since $\boldsymbol{Z}^0 = \mathcal{P}_{\mathcal{C}}(\tilde{\boldsymbol{Z}}^0)$ and $\boldsymbol{M}\boldsymbol{Q}_{\tilde{\boldsymbol{z}}^0} \in \mathcal{C}$, one has

$$\operatorname{dist}(\boldsymbol{Z}^0, \boldsymbol{M}) \leq \left\|\boldsymbol{Z}^0 - \boldsymbol{M}\boldsymbol{Q}_{\tilde{\boldsymbol{z}}^0}\right\|_F \leq \left\|\tilde{\boldsymbol{Z}}^0 - \boldsymbol{M}\boldsymbol{Q}_{\tilde{\boldsymbol{z}}^0}\right\|_F = \operatorname{dist}(\tilde{\boldsymbol{Z}}^0, \boldsymbol{M}). \tag{3.15}$$

Therefore, in order to show (3.13), it suffices to bound $\operatorname{dist}(\tilde{\boldsymbol{Z}}^0, \boldsymbol{M})$. We then use the following lemma from the literature.

Lemma 3.3.1 ([53, Lem. 5.4]). *For any* $\boldsymbol{Z},\ \boldsymbol{X} \in \mathbb{C}^{(n+1)\times r}$, *one has*

$$\operatorname{dist}^2(\boldsymbol{Z}, \boldsymbol{X}) \leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(\boldsymbol{X})} \left\|\boldsymbol{Z}\boldsymbol{Z}^* - \boldsymbol{X}\boldsymbol{X}^*\right\|_F^2 .$$

By this result, one has

$$\begin{aligned} \operatorname{dist}^2(\tilde{\boldsymbol{Z}}^0, \boldsymbol{M}) &\leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(\boldsymbol{M})} \left\|\tilde{\boldsymbol{Z}}^0(\tilde{\boldsymbol{Z}}^0)^* - \boldsymbol{M}\boldsymbol{M}^*\right\|_F^2 \\ &= \frac{1}{4(\sqrt{2}-1)\sigma_r(\mathcal{G}\boldsymbol{y})} \left\|\tilde{\boldsymbol{Z}}^0(\tilde{\boldsymbol{Z}}^0)^* - \boldsymbol{M}\boldsymbol{M}^*\right\|_F^2 . \end{aligned} \tag{3.16}$$

Let $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ be four $s \times r$ complex matrices with $s \geq r$. A simple calculation yields

$$\begin{aligned} \langle \boldsymbol{A}\boldsymbol{A}^*, \boldsymbol{B}\boldsymbol{B}^* \rangle + \langle \boldsymbol{C}\boldsymbol{C}^*, \boldsymbol{D}\boldsymbol{D}^* \rangle &= \left\langle \sum_{i=1}^r \boldsymbol{a}_i\boldsymbol{a}_i^*, \sum_{i=1}^r \boldsymbol{b}\boldsymbol{b}^* \right\rangle + \left\langle \sum_{i=1}^r \boldsymbol{c}_i\boldsymbol{c}_i^*, \sum_{i=1}^r \boldsymbol{d}_i\boldsymbol{d}_i^* \right\rangle \\ &= \sum_{i,j=1}^r \left(\left\langle \boldsymbol{a}_i\boldsymbol{a}_i^*, \boldsymbol{b}_j\boldsymbol{b}_j^* \right\rangle + \left\langle \boldsymbol{c}_i\boldsymbol{c}_i^*, \boldsymbol{d}_j\boldsymbol{d}_j^* \right\rangle \right) \\ &= \sum_{i,j=1}^r \left(\langle \boldsymbol{a}_i^*\boldsymbol{b}_j, \boldsymbol{a}_i^*\boldsymbol{b}_j \rangle + \langle \boldsymbol{c}_i^*\boldsymbol{d}_j, \boldsymbol{c}_i^*\boldsymbol{d}_j \rangle \right) \\ &\geq 2 \sum_{i,j=1} \operatorname{Re} \langle \boldsymbol{a}_i^*\boldsymbol{b}_j, \boldsymbol{c}_i^*\boldsymbol{d}_j \rangle \\ &= 2 \sum_{i,j=1} \operatorname{Re} \left\langle \boldsymbol{a}_i\boldsymbol{c}_i^*, \boldsymbol{b}_j\boldsymbol{d}_j^* \right\rangle \\ &= 2 \operatorname{Re} \langle \boldsymbol{A}\boldsymbol{C}^*, \boldsymbol{B}\boldsymbol{D}^* \rangle , \end{aligned} \tag{3.17}$$

where $\boldsymbol{a}_i$, $\boldsymbol{b}_i$, $\boldsymbol{c}_i$ and $\boldsymbol{d}_i$ are the i-th columns of $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ respectively. Then it

follows that

$$\begin{aligned}\left\|\tilde{\boldsymbol{Z}}^0(\tilde{\boldsymbol{Z}}^0)^* - \boldsymbol{M}\boldsymbol{M}^*\right\|_F^2 &= 2\left\|\boldsymbol{U}^0\boldsymbol{\Sigma}^0(\boldsymbol{V}^0)^* - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\right\|_F^2 + \left\|\boldsymbol{U}^0\boldsymbol{\Sigma}^0(\boldsymbol{U}^0)^* - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^*\right\|_F^2 \\ &\quad + \left\|\boldsymbol{V}^0\boldsymbol{\Sigma}(\boldsymbol{V}^0)^* - \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^*\right\|_F^2 \\ &\le 4\left\|\boldsymbol{U}^0\boldsymbol{\Sigma}^0(\boldsymbol{V}^0)^* - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\right\|_F^2 = 4\left\|\boldsymbol{L}^0 - \mathcal{G}\boldsymbol{y}\right\|_F^2 \end{aligned} \tag{3.18}$$

where the inequality follows from

$$\begin{aligned}&\left\|\boldsymbol{U}^0\boldsymbol{\Sigma}^0(\boldsymbol{U}^0)^* - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^*\right\|_F^2 + \left\|\boldsymbol{V}^0\boldsymbol{\Sigma}^0(\boldsymbol{V}^0)^* - \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^*\right\|_F^2 \\ \le& 2\left\|\boldsymbol{U}^0\boldsymbol{\Sigma}^0(\boldsymbol{V}^0)^* - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\right\|_F^2,\end{aligned}$$

which can be easily verified using (3.17). Substituting (3.18) into (3.16) gives

$$\operatorname{dist}^2(\tilde{\boldsymbol{Z}}^0, \boldsymbol{M}) \le \frac{1}{(\sqrt{2}-1)\sigma_r(\mathcal{G}\boldsymbol{y})}\left\|\boldsymbol{L}^0 - \mathcal{G}\boldsymbol{y}\right\|_F^2.$$

Since

$$\left\|\boldsymbol{L}^0 - \mathcal{G}\boldsymbol{y}\right\|_F \lesssim \sqrt{\frac{\mu c_s r^2 \log(n)}{m}}\left\|\mathcal{G}\boldsymbol{y}\right\|_2 \le \varepsilon_0\sigma_r(\mathcal{G}\boldsymbol{y}),$$

we finally have

$$\operatorname{dist}^2(\boldsymbol{Z}^0, \boldsymbol{M}) \le \operatorname{dist}^2(\tilde{\boldsymbol{Z}}^0, \boldsymbol{M}) \le 3\varepsilon_0^2\sigma_r(\mathcal{G}\boldsymbol{y}),$$

which completes the proof of (3.13). □

With Prop. 3.3.1 in place, the proof of Thm. 3.1.1 is complete if we can establish the local contraction property of Alg. 3.1, as stated in the following proposition.

Proposition 3.3.2 (Local Contraction)**.** *Assume* $\boldsymbol{M} \in \mathcal{C}$. *Let* ε_0 *be an absolute constant obeying* $0 < \varepsilon_0 \leq \frac{1}{11}$. *For any matrix* $\boldsymbol{Z} \in \mathcal{C}$, *define*

$$\tilde{\boldsymbol{Z}} = \boldsymbol{Z} - \eta \nabla F(\boldsymbol{Z}) \quad \textit{and} \quad \boldsymbol{Z}^+ = \mathcal{P}_{\mathcal{C}}(\tilde{\boldsymbol{Z}}).$$

There exists a numerical constant $\nu = \frac{1}{10}\sigma_r(\mathcal{G}\boldsymbol{y})$ *such that with probability at least* $1 - c_1 \cdot n^{-2}$,

$$\mathrm{dist}^2(\boldsymbol{Z}^+, \boldsymbol{M}) \leq (1 - \eta\nu)\mathrm{dist}^2(\boldsymbol{Z}, \boldsymbol{M})$$

holds for all $\boldsymbol{Z}$ *obeying* $\mathrm{dist}^2(\boldsymbol{Z}, \boldsymbol{M}) \leq 3\varepsilon_0^2\sigma_r(\mathcal{G}\boldsymbol{y})$ *provided*

$$m \geq c_2\, \varepsilon_0^{-2} \mu^2 c_s^2 \kappa^2 r^2 \log(n) \quad \textit{and} \quad \eta \leq \frac{\sigma_r(\mathcal{G}\boldsymbol{y})}{600(\mu c_s r)^2 \sigma_1^2(\mathcal{G}\boldsymbol{y})}.$$

Based on the same argument as in (3.15), one has $\mathrm{dist}(\boldsymbol{Z}^+, \boldsymbol{M}) \leq \mathrm{dist}(\tilde{\boldsymbol{Z}}, \boldsymbol{M})$. Hence, it suffices to show that

$$\mathrm{dist}^2(\tilde{\boldsymbol{Z}}, \boldsymbol{M}) \leq (1 - \eta\nu)\mathrm{dist}^2(\boldsymbol{Z}, \boldsymbol{M}) \tag{3.19}$$

holds for all matrices $\boldsymbol{Z}$ within a small neighborhood of $\boldsymbol{M}$. Let $\boldsymbol{H} = \boldsymbol{Z} - \boldsymbol{M}\boldsymbol{Q}_{\boldsymbol{Z}}$.

We follow a similar route as in [62] and instead establish the *regularity condition*

$$\operatorname{Re}\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\rangle \geq \frac{\eta}{2}\left\|\nabla F(\boldsymbol{Z})\right\|_F^2 + \frac{\nu}{2}\left\|\boldsymbol{H}\right\|_F^2 \tag{3.20}$$

for all matrices $\boldsymbol{Z}$ that are sufficiently close to $\boldsymbol{M}$. The notation of regularity condition was first introduced in [13] to show the convergence of a non-convex gradient descent algorithm for phase retrieval and since then has been extended to many other problems, see [62] and references therein. Once (3.20) is established, a little algebra yields

$$\begin{aligned}
\operatorname{dist}^2(\tilde{\boldsymbol{Z}}, \boldsymbol{M}) &= \left\|\tilde{\boldsymbol{Z}} - \boldsymbol{M}\boldsymbol{Q}_{\tilde{z}}\right\|_F^2 \leq \left\|\tilde{\boldsymbol{Z}} - \boldsymbol{M}\boldsymbol{Q}_{\boldsymbol{z}}\right\|_F^2 \\
&= \left\|\boldsymbol{H}\right\|_F^2 + \eta^2\left\|\nabla F(\boldsymbol{Z})\right\|_F^2 - 2\eta \operatorname{Re}\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\rangle \\
&\leq (1-\eta\nu)\left\|\boldsymbol{H}\right\|_F^2 \\
&= (1-\eta\nu)\operatorname{dist}^2(\boldsymbol{Z}, \boldsymbol{M}).
\end{aligned}$$

The proof of the regularity condition will occupy the remainder of this section. Even though the proof follows a well-established route, especially that in [62], the details of the proof are nevertheless quite involved and technical. Firstly, our objective function involves a transformation from the matrix domain to the vector domain, and an extra regularizer is also included to preserve the Hankel structure of the matrix. Secondly, we need to establish a key lemma which is closely related to the second largest eigenvalue of a special random graph, as presented in the next subsection.

3.3.1 A Key Ingredient

The following lemma will play a key role in the proof of the regularity condition.

Lemma 3.3.2. *Suppose $\Omega = \{a_k\}_{k=1}^m$, where each a_k is sampled from $\{0, \cdots, n-1\}$ independently and uniformly with replacement. Then for all $\boldsymbol{z} \in \mathbb{R}^{n_1}$ and $\boldsymbol{w} \in \mathbb{R}^{n_2}$,*

$$p^{-1} \sum_{k=1}^{m} \sum_{i+j=a_k} z_i w_j \le \|\boldsymbol{z}\|_1 \|\boldsymbol{w}\|_1 + \sqrt{\frac{24n\log(n)}{p}} \|\boldsymbol{z}\|_2 \|\boldsymbol{w}\|_2$$

holds with probability at least $1 - 2n^{-2}$ provided $m \ge \frac{8}{3}\log(n)$.

Proof. Let $\boldsymbol{H}_a,\ a = 0, \cdots, n-1$, be an $n_1 \times n_2$ matrix with the a-th skew-diagonal entries being equal to one and all the other entries being equal to zero. Notice that $p^{-1}\sum_{k=1}^m \sum_{i+j=a_k} z_i w_j$ can be written as

$$\begin{aligned} p^{-1} \sum_{k=1}^{m} \sum_{i+j=a_k} z_i w_j &= p^{-1} \sum_{k=1}^{m} \boldsymbol{z}^T \boldsymbol{H}_{a_k} \boldsymbol{w} = \boldsymbol{z}^T \left(\frac{n}{m} \sum_{k=1}^{m} \boldsymbol{H}_{a_k} \right) \boldsymbol{w} \\ &= \boldsymbol{z}^T \left(\mathbf{1}_{n_1} \mathbf{1}_{n_2}^T \right) \boldsymbol{w} + \boldsymbol{z}^T \left(\frac{n}{m} \sum_{k=1}^{m} \boldsymbol{H}_{a_k} - \mathbf{1}_{n_1} \mathbf{1}_{n_2}^T \right) \boldsymbol{w} \\ &\le \|\boldsymbol{z}\|_1 \|\boldsymbol{w}\|_1 + \left\| \sum_{k=1}^{m} \left(\frac{n}{m} \boldsymbol{H}_{a_k} - \frac{1}{m} \mathbf{1}_{n_1} \mathbf{1}_{n_2}^T \right) \right\|_2 \|\boldsymbol{z}\|_2 \|\boldsymbol{w}\|_2 . \quad (3.21) \end{aligned}$$

Let $\boldsymbol{Z}_k = \frac{n}{m} \boldsymbol{H}_{a_k} - \frac{1}{m} \mathbf{1}_{n_1} \mathbf{1}_{n_2}^T$. One can easily see that $\mathbb{E}[\boldsymbol{Z}_k] = 0$ and

$$\|\boldsymbol{Z}_k\|_2 \le \left\| \frac{n}{m} \boldsymbol{H}_{a_k} \right\|_2 + \left\| \frac{1}{m} \mathbf{1}_{n_1} \mathbf{1}_{n_2}^T \right\|_2 \le \frac{2n}{m}.$$

Moreover, one has

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right]\\
=&\mathbb{E}\left[\left(\frac{n}{m}\boldsymbol{H}_{a_k}-\frac{1}{m}\mathbf{1}_{n_1}\mathbf{1}_{n_2}^T\right)\left(\frac{n}{m}\boldsymbol{H}_{a_k}^T-\frac{1}{m}\mathbf{1}_{n_2}\mathbf{1}_{n_1}^T\right)\right]\\
=&\frac{n^2}{m^2}\mathbb{E}\left[\boldsymbol{H}_{a_k}\boldsymbol{H}_{a_k}^T\right]-\frac{n}{m^2}\left(\mathbf{1}_{n_1}\mathbf{1}_{n_2}^T\right)\mathbb{E}\left[\boldsymbol{H}_{a_k}\right]-\frac{n}{m^2}\mathbb{E}\left[\boldsymbol{H}_{a_k}\right]\left(\mathbf{1}_{n_2}\mathbf{1}{n_1}^T\right)+\frac{n_2}{m^2}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T\\
=&\frac{n}{m^2}\sum_{a=1}^{n-1}\boldsymbol{H}_a\boldsymbol{H}_a^T-\frac{n_2}{m^2}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T\\
=&\frac{n_2}{m^2}\left(n\boldsymbol{I}_{n_1}-\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T\right),
\end{aligned}$$

so $\left\|\mathbb{E}\left[\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right]\right\|_2\leq\frac{2n^2}{m^2}$. Similarly, one also has $\left\|\mathbb{E}\left[\boldsymbol{Z}_k^T\boldsymbol{Z}_k\right]\right\|_2\leq\frac{2n^2}{m^2}$. Consequently,

$$\max\left\{\left\|\sum_{k=1}^m\mathbb{E}\left[\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right]\right\|_2,\left\|\sum_{k=1}^m\mathbb{E}\left[\boldsymbol{Z}_k^T\boldsymbol{Z}_k\right]\right\|_2\right\}\leq\frac{2n^2}{m}.$$

Thus, the application of the Bernstein's inequality (see for example [51, Thm. 1.6]) yields

$$\mathbb{P}\left\{\left\|\sum_{k=1}^m\boldsymbol{Z}_k\right\|_2>t\right\}\leq(n_1+n_2)\exp\left(\frac{-t^2/2}{2n^2/m+2nt/3m}\right).$$

Letting $t=\sqrt{\frac{24n^2\log(n)}{m}}$ gives

$$\mathbb{P}\left\{\left\|\sum_{k=1}^m\boldsymbol{Z}_k\right\|_2>t\right\}\leq 2n^{-2}$$

provided $m\geq\frac{8}{3}\log(n)$. Substituting this result into (3.21) concludes the proof. □

Remark. Suppose n is odd and $n_1 = n_2 = (n+1)/2$. Let $\boldsymbol{H}$ be an $n_1 \times n_1$ random Hankel matrix, each skew-diagonal of which takes the value 1 with probability p and the value 0 with probability $1-p$. Then $\boldsymbol{H}$ can be viewed as the adjacency matrix corresponding a special random graph. Without rigorous justification, we can see that the largest eigenvalue of $\boldsymbol{H}$, denoted λ_1, is of order about $n_1 p$ as $\mathbb{E}[\boldsymbol{H}] = p\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T$. Let λ_2 be the second largest (in magnitude) eigenvalue of $\boldsymbol{H}$. Roughly speaking, Lem. 3.3.2 says that $|\lambda_2| \approx \sqrt{n_1 p \log(n_1)}$ since $|\lambda_2|$ can be approximated by $\left\|\boldsymbol{H} - p\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T\right\|_2$. Let $\boldsymbol{G}$ be an $n_1 \times n_1$ adjacency matrix of a random graph with n_1 vertex and every edge of which is connected with probability p. That is, each entry of $\boldsymbol{G}$ takes the value 1 with prfobability p and the value 0 with probability $1-p$. It was shown in [23] the second largest (in magnitude) eigenvalue of $\boldsymbol{G}$ is of order at most $\sqrt{n_1 p}$, which has also been extended to singular values in [33]. Thus, our analysis looses a $\sqrt{\log(n_1)}$ factor compared to the result for $\boldsymbol{G}$. However, we want to emphasize that the extra $\sqrt{\log(n)}$ factor in Lem. 3.3.2 does not affect our final result as a log factor will also appear in other place. That being said, we conjecture that the extra $\sqrt{\log(n)}$ factor for $\boldsymbol{H}$ is just an artifact of our analysis framework which uses the Bernstein's inequality under the sampling with replacement model, and it can be eliminated by the spectral techniques used in [23] under the Bernoulli model. We leave this for future work.

3.3.2 Proof of the Regularity Condition

The goal of this subsection is to show that the regularity condition (3.20) holds with high probability. Before proceeding to the formal proof, we first consider the expectation of $\mathrm{Re}\,\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\rangle$ and see what lower bound can be anticipated. With a slight abuse of notation, we denote $\boldsymbol{M}\boldsymbol{Q}_{\boldsymbol{Z}}$ by $\boldsymbol{M}$ throughout this subsection for ease of presentation. Since there exists a close solution for $\boldsymbol{Q}_{\boldsymbol{Z}}$, as presented in (3.8), one can easily verify that

$$\boldsymbol{H}^*\boldsymbol{M} = \boldsymbol{M}^*\boldsymbol{H} \quad \text{and} \quad \boldsymbol{M}^*\boldsymbol{Z} = \boldsymbol{Z}^*\boldsymbol{M} \succeq 0. \tag{3.22}$$

By noting that $\mathbb{E}\left[p^{-1}\mathcal{P}_\Omega\right] = \mathcal{I}$, the expectation of $\mathbb{E}\left[\mathrm{Re}\,\langle \nabla f(\boldsymbol{Z}), \boldsymbol{H}\rangle\right]$ can be bounded below as

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Re}\,\langle \nabla f(\boldsymbol{Z}), \boldsymbol{H}\rangle\right] \\
=&\,\mathrm{Re}\left\langle \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^* - \boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*, \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^* + \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\rangle \\
=&\,\mathrm{Re}\left\langle \boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^* + \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^* + \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*, \boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^* + \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^* + 2\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\rangle \\
=&\left\|\boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^* + \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*\right\|_F^2 + 3\,\mathrm{Re}\left\langle \boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^* + \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*, \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\rangle + 2\left\|\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\|_F^2 \\
\geq&\left\|\boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^* + \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*\right\|_F^2 - 3\left\|\boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^* + \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*\right\|\left\|\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\| + 2\left\|\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\|_F^2 \\
\geq&\frac{1}{2}\left\|\boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^* + \boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*\right\|_F^2 - \frac{5}{2}\left\|\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\|_F^2 \\
=&\frac{1}{2}\left(\left\|\boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\|_F^2 + \left\|\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*\right\|_F^2\right) - \frac{5}{2}\left\|\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\|_F^2 + \mathrm{Re}\left\langle \boldsymbol{H}_{\boldsymbol{U}}^*\boldsymbol{M}_{\boldsymbol{U}}, \boldsymbol{M}_{\boldsymbol{V}}^*\boldsymbol{H}_{\boldsymbol{V}}\right\rangle,
\end{aligned} \tag{3.23}$$

where in the second line we use $\boldsymbol{Z} = \boldsymbol{M} + \boldsymbol{H}$, and in the third line we use the

inequality $a^2 - 3ab + 2b^2 \geq \frac{1}{2}a^2 - \frac{5}{2}b^2$.

Before continuing to bound $\mathbb{E}\left[\mathrm{Re}\left\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\right\rangle\right]$ by adding $\lambda \,\mathrm{Re}\left\langle \nabla g(\boldsymbol{Z}), \boldsymbol{H}\right\rangle$ to $\mathbb{E}\left[\mathrm{Re}\left\langle \nabla f(\boldsymbol{Z}), \boldsymbol{H}\right\rangle\right]$, it might be better to examine the role of $g(\boldsymbol{Z})$ by studying a special case. Suppose $\boldsymbol{H}_{\boldsymbol{U}} = \delta \cdot \boldsymbol{M}_{\boldsymbol{U}}$ and $\boldsymbol{H}_{\boldsymbol{V}} = -\delta \cdot \boldsymbol{M}_{\boldsymbol{V}}$, where $\delta > 0$ is a small numerical constant. Then one has

$$\mathbb{E}\left[\mathrm{Re}\left\langle \nabla f(\boldsymbol{Z}), \boldsymbol{H}\right\rangle\right] = 2\left\|\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\|_F^2 = 2\delta^4\left\|\boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*\right\|_F^2 = 2\delta^4\left\|\boldsymbol{\Sigma}\right\|_F^2,$$

where the last equality follows from the fact $\boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. Since $\left\|\boldsymbol{H}\right\|_F^2 = \delta^2\left\|\boldsymbol{M}\right\|_F^2 = 2\delta^2\left\|\boldsymbol{\Sigma}\right\|_*$, the regularity condition (3.20) cannot be true for $f(\boldsymbol{Z})$ without the regularization function $g(\boldsymbol{Z})$. In this case, one can observe that the mismatch between $\boldsymbol{Z}_{\boldsymbol{U}}^*\boldsymbol{Z}_{\boldsymbol{U}}$ and $\boldsymbol{Z}_{\boldsymbol{V}}^*\boldsymbol{Z}_{\boldsymbol{V}}$ increases compared with the mismatch between $\boldsymbol{M}_{\boldsymbol{U}}^*\boldsymbol{M}_{\boldsymbol{U}}$ and $\boldsymbol{M}_{\boldsymbol{V}}^*\boldsymbol{M}_{\boldsymbol{V}}$ which is equal to zero. Because $g(\boldsymbol{Z})$ penalizes the mismatch between $\boldsymbol{Z}_{\boldsymbol{U}}^*\boldsymbol{Z}_{\boldsymbol{U}}$ and $\boldsymbol{Z}_{\boldsymbol{V}}^*\boldsymbol{Z}_{\boldsymbol{V}}$, one may intuitively expect that it can control the occurrence of this case so that $F(\boldsymbol{Z}) = f(\boldsymbol{Z}) + \lambda g(\boldsymbol{Z})$ could obey the regularity condition.

Let $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{I}_{n_1} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{I}_{n_2} \end{bmatrix}$. We can bound $\mathrm{Re}\left\langle \nabla g(\boldsymbol{Z}), \boldsymbol{H}\right\rangle$ from below as

$$\begin{aligned}
&\mathrm{Re}\left\langle \nabla g(\boldsymbol{Z}), \boldsymbol{H}\right\rangle \\
=&\,\mathrm{Re}\left\langle \boldsymbol{D}\boldsymbol{Z}(\boldsymbol{Z}^*\boldsymbol{D}\boldsymbol{Z}), \boldsymbol{H}\right\rangle = \mathrm{Re}\left\langle \boldsymbol{Z}^*\boldsymbol{D}\boldsymbol{Z}, \boldsymbol{Z}^*\boldsymbol{D}\boldsymbol{H}\right\rangle \\
=&\,\mathrm{Re}\left\langle \boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H} + \boldsymbol{H}^*\boldsymbol{D}\boldsymbol{M} + \boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}, \boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H} + \boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\rangle \\
=&\left\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2 + 3\,\mathrm{Re}\left\langle \boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}, \boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\rangle + \left\|\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2 + \mathrm{Re}\left\langle \boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}, \boldsymbol{H}^*\boldsymbol{D}\boldsymbol{M}\right\rangle
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{2}\left\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2+\frac{1}{2}\left\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}+3\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2-\frac{7}{2}\left\|\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2\\
&+\operatorname{Re}\left\langle\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H},\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{M}\right\rangle\\
=&\frac{1}{2}\left\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2+\frac{1}{2}\left\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}+3\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2-\frac{7}{2}\left\|\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2\\
&+\operatorname{Re}\left\langle\boldsymbol{M}^*\boldsymbol{H},\boldsymbol{H}^*\boldsymbol{M}\right\rangle-4\operatorname{Re}\left\langle\boldsymbol{H}_U^*\boldsymbol{M}_U,\boldsymbol{M}_V^*\boldsymbol{H}_V\right\rangle\\
\geq&\frac{1}{2}\left\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2-\frac{7}{2}\left\|\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2\\
&-4\operatorname{Re}\left\langle\boldsymbol{H}_U^*\boldsymbol{M}_U,\boldsymbol{M}_V^*\boldsymbol{H}_V\right\rangle,
\end{aligned}
\tag{3.24}
$$

where the third equality follows from $\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{M}=\boldsymbol{0}$, the fourth equality follows from

$$
\operatorname{Re}\left\langle\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{M},\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\rangle=\operatorname{Re}\left\langle\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H},\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\rangle=\operatorname{Re}\left\langle\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H},\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\right\rangle,
$$

the last equality follows from

$$
\operatorname{Re}\left\langle\boldsymbol{H}_U^*\boldsymbol{M}_U,\boldsymbol{M}_V^*\boldsymbol{H}_V\right\rangle=\operatorname{Re}\left\langle\boldsymbol{M}_V^*\boldsymbol{H}_V,\boldsymbol{H}_U^*\boldsymbol{M}_U\right\rangle=\operatorname{Re}\left\langle\boldsymbol{M}_U^*\boldsymbol{H}_U,\boldsymbol{H}_V^*\boldsymbol{M}_V\right\rangle,
$$

and the inequality follows from $\boldsymbol{H}^*\boldsymbol{M}=\boldsymbol{M}^*\boldsymbol{H}$, see (3.22).

If we take $\lambda=\frac{1}{4}$, then combining (3.23) and (3.24) together implies

$$
\begin{aligned}
&\mathbb{E}\left[\operatorname{Re}\left\langle\nabla F(\boldsymbol{Z}),\boldsymbol{H}\right\rangle\right]\\
\geq&\frac{1}{2}\left(\left\|\boldsymbol{M}_U\boldsymbol{H}_V^*\right\|_F^2+\left\|\boldsymbol{H}_U\boldsymbol{M}_V^*\right\|_F^2\right)-\frac{5}{2}\left\|\boldsymbol{H}_U\boldsymbol{H}_V^*\right\|_F^2-\frac{7}{8}\left\|\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2\\
&+\frac{1}{8}\left\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2\\
\gtrsim&\left(\sigma_r(\mathcal{G}\boldsymbol{y})-\left\|\boldsymbol{H}\right\|_F^2\right)\left\|\boldsymbol{H}\right\|_F^2+\left\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\right\|_F^2.
\end{aligned}
\tag{3.25}
$$

That is, we have established a lower bound for the expectation of $\mathrm{Re}\,\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\rangle$. As we will show later, $\mathrm{Re}\,\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\rangle$ obeys a similar lower bound with high probability. Moreover, the right hand side of (3.20) can be bounded from above by a similar bound. Therefore, $F(\boldsymbol{Z})$ obeys the regularity condition for sufficiently small $\boldsymbol{H}$. Specifically, we are going to show the following two bounds,

$$\mathrm{Re}\,\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\rangle \geq \frac{1}{10}\sigma_r(\mathcal{G}\boldsymbol{y})\,\|\boldsymbol{H}\|_F^2 + \frac{1}{8}\,\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2\,, \tag{3.26}$$

$$\|\nabla F(\boldsymbol{Z})\|_F^2 \leq 60(\mu c_s r)^2\sigma_1^2(\mathcal{G}\boldsymbol{y})\,\|\boldsymbol{H}\|_F^2 + \frac{1}{2}\sigma_1(\mathcal{G}\boldsymbol{y})\,\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2\,, \tag{3.27}$$

hold with high probability provided $\|\boldsymbol{H}\|_F^2 \leq 3\varepsilon_0^2\sigma_r(\mathcal{G}\boldsymbol{y})$ and $m \gtrsim \varepsilon_0^{-2}\mu^2 c_s^2\kappa^2 r^2\log(n)$ for $\varepsilon_0 \leq \frac{1}{11}$. The above two inequalities are typically referred to as the *local curvature property* and the *local smooth property* of the function $F(\boldsymbol{Z})$ in the literature, see for example [13, 62]. Once they are established, one can easily see that $F(\boldsymbol{Z})$ obeys the regularity condition (3.20) with

$$\eta \leq \frac{\sigma_r(\mathcal{G}\boldsymbol{y})}{600(\mu c_s r)^2\sigma_1^2(\mathcal{G}\boldsymbol{y})} \quad \text{and} \quad \nu = \frac{1}{10}\sigma_r(\mathcal{G}\boldsymbol{y}).$$

3.3.2.1 Proof of Local Curvature Property

Since $\mathrm{Re}\,\langle \nabla g(\boldsymbol{Z}), \boldsymbol{H}\rangle$ is deterministic and we have already obtained its lower bound in (3.24), it only remains to work out the lower bound for $\mathrm{Re}\,\langle \nabla f(\boldsymbol{Z}), \boldsymbol{H}\rangle$ and then combine it together with that for $\mathrm{Re}\,\langle \nabla g(\boldsymbol{Z}), \boldsymbol{H}\rangle$.

Note that

$$
\begin{aligned}
&\operatorname{Re}\langle \nabla f(\boldsymbol{Z}), \boldsymbol{H}\rangle \\
=&\operatorname{Re}\left\langle (\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U\boldsymbol{Z}_V^*)+p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*-\boldsymbol{M}_U\boldsymbol{M}_V^*), \boldsymbol{H}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{H}_V^*\right\rangle \\
=&\operatorname{Re}\left\langle (\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U\boldsymbol{Z}_V^*-\boldsymbol{M}_U\boldsymbol{M}_V^*), \boldsymbol{H}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{H}_V^*\right\rangle \\
&+\operatorname{Re}\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*-\boldsymbol{M}_U\boldsymbol{M}_V^*), \boldsymbol{H}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{H}_V^*\right\rangle \\
=&\operatorname{Re}\left\langle (\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{M}_V^*+\boldsymbol{M}_U\boldsymbol{H}_V^*+\boldsymbol{H}_U\boldsymbol{H}_V^*), \boldsymbol{H}_U\boldsymbol{M}_V^*+\boldsymbol{M}_U\boldsymbol{H}_V^*+2\boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle \\
&+\operatorname{Re}\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^*+\boldsymbol{M}_U\boldsymbol{H}_V^*+\boldsymbol{H}_U\boldsymbol{H}_V^*), \boldsymbol{H}_U\boldsymbol{M}_V^*+\boldsymbol{M}_U\boldsymbol{H}_V^*+2\boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle \\
:=&I_1+I_2, \qquad (3.28)
\end{aligned}
$$

where the second equality follows from the fact $(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{M}_U\boldsymbol{M}_V^*)=\boldsymbol{0}$.

Lower bound for I_1. The first term I_1 can be bounded directly as follows:

$$
\begin{aligned}
I_1 \geq& \left\|(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{M}_V^*+\boldsymbol{M}_U\boldsymbol{H}_V^*)\right\|_F^2 + +2\left\|(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{H}_V^*)\right\|_F^2 \\
&-3\left\|(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{M}_V^*+\boldsymbol{M}_U\boldsymbol{H}_V^*)\right\|_F\left\|(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{H}_V^*)\right\|_F \\
\geq& \frac{11}{20}\left\|(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{M}_V^*+\boldsymbol{M}_U\boldsymbol{H}_V^*)\right\|_F^2-3\left\|(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{H}_V^*)\right\|_F^2,
\end{aligned}
$$

where the second inequality follows from $a^2-3ab+2b^2\geq\frac{11}{20}a^2-3b^2$.

Lower bound for I_2. Let $\boldsymbol{G}_a,\ a=0,\cdots,n-1$, be an $n_1\times n_2$ matrix with the a-th skew-diagonal entries being equal to $1/\sqrt{w_a}$ and all the other entries being

equal to zero. Then,

$$\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*) = \left\{\left\langle \boldsymbol{G}_a, \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\rangle\right\}_{a=0}^{n-1}$$

and

$$\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{H}_V^*) = \left\{\left\langle \boldsymbol{G}_a, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right\}_{a=0}^{n-1}.$$

It follows that

$$\begin{aligned}
&\operatorname{Re}\left\langle \mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^* + \boldsymbol{H}_U\boldsymbol{H}_V^*), \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^* + 2\boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle \\
=&\operatorname{Re}\left\langle \mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^* + \boldsymbol{H}_U\boldsymbol{H}_V^*), \mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^* + 2\boldsymbol{H}_U\boldsymbol{H}_V^*)\right\rangle \\
=&\left\langle \mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*), \mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*)\right\rangle \\
&+ 2\left\langle \mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{H}_V^*), \mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{H}_V^*)\right\rangle + 3\operatorname{Re}\left\langle \mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*), \mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{H}_V^*)\right\rangle \\
=&\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\rangle\right|^2 + 2\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2 \\
&+ 3\operatorname{Re}\left(\sum_{k=1}^m \overline{\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\rangle}\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right) \\
\geq&\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\rangle\right|^2 + 2\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2 \\
&- 3\sqrt{\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\rangle\right|^2}\sqrt{\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2} \\
\geq&\frac{11}{20}\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\rangle\right|^2 - 3\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2 \\
=&\frac{11}{20}\left\langle \mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*), \mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*)\right\rangle - 3\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2,
\end{aligned}$$

where the third equality and the last equality follow from (3.12), the first inequality follows from the Hölder inequality, and the second inequality follows from $a^2 - 3ab + 2b^2 \geq \frac{11}{20}a^2 - 3b^2$. Consequently,

$$I_2 \geq \frac{11}{20}\left\langle p^{-1}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*), \mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*)\right\rangle - 3p^{-1}\sum_{k=1}^{m}\left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2. \tag{3.29}$$

We can bound $p^{-1}\sum_{k=1}^{m}\left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2$ from above by Lem. 3.3.2 as follows:

$$\begin{aligned}
&p^{-1}\sum_{k=1}^{m}\left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2 \\
=&p^{-1}\sum_{k=1}^{m}\left|\frac{1}{\sqrt{w_{a_k}}}\sum_{i+j=a_k}\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2 \\
\leq&p^{-1}\sum_{k=1}^{m}\sum_{i+j=a_k}\left|\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2 \\
\leq&p^{-1}\sum_{k=1}^{m}\sum_{i+j=a_k}\left\|\boldsymbol{H}_U^{(i,:)}\right\|_2^2\left\|\boldsymbol{H}_V^{(j,:)}\right\|_2^2 \\
\leq&\left\|\boldsymbol{H}_U\right\|_F^2\left\|\boldsymbol{H}_V\right\|_F^2 + \sqrt{\frac{24n\log(n)}{p}}\sqrt{\sum_{i=1}^{n_1}\left\|\boldsymbol{H}_U^{(i,:)}\right\|_2^4}\sqrt{\sum_{j=1}^{n_2}\left\|\boldsymbol{H}_V^{(j,:)}\right\|_2^4} \\
\leq&\left\|\boldsymbol{H}_U\right\|_F^2\left\|\boldsymbol{H}_V\right\|_F^2 + \sqrt{\frac{24n\log(n)}{p}}\left(\left\|\boldsymbol{H}_U\right\|_{2,\infty}\left\|\boldsymbol{H}_U\right\|_F\right)\left(\left\|\boldsymbol{H}_V\right\|_{2,\infty}\left\|\boldsymbol{H}_V\right\|_F\right) \\
\leq&\left\|\boldsymbol{H}_U\right\|_F^2\left\|\boldsymbol{H}_V\right\|_F^2 + \sqrt{\frac{24n\log(n)}{p}}\left(\frac{4\mu c_s r}{n}\sigma\right)\left\|\boldsymbol{H}_U\right\|_F\left\|\boldsymbol{H}_V\right\|_F \\
\leq&\frac{1}{4}\left\|\boldsymbol{H}\right\|_F^4 + \sqrt{\frac{96\mu^2c_s^2r^2\log(n)}{m}}\frac{\sigma_1(\boldsymbol{L}^0)}{1-\varepsilon_0}\left\|\boldsymbol{H}\right\|_F^2 \\
\leq&\left(\frac{3\varepsilon_0^2}{4} + \frac{\varepsilon_0(1+\varepsilon_0)}{1-\varepsilon_0}\right)\sigma_r(\mathcal{G}\boldsymbol{y})\left\|\boldsymbol{H}\right\|_F^2,
\end{aligned}$$

where the fourth line follows from Lem. 3.3.2, the sixth line follows from

$$\max\left\{\|\boldsymbol{H}_U\|_{2,\infty}, \|\boldsymbol{H}_V\|_{2,\infty}\right\} = \|\boldsymbol{H}\|_{2,\infty} \le \|\boldsymbol{M}\|_{2,\infty} + \|\boldsymbol{Z}\|_{2,\infty} \le 2\sqrt{\frac{\mu c_s r}{n}}\sigma,$$

and the last line follows from (3.14) and the assumptions on $\|\boldsymbol{H}\|_F^2$ and m.

Lower bound for $\operatorname{Re}\langle\nabla f(\boldsymbol{Z}), \boldsymbol{H}\rangle$**.** Recall the tangent space of the rank r matrix manifold at $\mathcal{G}\boldsymbol{y}$ is denoted as $\mathcal{S}$. One can easily see that $\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^* \in \mathcal{S}$. Substituting the bound for $p^{-1}\sum_{k=1}^m \left|\left\langle \boldsymbol{G}_{a_k}, \boldsymbol{H}_U\boldsymbol{H}_V^*\right\rangle\right|^2$ into (3.29) and then combining the lower bounds for I_1 and I_2 together yields

$$\begin{aligned}
&\operatorname{Re}\langle\nabla f(\boldsymbol{Z}), \boldsymbol{H}\rangle \\
\ge&\frac{11}{20}\left\|(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*)\right\|_F^2 - 3\left\|(\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{H}_U\boldsymbol{H}_V^*)\right\|_F^2 \\
&+\frac{11}{20}\left\langle p^{-1}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*), \mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*)\right\rangle \\
&-\left(\frac{9\varepsilon_0^2}{4} + \frac{3\varepsilon_0(1+\varepsilon_0)}{1-\varepsilon_0}\right)\sigma_r(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2 \\
\ge&\frac{11}{20}\left\|\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\|_F^2 - 3\left\|\boldsymbol{H}_U\boldsymbol{H}_V^*\right\|_F^2 - \left(\frac{9\varepsilon_0^2}{4} + \frac{3\varepsilon_0(1+\varepsilon_0)}{1-\varepsilon_0}\right)\sigma_r(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2 \\
&-\frac{11}{20}\left\langle\mathcal{G}(\mathcal{I}-p^{-1}\mathcal{P}_\Omega)\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*), \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\rangle \\
=&\frac{11}{20}\left\|\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\|_F^2 - 3\left\|\boldsymbol{H}_U\boldsymbol{H}_V^*\right\|_F^2 - \left(\frac{9\varepsilon_0^2}{4} + \frac{3\varepsilon_0(1+\varepsilon_0)}{1-\varepsilon_0}\right)\sigma_r(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2 \\
&-\frac{11}{20}\left\langle\mathcal{P}_\mathcal{S}\mathcal{G}(\mathcal{I}-p^{-1}\mathcal{P}_\Omega)\mathcal{G}^*\mathcal{P}_\mathcal{S}(\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*), \boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\rangle \\
\ge&\frac{11}{20}(1-\varepsilon_0)\left\|\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\|_F^2 - \left(\frac{9\varepsilon_0^2}{2} + \frac{3\varepsilon_0(1+\varepsilon_0)}{1-\varepsilon_0}\right)\sigma_r(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2 \\
\ge&\frac{1}{2}\left\|\boldsymbol{H}_U\boldsymbol{M}_V^* + \boldsymbol{M}_U\boldsymbol{H}_V^*\right\|_F^2 - \left(\frac{9\varepsilon_0^2}{2} + \frac{3\varepsilon_0(1+\varepsilon_0)}{1-\varepsilon_0}\right)\sigma_r(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2 \\
\ge&\frac{1}{8}\sigma_r(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2 + \operatorname{Re}\left\langle\boldsymbol{H}_U^*\boldsymbol{M}_U, \boldsymbol{M}_V^*\boldsymbol{H}_V\right\rangle,
\end{aligned} \tag{3.30}$$

where the second inequality follows from the fact $\mathcal{G}\mathcal{G}^*$ is a projection operator, the third inequality holds with probability at least $1 - n^{-2}$ (see Lem. 2.3.2) under the assumption on m and $\|\boldsymbol{H}\|_F^2$, and the last inequality follows from $\left\|\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*\right\|_F \geq \sigma_r(\boldsymbol{M}_{\boldsymbol{V}}) \|\boldsymbol{H}_{\boldsymbol{U}}\|_F$, $\left\|\boldsymbol{M}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\|_F \geq \sigma_r(\boldsymbol{M}_{\boldsymbol{U}}) \|\boldsymbol{H}_{\boldsymbol{V}}\|_F$, and the assumption $\varepsilon_0 \leq \frac{1}{11}$.

Lower bound for $\operatorname{Re}\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\rangle$**.** Let $\lambda = \frac{1}{4}$. Combining the lower bound in (3.30) for $\operatorname{Re}\langle \nabla f(\boldsymbol{Z}), \boldsymbol{H}\rangle$ and the lower bound in (3.24) for $\operatorname{Re}\langle \nabla g(\boldsymbol{Z}), \boldsymbol{H}\rangle$ together gives

$$\begin{aligned}\operatorname{Re}\langle \nabla F(\boldsymbol{Z}), \boldsymbol{H}\rangle &\geq \frac{1}{8}\sigma_r(\mathcal{G}\boldsymbol{y}) \|\boldsymbol{H}\|_F^2 - \frac{7}{8} \|\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\|_F^2 + \frac{1}{8} \|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2 \\ &\geq \frac{1}{10}\sigma_r(\mathcal{G}\boldsymbol{y}) \|\boldsymbol{H}\|_F^2 + \frac{1}{8} \|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2,\end{aligned}$$

where the second inequality follows from

$$\|\boldsymbol{H}^*\boldsymbol{D}\boldsymbol{H}\|_F^2 \leq \|\boldsymbol{H}\|_F^4 \leq 3\varepsilon_0^2\sigma_r(\mathcal{G}\boldsymbol{y}) \|\boldsymbol{H}\|_F^2$$

and the assumption $\varepsilon_0 \leq \frac{1}{11}$. This concludes the proof of (3.26).

3.3.2.2 Proof of Local Smoothness Property

Since

$$\|\nabla F(\boldsymbol{Z})\|_F^2 \leq 2 \|\nabla f(\boldsymbol{Z})\|_F^2 + 2\lambda^2 \|\nabla g(\boldsymbol{Z})\|_F^2, \tag{3.31}$$

it suffices to bound $\|\nabla f(\boldsymbol{Z})\|_F^2$ and $\|\nabla g(\boldsymbol{Z})\|_F^2$ separately.

Upper bound for $\|\nabla g(\boldsymbol{Z})\|_F^2$. We begin with the upper bound for $\|\nabla g(\boldsymbol{Z})\|_F^2$, which can be obtained in a straightforward way,

$$\begin{aligned}
\|\nabla g(\boldsymbol{Z})\|_F^2 &= \|\boldsymbol{D}\boldsymbol{Z}\boldsymbol{Z}^*\boldsymbol{D}\boldsymbol{Z}\|_F^2 = \|\boldsymbol{D}(\boldsymbol{Z}\boldsymbol{Z}^* - \boldsymbol{M}\boldsymbol{M}^*)\boldsymbol{D}\boldsymbol{Z} + \boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{Z}\|_F^2 \\
&\le 2\|\boldsymbol{D}(\boldsymbol{Z}\boldsymbol{Z}^* - \boldsymbol{M}\boldsymbol{M}^*)\boldsymbol{D}\boldsymbol{Z}\|_F^2 + 2\|\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{Z}\|_F^2 \\
&\le 2\|\boldsymbol{Z}\|_2^2\|\boldsymbol{Z}\boldsymbol{Z}^* - \boldsymbol{M}\boldsymbol{M}^*\|_F^2 + 2\|\boldsymbol{M}\|_2^2\|\boldsymbol{M}^*\boldsymbol{D}(\boldsymbol{M}+\boldsymbol{H})\|_F^2 \\
&= 2\|\boldsymbol{Z}\|_2^2\|\boldsymbol{M}\boldsymbol{H}^* + \boldsymbol{H}\boldsymbol{M}^* + \boldsymbol{H}\boldsymbol{H}^*\|_F^2 + 2\|\boldsymbol{M}\|_2^2\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2 \\
&\le 6\|\boldsymbol{Z}\|_2^2\left(2\|\boldsymbol{M}\|_2^2\|\boldsymbol{H}\|_F^2 + \|\boldsymbol{H}\|_F^4\right) + 2\|\boldsymbol{M}\|_2^2\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2 \\
&\le 6\left(\sqrt{3\varepsilon_0^2\sigma_r(\mathcal{G}\boldsymbol{y})} + \sqrt{2\sigma_1(\mathcal{G}\boldsymbol{y})}\right)^2\left(4\sigma_1(\mathcal{G}\boldsymbol{y}) + 3\varepsilon_0^2\sigma_r(\mathcal{G}\boldsymbol{y})\right)\|\boldsymbol{H}\|_F^2 \\
&\quad + 4\sigma_1(\mathcal{G}\boldsymbol{y})\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2 \\
&\le 60\sigma_1^2(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2 + 4\sigma_1(\mathcal{G}\boldsymbol{y})\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2, \qquad (3.32)
\end{aligned}$$

where the third equality follows from $\boldsymbol{Z} = \boldsymbol{M} + \boldsymbol{H}$ and $\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{M} = \boldsymbol{0}$, the fourth inequality follows from $\|\boldsymbol{M}\|_2 = \sqrt{2\sigma_1(\mathcal{G}\boldsymbol{y})}$ and

$$\|\boldsymbol{Z}\|_2 \le \|\boldsymbol{Z} - \boldsymbol{M}\|_2 + \|\boldsymbol{M}\|_2 \le \|\boldsymbol{Z} - \boldsymbol{M}\|_F + \|\boldsymbol{M}\|_2,$$

and the last line follows from the assumption $\varepsilon_0 \le \frac{1}{11}$.

In order to bound $\|\nabla f(\boldsymbol{Z})\|_F^2$, we consider $|\langle \nabla f(\boldsymbol{Z}), \boldsymbol{X}\rangle|^2$ for matrices $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_U & \boldsymbol{X}_V\end{bmatrix}^T$ with unit Frobenius norm (i.e., $\|\boldsymbol{X}_U\|_F^2 + \|\boldsymbol{X}_V\|_F^2 = 1$).

Note that

$$
\begin{aligned}
&|\langle \nabla f(\boldsymbol{Z}), \boldsymbol{X}\rangle|^2 \\
=&\left|\left\langle (\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U\boldsymbol{Z}_V^*)+p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*-\boldsymbol{M}_U\boldsymbol{M}_V^*), \boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|^2 \\
=&\left|\left\langle (\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U\boldsymbol{Z}_V^*-\boldsymbol{M}_U\boldsymbol{M}_V^*)+p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*-\boldsymbol{M}_U\boldsymbol{M}_V^*), \boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|^2 \\
\le&2\left|\left\langle (\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U\boldsymbol{Z}_V^*-\boldsymbol{M}_U\boldsymbol{M}_V^*), \boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|^2 \\
&+2\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{Z}_V^*-\boldsymbol{M}_U\boldsymbol{M}_V^*), \boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|^2 \\
=&2\left|\left\langle (\mathcal{I}-\mathcal{G}\mathcal{G}^*)(\boldsymbol{Z}_U\boldsymbol{H}_V^*+\boldsymbol{H}_U\boldsymbol{M}_V^*), \boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|^2 \\
&+2\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{H}_V^*+\boldsymbol{H}_U\boldsymbol{M}_V^*), \boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|^2 \\
:=&2\cdot I_3+2\cdot I_4. \qquad (3.33)
\end{aligned}
$$

Upper bound for I_3. Since

$$
\begin{aligned}
&\|\boldsymbol{Z}_U\|_2 \le \|\boldsymbol{M}_U\|_2+\|\boldsymbol{H}_U\|_2 \le \|\boldsymbol{M}_U\|_2+\|\boldsymbol{H}\|_F \le (1+\sqrt{3}\varepsilon_0)\sqrt{\sigma_1(\mathcal{G}\boldsymbol{y})}, \\
&\|\boldsymbol{Z}_V\|_2 \le \|\boldsymbol{M}_V\|_2+\|\boldsymbol{H}_V\|_2 \le \|\boldsymbol{M}_U\|_2+\|\boldsymbol{H}\|_F \le (1+\sqrt{3}\varepsilon_0)\sqrt{\sigma_1(\mathcal{G}\boldsymbol{y})}.
\end{aligned}
$$

one has

$$
\begin{aligned}
\left\|\boldsymbol{Z}_U\boldsymbol{H}_V^*+\boldsymbol{H}_U\boldsymbol{M}_V^*\right\|_F^2 &\le 2\left(\left\|\boldsymbol{Z}_U\boldsymbol{H}_V^*\right\|_F^2+\left\|\boldsymbol{H}_U\boldsymbol{M}_V^*\right\|_F^2\right) \\
&\le 2\left(\|\boldsymbol{Z}_U\|_2^2\|\boldsymbol{H}_V\|_F^2+\|\boldsymbol{M}_V\|_2^2\|\boldsymbol{H}_U\|_F^2\right) \\
&\le 2(1+\sqrt{3}\varepsilon_0)^2\sigma_1(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2
\end{aligned}
$$

and

$$\begin{aligned}\left\|\boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\|_F^2 &\le 2\left(\left\|\boldsymbol{X}_U\boldsymbol{Z}_V^*\right\|_F^2+\left\|\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\|_F^2\right)\\ &\le 2\left(\|\boldsymbol{Z}_V\|_2^2\|\boldsymbol{X}_U\|_F^2+\|\boldsymbol{Z}_U\|_2^2\|\boldsymbol{X}_V\|_F^2\right)\\ &\le 2(1+\sqrt{3}\varepsilon_0)^2\sigma_1(\mathcal{G}\boldsymbol{y}),\end{aligned}$$

where in the last line we have utilized $\|\boldsymbol{X}_U\|_F^2+\|\boldsymbol{X}_V\|_F^2=1$. Because $\mathcal{I}-\mathcal{G}\mathcal{G}^*$ is a projection operator, I_3 can be bounded as follows:

$$I_3\le\left\|\boldsymbol{Z}_U\boldsymbol{H}_V^*+\boldsymbol{H}_U\boldsymbol{M}_V^*\right\|_F^2\cdot\left\|\boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\|_F^2\le 4(1+\sqrt{3}\varepsilon_0)^4\sigma_1^2(\mathcal{G}\boldsymbol{y})\|\boldsymbol{H}\|_F^2.$$

Upper bound for I_4. Notice that

$$\begin{aligned}&\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{H}_V^*+\boldsymbol{H}_U\boldsymbol{M}_V^*),\boldsymbol{X}_U\boldsymbol{Z}_V^*+\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|\\ \le&\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{H}_V^*),\boldsymbol{X}_U\boldsymbol{Z}_V^*\right\rangle\right|+\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{H}_V^*),\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|\\ &+\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^*),\boldsymbol{X}_U\boldsymbol{Z}_V^*\right\rangle\right|+\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_U\boldsymbol{M}_V^*),\boldsymbol{Z}_U\boldsymbol{X}_V^*\right\rangle\right|.\end{aligned}\tag{3.34}$$

We can bound $\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{H}_V^*),\boldsymbol{X}_U\boldsymbol{Z}_V^*\right\rangle\right|$ as follows:

$$\begin{aligned}&\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{H}_V^*),\boldsymbol{X}_U\boldsymbol{Z}_V^*\right\rangle\right|\\ =&p^{-1}\left|\left\langle\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_U\boldsymbol{H}_V^*),\mathcal{G}^*(\boldsymbol{X}_U\boldsymbol{Z}_V^*)\right\rangle\right|\\ \le&p^{-1}\sum_{k=1}^m\left\{\left|\left\langle\boldsymbol{G}_{a_k},\boldsymbol{Z}_U\boldsymbol{H}_V^*\right\rangle\right|\left|\left\langle\boldsymbol{G}_{a_k},\boldsymbol{X}_U\boldsymbol{Z}_V^*\right\rangle\right|\right\}\end{aligned}$$

$$
\begin{aligned}
=&p^{-1}\sum_{k=1}^{m}\left\{\left|\frac{1}{\sqrt{w_{a_k}}}\sum_{i+j=a_k}\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\rangle\right|\left|\frac{1}{\sqrt{w_{a_k}}}\sum_{i+j=a_k}\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*\right\rangle\right|\right\}\\
\leq&p^{-1}\sum_{k=1}^{m}\left\{\frac{1}{\sqrt{w_{a_k}}}\sum_{i+j=a_k}\left|\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\rangle\right|\frac{1}{\sqrt{w_{a_k}}}\sum_{i+j=a_k}\left|\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*\right\rangle\right|\right\}\\
=&p^{-1}\sum_{k=1}^{m}\left\{\sqrt{\left(\frac{1}{\sqrt{w_{a_k}}}\sum_{i+j=a_k}\left|\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\rangle\right|\right)^2}\sqrt{\left(\frac{1}{\sqrt{w_{a_k}}}\sum_{i+j=a_k}\left|\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*\right\rangle\right|\right)^2}\right\}\\
\leq&p^{-1}\sum_{k=1}^{m}\left\{\left(\sum_{i+j=a_k}\left|\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*\right\rangle\right|^2\right)^{1/2}\left(\sum_{i+j=a_k}\left|\left\langle \boldsymbol{e}_i\boldsymbol{e}_j^T, \boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*\right\rangle\right|^2\right)^{1/2}\right\}\\
\leq&p^{-1}\sum_{k=1}^{m}\left\{\left(\sum_{i+j=a_k}\left\|\boldsymbol{Z}_{\boldsymbol{U}}^{(i,:)}\right\|_2^2\left\|\boldsymbol{H}_{\boldsymbol{V}}^{(j,:)}\right\|_2^2\right)^{1/2}\left(\sum_{i+j=a_k}\left\|\boldsymbol{X}_{\boldsymbol{U}}^{(i,:)}\right\|_2^2\left\|\boldsymbol{Z}_{\boldsymbol{V}}^{(j,:)}\right\|_2^2\right)^{1/2}\right\}\\
\leq&p^{-1}\sum_{k=1}^{m}\left\{\left(\|\boldsymbol{Z}\|_{2,\infty}\|\boldsymbol{H}_{\boldsymbol{V}}\|_F\right)\left(\|\boldsymbol{Z}\|_{2,\infty}\|\boldsymbol{X}_{\boldsymbol{U}}\|_F\right)\right\}\\
\leq&\mu c_s r\sigma\|\boldsymbol{H}_{\boldsymbol{V}}\|_F\|\boldsymbol{X}_{\boldsymbol{U}}\|_F,
\end{aligned}
$$

where in the last line, we utilize $\|\boldsymbol{Z}\|_{2,\infty}^2 \leq \mu c_s r\sigma/n$. Similar upper bounds can be established for the other three terms in (3.34). That is,

$$
\begin{aligned}
&\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{H}_{\boldsymbol{V}}^*), \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{X}_{\boldsymbol{V}}^*\right\rangle\right| \leq \mu c_s r\sigma\|\boldsymbol{H}_{\boldsymbol{V}}\|_F\|\boldsymbol{X}_{\boldsymbol{V}}\|_F,\\
&\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*), \boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{Z}_{\boldsymbol{V}}^*\right\rangle\right| \leq \mu c_s r\sigma\|\boldsymbol{H}_{\boldsymbol{U}}\|_F\|\boldsymbol{X}_{\boldsymbol{U}}\|_F,\\
&\left|\left\langle p^{-1}\mathcal{G}\mathcal{P}_\Omega\mathcal{G}^*(\boldsymbol{H}_{\boldsymbol{U}}\boldsymbol{M}_{\boldsymbol{V}}^*), \boldsymbol{Z}_{\boldsymbol{U}}\boldsymbol{X}_{\boldsymbol{V}}^*\right\rangle\right| \leq \mu c_s r\sigma\|\boldsymbol{H}_{\boldsymbol{U}}\|_F\|\boldsymbol{X}_{\boldsymbol{V}}\|_F.
\end{aligned}
$$

Combining these four upper bounds together yields

$$
\begin{aligned}
I_4 &\leq (\mu c_s r\sigma)^2\left(\|\boldsymbol{H}_{\boldsymbol{V}}\|_F\|\boldsymbol{X}_{\boldsymbol{U}}\|_F+\|\boldsymbol{H}_{\boldsymbol{V}}\|_F\|\boldsymbol{X}_{\boldsymbol{V}}\|_F+\|\boldsymbol{H}_{\boldsymbol{U}}\|_F\|\boldsymbol{X}_{\boldsymbol{U}}\|_F+\|\boldsymbol{H}_{\boldsymbol{U}}\|_F\|\boldsymbol{X}_{\boldsymbol{V}}\|_F\right)^2\\
&= (\mu c_s r\sigma)^2\left(\|\boldsymbol{H}_{\boldsymbol{U}}\|_F+\|\boldsymbol{H}_{\boldsymbol{V}}\|_F\right)^2\left(\|\boldsymbol{X}_{\boldsymbol{U}}\|_F+\|\boldsymbol{X}_{\boldsymbol{V}}\|_F\right)^2 \leq 4(\mu c_s r\sigma)^2\|\boldsymbol{H}\|_F^2
\end{aligned}
$$

where in the last line we have used the fact $\|\boldsymbol{X}_U\|_F^2 + \|\boldsymbol{X}_V\|_F^2 = 1$.

Upper bound for $\|\nabla f(\boldsymbol{Z})\|_F^2$. Substituting the upper bounds for I_3 and I_4 into (3.33) give the upper bound for $\|\nabla f(\boldsymbol{Z})\|_F^2$,

$$\|\nabla f(\boldsymbol{Z})\|_F^2 \leq 8\left((1+\sqrt{3}\varepsilon_0)^4\sigma_1^2(\mathcal{G}\boldsymbol{y}) + (\mu c_s r\sigma)^2\right)\|\boldsymbol{H}\|_F^2 .$$

Upper bound for $\|\nabla F(\boldsymbol{Z})\|_F^2$. Noting $\lambda = 1/4$, $\sigma \leq (1+\varepsilon_0)\sigma_1(\mathcal{G}\boldsymbol{y})/(1-\varepsilon_0)$, and $\varepsilon_0 \leq 1/11$, after substituting the upper bounds for $\|\nabla f(\boldsymbol{Z})\|_F^2$ and $\|\nabla g(\boldsymbol{Z})\|_F^2$ into (3.31), we get

$$\begin{aligned}
&\|\nabla F(\boldsymbol{Z})\|_F^2 \\
\leq& 16\left((1+\sqrt{3}\varepsilon_0)^4\sigma_1^2(\mathcal{G}\boldsymbol{y}) + (\mu c_s r\sigma)^2 + \frac{60}{128}\sigma_1^2(\mathcal{G}\boldsymbol{y})\right)\|\boldsymbol{H}\|_F^2 + \frac{1}{2}\sigma_1(\mathcal{G}\boldsymbol{y})\,\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2 \\
\leq& 60(\mu c_s r)^2\sigma_1^2(\mathcal{G}\boldsymbol{y})\,\|\boldsymbol{H}\|_F^2 + \frac{1}{2}\sigma_1(\mathcal{G}\boldsymbol{y})\,\|\boldsymbol{M}^*\boldsymbol{D}\boldsymbol{H}\|_F^2 ,
\end{aligned}$$

which completes the proof of (3.27).

CHAPTER 4
CONCLUSIONS

Spectrally sparse signals arise in many applications of signal processing. We consider the problem of reconstructing a spectrally sparse signal composed of r undamped or damped complex sinusoids from partial samples. Previous convex methods do not scale well with respect to the signal length n. After converting to a low-rank Hankel matrix completion problem, we propose some non-convex methods [8, 9] to solve it.

In Chapter 2, we first propose an iterative hard thresholding (IHT) algorithm for low-rank Hankel matrix completion. By adding a projection onto a low-dimensional subspace before rank truncation every iteration, a fast iterative hard thresholding (FIHT) algorithm is derived. The projection step also helps us derive the theoretical guarantees. Overall, FIHT has the following appealing features:

- provided there are $O(r^2 \log^2(n))$ samples and FIHT is initialized via resampling and trimming, the algorithm converges linearly to the ground truth with overwhelming probability;
- the computation cost is $O(r^2 n + rn \log(n) + r^3)$ per iteration, which scales well with respect to the length n of the signal;
- robust to additive noise, not sensitive to rank mis-specification.

In Chapter 3, we adopt the Burer-Monteiro representation of low-rank matrices and try to find a low-rank matrix that is of Hankel structure by minimizing a

non-convex function. We solve the minimization via gradient descent, followed by projection onto a set of incoherence property every iteration. Overall, this algorithm has the following appealing features:

- provided there are $O(r^2 \log(n))$ samples, PGD with the proposed initialization scheme converges to the ground truth with overwhelming probability;
- the computation cost is $O(r^2 n + rn \log(n))$ per iteration, which scales well with respect to the length n of the signal;
- robust to additive noise, not sensitive to rank mis-specification.

By comparison, FIHT is faster for easy problems while PGD has higher recovery rates for hard problems.

We could extend the proposed algorithms to solve problems including super-resolution [12], image inpainting [32], etc. There is also follow-up work to do based on our numerical findings about FIHT and PGD.

In the analysis of FIHT, we derive the sampling complexity $O(r^2 \log^2(n))$ using resampling initialization. The derived sampling complexity for FIHT with one step hard thresholding initialization is highly pessimistic when compared with the empirical observations, which suggests the possibility of improving this result. The leave-one-out technique [20, 39] may help us achieve this goal.

Recently, a line of research work has been devoted to the geometric analysis of non-convex optimization problems including dictionary learning [47], phase retrieval [48], low rank matrix sensing and matrix completion [1, 26, 27, 42], tensor completion

[25] and robust PCA [26]. It has been shown that the non-convex functions for those problems have well-behaved optimization landscape: all local minima are also globally optimal. Preliminary numerical results show that our projected gradient descent algorithm works equally well with random initialization, which suggests the geometric landscape of the objective function $F(\boldsymbol{Z})$ introduced in Chapter 3 may be similarly well-behaved.

REFERENCES

[1] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv:1605.07221*, 2016.

[2] J. Blanchard, J. Tanner, and K. Wei. CGIHT: Conjugate gradient iterative hard thresholding for compressed sensing and matrix completion. *Information and Inference*, 4(4):289–327, 2015.

[3] J. Blanchard, J. Tanner, and K. Wei. Conjugate gradient iterative hard thresholding: Observed noise stability for compressed sensing. *IEEE Transactions on Signal Processing*, 63(2):528–537, 2015.

[4] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[5] T. Blumensath and M. E. Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):298–309, 2010.

[6] Jian-Feng Cai, Suhui Liu, and Weiyu Xu. A fast algorithm for reconstruction of spectrally sparse signals in super-resolution. In *Wavelets and Sparsity XVI*, volume 9597, page 95970A. International Society for Optics and Photonics, 2015.

[7] Jian-Feng Cai, Xiaobo Qu, Weiyu Xu, and Gui-Bo Ye. Robust recovery of complex exponential signals from random Gaussian projections via low rank Hankel matrix reconstruction. *Applied and Computational Harmonic Analysis*, 41(2):470–490, 2016.

[8] Jian-Feng Cai, Tianming Wang, and Ke Wei. Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank Hankel matrix completion. *Applied and Computational Harmonic Analysis (to appear)*, 2017.

[9] Jian-Feng Cai, Tianming Wang, and Ke Wei. Spectral compressed sensing via projected gradient descent. *arXiv preprint arXiv:1707.09726*, 2017.

[10] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[11] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[12] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

[13] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[14] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[15] Y. Chen and Y. Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576–6601, 2014.

[16] Yudong Chen and Martin J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv:1509.03025*, 2015.

[17] Yuxin Chen and Emmanuel J Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5):822–883, 2017.

[18] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank. Sensitivity to basis mismatch in compressed sensing. *IEEE Transactions on Signal Processing*, 59(5):2182–2195, 2011.

[19] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

[20] Lijun Ding and Yudong Chen. The leave-one-out approach for matrix completion: Primal and dual analysis. *arXiv preprint arXiv:1803.07554*, 2018.

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

[22] M. Fazel, T. K. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications in system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

[23] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures and Algorithms*, 27(2):251–275, 2005.

[24] S. Foucart. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

[25] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. *arXiv:1503.02101*, 2015.

[26] Rong Ge, Chi Jin, and Yi Zhang. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv:1704.00708*, 2017.

[27] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *NIPS*, pages 2973–2981, 2016.

[28] D. Goldfarb and S. Ma. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2):183–210, 2011.

[29] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[30] M. Herman and T. Strohmer. General deviants: An analysis of perturbations in compressed sensing. *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Compressive Sensing*, 4(2):342–349, 2010.

[31] P. Jain, R. Meka, and I. Dhillon. Guaranteed rank minimization via singular value projection. In *Proceedings of the Neural Information Processing Systems Conference*, 2010.

[32] Kyong Hwan Jin and Jong Chul Ye. Annihilating filter-based low-rank Hankel matrix approach for image inpainting. *IEEE Transactions on Image Processing*, 24(11):3498–3511, 2015.

[33] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[34] R. Larsen. PROPACK - software for large and sparse SVD calculations, version 2.1. `http://sun.stanford.edu/~rmunk/PROPACK/`, April 2005.

[35] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv:1606.04933*, 2016.

[36] W. Liao. MUSIC for multidimensional spectral estimation: Stability and super-resolution. *IEEE Transactions on Signal Processing*, 63(23):6395–6406, 2015.

[37] W. Liao and A. Fannjiang. MUSIC for single-snapshot spectral estimation: Stability and super-resolution. *Applied and Computational Harmonic Analysis*, 40(1):33–67, 2016.

[38] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[39] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

[40] L. Mirsky. A trace inequality of John von Neumann. *Monatshefte für mathematik*, 79(4):303–306, 1975.

[41] Deanna Needell and Joel Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[42] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. *arXiv:1609.03240*, 2016.

[43] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin. Sparsity and compressed sensing in radar imaging. *Proceedings of the IEEE*, 98(6):1006–1020, 2010.

[44] Xiaobo Qu, Maxim Mayzel, Jian-Feng Cai, Zhong Chen, and Vladislav Orekhov. Accelerated NMR spectroscopy with low-rank reconstruction. *Angewandte Chemie International Edition*, 54(3):852–854, 2015.

[45] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[46] L. Schermelleh, R. Heintzmann, and H. Leonhardt. A guide to super-resolution fluorescence microscopy. *The Journal of Cell Biology*, 190(2):165–175, 2010.

[47] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[48] Ju Sun, Qing Qu, and John Wright. A geometrical analysis of phase retrieval. *Foundations of Computational Mathematics (to appear)*, 2017.

[49] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *IEEE Transactions on Information Theory*, 59(11):7465–7490, 2013.

[50] J. Tanner and K. Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

[51] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[52] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk. Beyond Nyquist: Efficient sampling of sparse bandlimited signals. *IEEE Transactions on Information Theory*, 56(1):520–544, 2010.

[53] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *arXiv:1507.03566*, 2015.

[54] B. Vandereycken. Low rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[55] K. Wei, J. F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix completion. *arXiv:1603.06610*, 2016.

[56] K. Wei, J. F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[57] Wei Xu and Sanzheng Qiao. A fast symmetric SVD algorithm for square Hankel matrices. *Linear Algebra and its Applications*, 428(2):550–563, 2008.

[58] Weiyu Xu, Jirong Yi, Soura Dasgupta, Jian-Feng Cai, Mathews Jacob, and Myung Cho. Separation-free super-resolution from compressed measurements is possible: an orthonormal atomic norm minimization approach. *arXiv preprint arXiv:1711.01396*, 2017.

[59] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. *arXiv:1605.07784*, 2016.

[60] J. Ying, H. Lu, Q. Wei, J.-F. Cai, D. Guo, J. Wu, Z. Chen, and X. Qu. Hankel matrix nuclear norm regularized tensor completion for N-dimensional exponential signals. *IEEE Transactions on Signal Processing*, 65(14):3702–3717, 2017.

[61] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *arXiv:1506.06081*, 2015.

[62] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv:1605.07051*, 2016.