Sparse Blind Deconvolution via Thresholded Wirtinger Flow

Mengting Chen Haitong Lan Ziping Zhao
School of Information Science and Technology
ShanghaiTech University
Shanghai, China 201210
{chenmt, lanht2026, zipingzhao}@shanghaitech.edu.cn

Abstract

Blind deconvolution is a classical problem arising in many signal processing and machine learning applications, where one aims to recover an unknown kernel and an unknown signal from their convolution. Since both the kernel and the signal are unknown, the problem is intrinsically ill-posed, and meaningful recovery is possible only under suitable structural assumptions. In this paper, we study sparse blind deconvolution, where the kernel is s_h-sparse and the signal admits an s_x-sparse representation in a known dictionary. Although sparsity is a natural and practically relevant prior, its interaction with the bilinear observation model creates substantial challenges for both computation and analysis. To address these challenges, we propose Thresholded Wirtinger Flow (ThWF), a simple and scalable iterative algorithm equipped with a sparse spectral initialization. For noiseless observations, we prove that the proposed initialization achieves exact support recovery with sample complexity $M \gtrsim s_h s_x$ and is close to the ground truth with $M \gtrsim s_h^2 + s_x^2$, up to logarithmic factors. Building on this initialization, we establish what is, to the best of our knowledge, the first algorithmic convergence guarantee for sparse blind deconvolution: ThWF converges linearly to the ground truth, up to the inherent scaling ambiguity, provided that $M \gtrsim s_h + s_x$ up to logarithmic factors. Our analysis further extends to noisy measurements, showing that ThWF is robust and contracts linearly up to a statistical error under a bounded noise level. Experiments on synthetic data and real-world image deblurring tasks corroborate the predicted linear convergence and phase transition behavior, and demonstrate the restoration performance of the proposed method.

1 Introduction

1.1 Motivation

Convolutional observations are ubiquitous in science and engineering, with applications ranging from medical imaging and seismology to wireless communications. In these settings, the observed data $\mathbf{f} \in \mathbb{C}^M$ is often modeled as the circular convolution of an unknown kernel $\mathbf{h} \in \mathbb{C}^M$ and an unknown signal $\mathbf{g} \in \mathbb{C}^M$:

$$\mathbf{f} = \mathbf{h} \circledast \mathbf{g}, \tag{1}$$

where $\circledast$ denotes length-M circular convolution. For example, in image deblurring, $\mathbf{f}$ is a blurry image, $\mathbf{h}$ corresponds to the point spread function, and $\mathbf{g}$ is the unknown sharp image to be recovered. Blind deconvolution aims to recover both $\mathbf{h}$ and $\mathbf{g}$ from their convolution $\mathbf{f}$. This problem is severely underdetermined, since the observation provides only M measurements for two unknown length-M vectors. Moreover, the bilinear convolutional structure introduces inherent non-identifiability, including scaling and shift ambiguities, and more generally, the redistribution of filtering effects

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

between the two factors. Therefore, meaningful recovery is impossible without additional structural assumptions [23, 7].

A common strategy for making blind deconvolution tractable is to impose subspace priors, assuming that the unknown components lie in known low-dimensional subspaces [3, 27, 31]. This line of work has led to important theoretical and algorithmic advances, but the subspace model can be restrictive when the active degrees of freedom are not known in advance. In many applications, the kernel itself is sparse, and the signal may be sparse after expansion over a known dictionary. For instance, in astronomical imaging or multipath channel estimation, the locations of the active kernel entries or the active signal coefficients are typically unknown and must be inferred from the data. This motivates sparse blind deconvolution under sparsity priors, where sparsity is imposed directly on the unknown components rather than encoded through a fixed low-dimensional subspace [20, 21, 27, 11].

In this paper, we study sparse blind deconvolution where the kernel $\mathbf{h} \in \mathbb{C}^M$ is s_h-sparse and the signal admits the representation $\mathbf{g} = \mathbf{C}\mathbf{x}^*$ over a known dictionary $\mathbf{C} \in \mathbb{C}^{M\times N}$, with an s_x-sparse coefficient vector $\mathbf{x} \in \mathbb{C}^N$. Our goal is to recover the sparse kernel $\mathbf{h}$ and the sparse coefficient vector $\mathbf{x}$ from a single observation $\mathbf{f}$, with both computational efficiency and provable algorithmic guarantees.

1.2 Related Work

Subspace priors. Blind deconvolution under subspace priors assumes that the unknown components lie in known low-dimensional subspaces. This setting has led to substantial theoretical progress. A convex lifting approach was pioneered in [3], where blind deconvolution is reformulated as a low-rank matrix recovery problem and exact recovery is established with $\mathcal{O}(N \log^2 M)$ samples. However, the resulting semidefinite program can be computationally expensive for large-scale problems. To improve scalability, subsequent work has developed nonconvex formulations. A regularized Wirtinger flow method with spectral initialization was proposed in [24], establishing linear convergence under comparable sample complexity. More recently, it was shown in [31] that vanilla Wirtinger flow can succeed through implicit regularization, preserving incoherence along the optimization trajectory and yielding linear convergence with sample complexity $\mathcal{O}(N \log^9 M)$.

Sparsity priors. Sparse blind deconvolution goes beyond the fixed-subspace setting by requiring the active supports of the unknown components to be identified from the data. A parallel line of work studies this problem under sparsity priors, where both the signal and the kernel are assumed to be sparse or sparse in known dictionaries. Alternating minimization methods with sparsity-enforcing projection steps were developed in [20, 21]. However, the projection step can be computationally intractable due to its combinatorial nature, and the accompanying theory focuses on RIP-based stable recovery rather than convergence of the proposed algorithm. The bisparse model was further studied in [11] by representing the kernel and signal coefficients through a hierarchically sparse Kronecker product and applying hard-thresholding pursuit. This approach provides stable recovery guarantees under RIP-type conditions, but it does not explicitly decouple the kernel and signal factors and does not establish a convergence theory for the iterative process.

Other variants. Several other variants of blind deconvolution have also been investigated. Multichannel blind deconvolution uses multiple observations to reduce the degrees of freedom in the ambiguity space and improve recovery robustness [38, 43, 32, 36, 22, 1, 12]. Blind spike deconvolution in the continuous domain focuses on recovering point sources from low-pass filtered data [6]. Blind deconvolution and demixing considers observations formed by superpositions of multiple signals convolved with distinct kernels [29, 30, 16, 9, 10]. Finally, a line of work has analyzed the landscape and geometry of sparse blind deconvolution and developed geometric methods for single-channel and multichannel settings [17, 45, 44, 18, 25, 26, 33, 34, 37].

1.3 Our Contributions

Despite the progress reviewed above, existing results for sparse blind deconvolution mainly establish identifiability or stable recovery guarantees, often under RIP-type conditions, while a convergence theory for a practical iterative algorithm remains largely unavailable. A central difficulty is that sparse blind deconvolution requires the algorithm to simultaneously identify the active supports of the unknown factors and refine their bilinearly coupled amplitudes. In this paper, we address this gap

by studying sparse blind deconvolution under the assumption that the kernel $\mathbf{h}$ is sparse and the signal admits a sparse representation $\mathbf{g} = \mathbf{C}\mathbf{x}^*$ in a known dictionary. We develop a thresholded Wirtinger flow method with sparse spectral initialization and prove its linear convergence to the ground truth up to the inherent scaling ambiguity. Our main contributions are summarized as follows:

- **Exact Support Recovery via Sparse Initialization:** We introduce a sparse spectral initialization scheme to estimate the supports of both the kernel $\mathbf{h}$ and the sparse coefficient vector $\mathbf{x}$. We prove that this scheme achieves exact support recovery if $M \gtrsim s_h s_x$ and provides an initial estimate within a small neighborhood of the ground truth with high probability, provided that $M \gtrsim s_h^2 + s_x^2$ up to logarithmic factors.
- **First Algorithmic Convergence Guarantee:** We propose Thresholded Wirtinger Flow (ThWF) and establish its convergence from the proposed initialization. We show that the iterates converge linearly to the ground truth up to the inherent scaling ambiguity, provided that $M \gtrsim (s_h + s_x) \operatorname{poly}\log M$. To the best of our knowledge, this is the first algorithmic linear convergence guarantee for sparse blind deconvolution with explicit sparse factors.
- **Theoretical Robustness to Noise:** We extend the analysis to noisy measurements and prove that ThWF remains stable under additive complex Gaussian noise. Specifically, with step size η, the estimation error contracts linearly up to an additive term of order $\mathcal{O}(\sqrt{M}\sigma)$, where σ is the noise standard deviation, provided that $\sigma \lesssim 1/\sqrt{M \log M}$.
- **Empirical Validation on Synthetic and Real Data:** Experiments on synthetic data validate the predicted phase transition behavior and linear convergence rate. We further demonstrate the practical performance of ThWF on real-world image deblurring tasks, where it restores sharp images under several representative blur kernels.

2 Technical Overview

The proposed method consists of two stages: a sparse spectral initialization that estimates the active supports, followed by ThWF that refines the estimates while preserving sparsity.

The primary challenge in sparse blind deconvolution is constructing an initial point within a basin of attraction where local contraction can be established. To address this, we leverage the structure of a matrix $\mathbf{Z}$ (defined in Section 3.1), whose expectation is aligned with the lifting rank-one matrix $\mathbf{h}\mathbf{x}^{\mathsf{H}}$. Motivated by [15], we recover the active supports of the kernel and the signal coefficients by analyzing the marginal energies, namely the row- and column-wise l_2-norms, of $\mathbf{Z}$. We prove that entries corresponding to the true supports exhibit substantially larger energy than off-support entries, enabling reliable support identification and effective search-space reduction. We then compute the leading singular vectors of the restricted matrix to obtain an initialization that is provably close to the ground truth up to the intrinsic scaling ambiguity.

For iterative refinement, we develop a scale-invariant Wirtinger gradient descent procedure that mitigates the imbalance induced by the scaling ambiguity. Specifically, each gradient update is scaled by the squared norm of the paired variable. After each gradient step, we apply a hard-thresholding operator that retains only the largest-magnitude entries, thereby enforcing sparsity throughout the iterations.

The key consequence of the sparse spectral initialization is that, once the correct supports are identified, the subsequent analysis of ThWF can be carried out entirely within the corresponding low-dimensional restricted subspace. This allows us to adapt the local geometry framework of [31] to the sparse setting. In particular, we show that within the basin of attraction, the empirical loss satisfies restricted strong convexity and smoothness properties, which implies linear contraction of the iterations to ground truth. The hard-thresholding step does not create the contraction itself; rather, it preserves the correct support structure so that the iterates remain within the region where the local geometric conditions continue to hold. To further control the statistical dependencies between the iterates and the random measurements, we incorporate a leave-one-out analysis, which guarantees uniform concentration along the entire optimization trajectory.

Notation. Throughout this paper, vectors by bold lowercase letters, e.g., $\mathbf{a}$, and matrices are denoted by bold uppercase letters, e.g., $\mathbf{A}$. Unless otherwise specified, all vectors are column vectors. For a vector $\mathbf{a}$, its transpose, conjugate transpose, and elementwise complex conjugate are denoted

by $\mathbf{a}^\top$, $\mathbf{a}^\mathsf{H}$, and $\mathbf{a}^*$, respectively. For a vector $\mathbf{a}$, its i-th entry is denoted by a_i; for a matrix $\mathbf{A}$, its i-th column is denoted by $\mathbf{a}_i$. For a vector $\mathbf{a}$, we use $\|\mathbf{a}\|_0$ for the number of nonzero entries, $\|\mathbf{a}\|$ for the Euclidean norm, and $\|\mathbf{a}\|_\infty$ for the infinity norm. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|$ denotes the spectral norm and $\|\mathbf{A}\|_F$ denotes the Frobenius norm. The Frobenius inner product of two matrices is $\langle \mathbf{A}, \mathbf{B} \rangle_F = \mathrm{Tr}(\mathbf{A}^\mathsf{H}\mathbf{B})$. For an index set $I \subseteq [K]$ where $[K] = \{1, \ldots, K\}$ denotes the ambient index set, $|I|$ denotes the cardinality of I and I^c denotes the complement $[K] \setminus I$. The symbols $\odot$ and $\otimes$ denote the Hadamard and Kronecker products, respectively. We use $\mathrm{Re}(\cdot)$ for the real part and $\mathbb{C}$ for the complex field. For two nonnegative functions $f(n)$ and $g(n)$, $f(n) \lesssim g(n)$, equivalently $f(n) = \mathcal{O}(g(n))$, means that $f(n) \le cg(n)$ for some absolute constant $c > 0$, and $f(n) \asymp g(n)$ means that $c_1 g(n) \le f(n) \le c_2 g(n)$ for some constants $c_1, c_2 > 0$. We use $\mathrm{poly}\log(M)$ to denote a polynomial in $\log M$. Finally, $\mathcal{CN}(0, \sigma^2)$ denotes the circularly symmetric complex Gaussian distribution whose real and imaginary parts are independent with zero mean and equal variance $\sigma^2/2$.

3 Sparse Blind Deconvolution

We now formally introduce the sparse blind deconvolution problem. Recall from Section 1.1 that the observation is modeled as

$$\mathbf{f} = \mathbf{h} \circledast \mathbf{C}\mathbf{x}^*, \tag{2}$$

where $\mathbf{h} \in \mathbb{C}^M$ is the unknown sparse kernel, $\mathbf{C} \in \mathbb{C}^{M \times N}$ is a known dictionary, and $\mathbf{x} \in \mathbb{C}^N$ is the unknown sparse coefficient vector. By applying the unitary discrete Fourier transform matrix $\mathbf{B}^\mathsf{H} \in \mathbb{C}^{M \times M}$ to both sides of (2), the convolution in the time domain is transformed into an elementwise product in the frequency domain:

$$\mathbf{y} = \mathbf{B}^\mathsf{H}\mathbf{f} = \sqrt{M}\left(\mathbf{B}^\mathsf{H}\mathbf{h}\right) \odot \left(\mathbf{B}^\mathsf{H}\mathbf{C}\mathbf{x}^*\right).$$

Define $\mathbf{A} := \sqrt{M}\mathbf{C}^\top\mathbf{B}^*$. The j-th entry of the measurements is

$$y_j = \mathbf{b}_j^\mathsf{H}\mathbf{h}\mathbf{x}^\mathsf{H}\mathbf{a}_j, \tag{3}$$

where $\mathbf{b}_j$ and $\mathbf{a}_j$ is the j-th column of $\mathbf{B}$ and $\mathbf{A}$, respectively. To account for practical scenarios, we further consider the noisy case, where the frequency-domain measurements are given by $\mathbf{y} = \left(\mathbf{B}^\mathsf{H}\mathbf{h}\right) \odot \left(\mathbf{A}^\top\mathbf{x}^*\right) + \mathbf{e}$, where $\mathbf{e} \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2\mathbf{I}_M\right)$ is the additive Gaussian noise.

Based on the above models, we estimate $\mathbf{h}$ and $\mathbf{x}$ by solving

$$\begin{aligned} \underset{\mathbf{h},\mathbf{x}}{\text{minimize}} \quad & \sum_{j=1}^{M} \left|\mathbf{b}_j^\mathsf{H}\mathbf{h}\,\mathbf{x}^\mathsf{H}\mathbf{a}_j - y_j\right|^2 \\ \text{subject to} \quad & \|\mathbf{h}\|_0 \le s_h,\ \|\mathbf{x}\|_0 \le s_x, \end{aligned} \tag{4}$$

where s_h and s_x denote the sparsity levels of the kernel $\mathbf{h}$ and the coefficient vector $\mathbf{x}$, respectively.

We solve (4) using a two-stage procedure. The first stage constructs a sparse spectral initialization that is close to the ground truth pair $(\mathbf{h}^\star, \mathbf{x}^\star)$. The second stage applies ThWF to refine the estimates while preserving sparsity and achieving linear convergence.

3.1 Sparse Spectral Initialization

Support Estimation. Define the adjoint matrix $\mathbf{Z} := \sum_{j=1}^{M} y_j\,\mathbf{b}_j\mathbf{a}_j^\mathsf{H}$. Under our measurement model, this matrix satisfies $\mathbb{E}[\mathbf{Z}] = \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}$, which suggests that the maximum absolute entries along the rows and columns of $\mathbf{Z}$ carry information about the supports of $\mathbf{h}^\star$ and $\mathbf{x}^\star$, respectively. Specifically, we estimate the support of $\mathbf{h}^\star$ by computing the row-wise maximum magnitudes $c_l = \max_{i \in [N]} |Z_{l,i}|$ and selecting the indices corresponding to the largest s_h values, denoting the resulting set by $\widehat{I}_h$. Similarly, we estimate the support of $\mathbf{x}^\star$ by computing the column-wise maximum magnitudes $r_i = \max_{l \in [M]} |Z_{l,i}|$ and selecting the indices corresponding to the largest s_x values, denoting the resulting set by $\widehat{I}_x$.

Spectral Initialization. Given the estimated supports $\widehat{I}_h$ and $\widehat{I}_x$, we form the restricted matrix $\mathbf{Z}_{\widehat{I}_h,\widehat{I}_x} = \sum_{j=1}^{M} y_j\,\mathbf{b}_{j,\widehat{I}_h}\mathbf{a}_{j,\widehat{I}_x}^\mathsf{H}$. Let σ_1, $\mathbf{p}$, and $\mathbf{q}$ denote the leading singular value and the associated

Algorithm 1 Sparse Spectral Initialization

Input: Observation $\mathbf{f}$, sensing matrices $\mathbf{C}$ and $\mathbf{B}$, sparsity levels s_h and s_x.
Compute $\mathbf{Z} \leftarrow \sum_{j=1}^{M} y_j \, \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}}$; $c_l = \max_i |Z_{l,i}|$ and $r_i = \max_l |Z_{l,i}|$.
$(\widehat{I}_h, \widehat{I}_x) \leftarrow$ indices of the largest s_h values of $\{c_l\}_{l=1}^{M}$ and largest s_x values of $\{r_i\}_{i=1}^{N}$.
Compute $(\sigma_1, \mathbf{p}, \mathbf{q})$, the leading singular triple of $\mathbf{Z}_{\widehat{I}_h, \widehat{I}_x}$.
Set $\mathbf{h}^0_{\widehat{I}_h} \leftarrow \sqrt{\sigma_1}\mathbf{p}$ and $\mathbf{h}^0_{\widehat{I}_h^c} \leftarrow \mathbf{0}$.
Set $\mathbf{x}^0_{\widehat{I}_x} \leftarrow \sqrt{\sigma_1}\mathbf{q}$ and $\mathbf{x}^0_{\widehat{I}_x^c} \leftarrow \mathbf{0}$.
Output: Initial estimates $\mathbf{h}^0$ and $\mathbf{x}^0$.

Algorithm 2 Thresholded Wirtinger Flow (ThWF)

Input: $\mathbf{y}$, $\{\mathbf{a}_j\}, \{\mathbf{b}_j\}$, sparsity levels s_h, s_x, step size η, initial $(\mathbf{h}^0, \mathbf{x}^0)$ from Algorithm 1.
for $k = 0, 1, 2, \ldots$ until convergence **do**
    $\mathbf{h}^{k+1} \leftarrow \mathcal{T}_{s_h}(\mathbf{h}^k - \frac{\eta}{\|\mathbf{x}^k\|^2} \nabla_{\mathbf{h}} f(\mathbf{x}^k, \mathbf{h}^k))$.
    $\mathbf{x}^{k+1} \leftarrow \mathcal{T}_{s_x}(\mathbf{x}^k - \frac{\eta}{\|\mathbf{h}^k\|^2} \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{h}^k))$.
end for
Output: Estimated sparse vectors $\widehat{\mathbf{h}}$, $\widehat{\mathbf{x}}$.

left and right singular vectors of $\mathbf{Z}_{\widehat{I}_h, \widehat{I}_x}$, respectively. The spectral initialization is defined as

$$\mathbf{h}^0_{\widehat{I}_h} = \sqrt{\sigma_1}\mathbf{p}, \qquad \mathbf{x}^0_{\widehat{I}_x} = \sqrt{\sigma_1}\mathbf{q}, \tag{5}$$

with $\mathbf{h}^0_{\widehat{I}_h^c} = \mathbf{0}$ and $\mathbf{x}^0_{\widehat{I}_x^c} = \mathbf{0}$.

The overall initialization procedure is summarized in Algorithm 1.

3.2 Thresholded Wirtinger Flow

We develop the ThWF algorithm to solve (4), where each iteration consists of a normalized Wirtinger gradient step followed by hard thresholding. A key issue in blind deconvolution is the intrinsic scaling ambiguity: for any nonzero $\alpha \in \mathbb{C}$, the pair $\left(\frac{1}{\alpha^*}\mathbf{h}, \alpha\mathbf{x}\right)$ produces the same measurements as $(\mathbf{h}, \mathbf{x})$ under the model (2). Consequently, the magnitudes of the partial gradients can become unbalanced under rescaling, scaling as $|\alpha|^{\pm 2}$; see the gradients in (9).

To address this imbalance, we employ a scale-invariant gradient descent scheme that normalizes each partial gradient by the squared norm of the paired variable. Define the objective in (4) as $f(\mathbf{x}, \mathbf{h})$. At iteration k, the intermediate estimates are computed as

$$\mathbf{h}^{k+\frac{1}{2}} = \mathbf{h}^k - \frac{\eta}{\left\|\mathbf{x}^k\right\|^2} \nabla_{\mathbf{h}} f\left(\mathbf{x}^k, \mathbf{h}^k\right), \quad \mathbf{x}^{k+\frac{1}{2}} = \mathbf{x}^k - \frac{\eta}{\left\|\mathbf{h}^k\right\|^2} \nabla_{\mathbf{x}} f\left(\mathbf{x}^k, \mathbf{h}^k\right), \tag{6}$$

where $\eta > 0$ is the step size. This normalization makes the update compatible with the scaling equivalence class and yields balanced progress for the two factors.

Finally, sparsity is enforced through the hard-thresholding operator $\mathcal{T}_s(\cdot)$, which retains the largest s entries in magnitude and sets the remaining entries to zero:

$$\mathbf{h}^{k+1} = \mathcal{T}_{s_h}\left(\mathbf{h}^{k+\frac{1}{2}}\right), \qquad \mathbf{x}^{k+1} = \mathcal{T}_{s_x}\left(\mathbf{x}^{k+\frac{1}{2}}\right). \tag{7}$$

The complete procedure is summarized in Algorithm 2.

4 Main Results

Before stating our theoretical guarantees, we introduce a distance metric that accounts for the inherent scaling ambiguity. For any estimate $\mathbf{z}^k = \left[\mathbf{h}^{k\mathsf{H}}, \mathbf{x}^{k\mathsf{H}}\right]^{\mathsf{H}}$, define the distance to the ground truth pair $\mathbf{z}^\star = \left[\mathbf{h}^{\star\mathsf{H}}, \mathbf{x}^{\star\mathsf{H}}\right]^{\mathsf{H}}$ as:

$$\operatorname{dist}\left(\mathbf{z}^k, \mathbf{z}^\star\right) = \min_{\alpha \in \mathbb{C}} \sqrt{\left\| \frac{1}{\alpha^*}\mathbf{h}^k - \mathbf{h}^\star \right\|^2 + \left\|\alpha\mathbf{x}^k - \mathbf{x}^\star\right\|^2}, \tag{8}$$

where α represents the scaling factor that accounts for the scaling ambiguity in the problem.

For the model in (3), we assume the following conditions hold.

Assumption 1. *The columns $\{\mathbf{a}_j\}$ of $\mathbf{A}$ are independent and identically distributed complex Gaussian vectors, i.e., $\mathbf{a}_j \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$.*

While the Gaussian assumption on $\mathbf{A}$ is adopted for theoretical tractability, our method demonstrates empirical robustness across various structured matrices. For instance, in image deblurring where $\mathbf{A}$ represents a wavelet basis, or in communications with Hadamard-type matrices, the proposed approach maintains high performance. Extending theoretical guarantees to these realistic settings remains a subject for future work.

Assumption 2. *The ground-truth kernel $\mathbf{h}^\star$ is μ-spectrally flat. Specifically, the incoherence parameter μ is defined as the smallest constant such that*

$$\max_{1\le j\le M} |\mathbf{b}_j^{\mathsf{H}}\mathbf{h}^\star| \le \frac{\mu}{\sqrt{M}}\|\mathbf{h}^\star\|, \quad 1 \le \mu \le \sqrt{M}.$$

Assumption 2 is on the spectral flatness, which ensures that the energy of $\mathbf{h}^\star$ is well-distributed across frequency bins, preventing the problem from becoming ill-conditioned. A small μ signifies superior spectral balance and is a standard assumption in the blind deconvolution literature, see [3, 2, 31, 21].

Assumption 3. *There exist constants $C_h, C_x > 0$ such that the non-zero entries of $\mathbf{h}^\star$ and $\mathbf{x}^\star$ satisfy*

$$\min_{l\in I_h} |h_l^\star| \ge \frac{C_h}{\sqrt{s_h}}\|\mathbf{h}^\star\| \quad \textit{and} \quad \min_{i\in I_x} |x_i^\star| \ge \frac{C_x}{\sqrt{s_x}}\|\mathbf{x}^\star\|.$$

In the noisy case, we further require these entries to exceed a noise-dependent threshold:

$$\min_{l\in I_h} |h_l^\star| \min_{i\in I_x} |x_i^\star| \ge C'\sigma\sqrt{\log(MN)}\|\mathbf{h}^\star\|\|\mathbf{x}^\star\|,$$

for constants $C' > 0$ and where σ is the noise standard deviation.

Assumption 3 ensures that the true supports I_h and I_x can be reliably identified by hard-thresholding. In the noisy setting, this condition implies that the signal's marginal energy dominate the noise-induced bias, which is essential for maintaining the stability of the iterative trajectory.

The following theorem guarantees that Algorithm 1 recovers the true supports and provides a sufficiently accurate initial guess.

Theorem 1. *Under the Assumptions 1, 2, and 3. Suppose the number of measurements obeys $M \ge C_0 s_h s_x \log M$ for some constant $C_0 > 0$. Algorithm 1 achieves exact support recovery with probability at least $1 - \mathcal{O}\left(M^{-10}\right)$.*

Moreover, under the Assumptions 1 and 2. Suppose $M \ge C_0'\left(s_h^2 + \mu^4 s_x^2\right)\log M$ for some constant $C_0' > 0$. The initial point $\mathbf{z}^0 = \left[\mathbf{h}^{0\mathsf{H}}, \mathbf{x}^{0\mathsf{H}}\right]^{\mathsf{H}}$ generated by Algorithm 1 is close to the ground truth $\mathbf{z}^\star = \left[\mathbf{h}^{\star\mathsf{H}}, \mathbf{x}^{\star\mathsf{H}}\right]^{\mathsf{H}}$ with probability at least $1 - \mathcal{O}\left(M^{-1}\right)$, i.e.,

$$\operatorname{dist}\left(\mathbf{z}^0, \mathbf{z}^\star\right) \le \lambda\left\|\mathbf{z}^\star\right\|, \quad 0 < \lambda < 1.$$

The two guarantees in Theorem 1 have distinct sample requirements. Under the minimum-amplitude condition, the entrywise support criterion yields the multiplicative requirement $M \gtrsim s_h s_x \log M$ for exact support recovery. The second part of the theorem gives a separate sufficient bound $(s_h^2 + \mu^4 s_x^2)\log M$ for accurate initialization without the minimum-amplitude condition. Although μ characterizes the kernel spectrum, it controls the fluctuations in estimating the signal coefficients, explaining the asymmetric $\mu^4 s_x^2$ term.

4.1 Noiseless Case

We first establish the linear convergence of ThWF in the absence of noise, when initialized by the proposed sparse spectral method.

Theorem 2. *Suppose Assumptions 1, 2, and 3 hold. Let $\mathbf{z}^0$ be the initial estimate generated by Algorithm 1, and let $\{\mathbf{z}^k\}_{k\geq 0}$ be the sequence generated by Algorithm 2 from the noiseless frequency-domain measurements $\mathbf{y}$. Choose the step size $\eta \leq 1$. If $M \geq C_1\mu^2(s_h+s_x)\log^9 M$ for some constant $C_1 > 0$, then, with probability at least $1-\mathcal{O}\big(M^{-10}+e^{-(s_h+s_x)\log M}\big)$, the iterates contract linearly:*

$$\operatorname{dist}\left(\mathbf{z}^{k+1},\mathbf{z}^\star\right) \leq (1-\eta/16)\operatorname{dist}\left(\mathbf{z}^k,\mathbf{z}^\star\right), \quad \forall k \geq 0.$$

Theorem 2 shows that sparse spectral initialization places the iterates in a contraction region where ThWF converges geometrically to the ground truth, modulo the inherent scaling ambiguity. The refinement condition is nearly linear in the total sparsity level $s_h + s_x$, up to logarithmic factors and the spectral flatness parameter μ. To the best of our knowledge, this is the first algorithmic linear convergence guarantee for sparse blind deconvolution with explicit sparse factors.

The factor μ^2 multiplies both sparsity levels because our analysis controls the local geometry of the concatenated variable $\mathbf{z} = [\mathbf{h}^{\mathsf{H}}, \mathbf{x}^{\mathsf{H}}]^{\mathsf{H}}$ using uniform worst-case bounds. This symmetric dependence is a feature of the present proof and is not established as an intrinsic requirement of the problem. Tracking coupled errors for $\mathbf{h}$ and $\mathbf{x}$, with each update depending explicitly on both current errors, may yield a sharper, asymmetric dependence on μ. Establishing such bounds requires a different convergence analysis and is left for future work.

4.2 Noisy Case

We next consider noisy observations. The following theorem shows that ThWF remains stable under additive complex Gaussian noise.

Theorem 3. *Suppose Assumptions 1, 2, and 3 hold. Let $\mathbf{z}^0$ be the initial estimate generated by Algorithm 1, and let $\{\mathbf{z}^k\}_{k\geq 0}$ be the sequence generated by Algorithm 2 from the noisy frequency-domain measurements $\mathbf{y}_{\text{obs}}$. Choose the step size $\eta \leq 1$. Suppose the noise level satisfies $\sigma \leq C_\sigma\|\mathbf{x}^\star\|\|\mathbf{h}^\star\|/\sqrt{M\log M}$ for some constant $C_\sigma > 0$, and suppose the sample complexity condition $M \geq C_2\mu^2(s_h+s_x)\log^9 M$ holds for some constant $C_2 > 0$. Then, with probability at least $1-\mathcal{O}\big(M^{-10}+e^{-(s_h+s_x)\log M}\big)$, the iterates satisfy*

$$\operatorname{dist}^2\left(\mathbf{z}^{k+1},\mathbf{z}^\star\right) \leq (1-\eta/16)\operatorname{dist}^2\left(\mathbf{z}^k,\mathbf{z}^\star\right) + \mathcal{O}\left(\eta M\sigma^2\right), \quad \forall k \geq 0.$$

Theorem 3 shows that the linear contraction of ThWF persists in the presence of additive noise, up to a noise-dependent error floor. In particular, iterating the recursion yields a limiting squared error of order $\mathcal{O}(M\sigma^2)$, up to constants and logarithmic factors. The condition $\sigma \lesssim \|\mathbf{x}^\star\|\|\mathbf{h}^\star\|/\sqrt{M\log M}$ is mainly imposed to guarantee reliable sparse spectral initialization. Indeed, the initialization is based on marginal energies obtained from squared noisy observations, which introduces a positive bias of order σ^2 per measurement. Accumulated over M measurements, this bias is of order $\sigma^2 M$ and may obscure the on-support and off-support energy gap when the SNR is too low. Developing a debiased initialization that relaxes this requirement is an interesting direction for future work.

5 Numerical Experiments

We conduct experiments on both synthetic and real-world data to validate the theoretical findings and the effectiveness of our proposed algorithm.[1] All experiments were conducted using MATLAB 2025b on a computer with a 3.3 GHz 12-Core Intel Xeon W processor and 48 GB RAM.

5.1 Synthetic Data

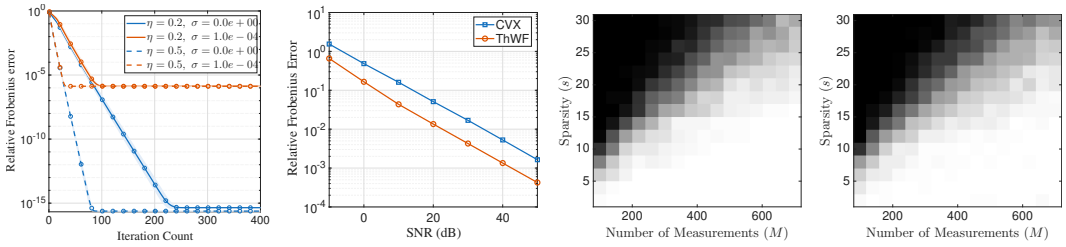
For each experimental configuration, we perform 100 independent Monte Carlo trials. In each trial, the ground truth pair $(\mathbf{h}^\star, \mathbf{x}^\star)$ is generated with i.i.d. standard complex Gaussian entries restricted to the support, followed by normalization to unit ℓ_2-norm. The subspace matrix $\mathbf{C}$ is constructed as described in Section 3. We terminate the algorithm when either the relative Frobenius error

$$\Delta^k = \left\|\mathbf{h}^k\mathbf{x}^{k\mathsf{H}} - \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\right\|_F / \left\|\mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\right\|_F$$

[1] Our code is open access at `https://anonymous.4open.science/r/ThWF-DF13/README.md`

falls below 10^{-8}, or the maximum number of iterations $k_{\max} = 500$ is reached. Since $\left\|\mathbf{h}^k\mathbf{x}^{k\mathsf{H}} - \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\right\|_F \lesssim \operatorname{dist}\left(\mathbf{z}^k, \mathbf{z}^\star\right)$, we directly use the relative Frobenius error without resolving the scaling ambiguity.

We first examine the linear convergence behavior of the proposed ThWF algorithm (Algorithm 2) initialized via the sparse spectral method (Algorithm 1). We fix the signal dimension to $N = 600$ and set the number of measurements to $M = N$. The sparsity levels are chosen as $s_h = s_x = 10$. We vary the step size $\eta \in \{0.2, 0.5\}$ and the noise standard variance $\sigma \in \left\{0, 10^{-4}\right\}$. Figure 1(a) plots the mean relative error versus the iteration count across trials. The shaded regions indicate the standard deviation, providing a robust characterization of variability. The results show a clear linear decay of the relative Frobenius error on the semilogarithmic scale, while the presence of noise leads to a higher statistical error. These observations are consistent with the linear convergence rate established in Theorem 2 and the additive noise effect characterized in Theorem 3.



(a) (b) (c) (d)

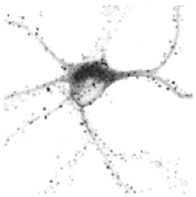
Figure 1: Synthetic data experiments. (a) Linear convergence; (b) SNR; (c) Phase transition in the noiseless case; (d) Phase transition when $\sigma = 10^{-4}$.

We then evaluate the robustness of the proposed ThWF algorithm against noise. We consider noisy observations with additive Gaussian noise $\mathbf{e} \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2\mathbf{I}_M\right)$. The signal-to-noise ratio (SNR) in decibels is defined as $\text{SNR (dB)} = 20\log_{10}\left(\|\mathbf{y}\| / \|\mathbf{e}\|\right)$. We set $M = 10N$, $N = 60$, and $s_h = s_x = 6$, with the step size is $\eta = 0.5$. Figure 1(b) reports the average final relative Frobenius error Δ^{final} over 100 independent trials for ThWF, alongside the mean error over 5 trials for the convex baseline [10, 28], which is solved via CVX, a package for specifying and solving convex programs [14, 13]. We adopt the convex lifting approach of [10, 28] as our primary baseline. The literature on sparse blind deconvolution is relatively limited, and existing methods either focus on subsampled convolution models [20] or do not explicitly separate the signal and kernel during recovery [11]. The empirical behaviors in Figure 1(b) are closely aligned with the theoretical limit established in Theorem 3. Specifically, as the iteration $k \to \infty$, the squared distance contracts to a statistical error $\mathcal{O}\left(\sigma^2 M\right)$, implying that the relative error scales as $\Delta^{\text{final}} \propto \frac{\sigma\sqrt{M}}{\|\mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\|_F}$. Given that the linear SNR satisfies $\text{SNR} \propto \frac{\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2}{M\sigma^2}$, substituting $\sigma \propto \frac{1}{\sqrt{M \cdot \text{SNR}}}$ into the relative error expression yields the theoretical scaling $\Delta^{\text{final}} \propto \frac{1}{\sqrt{\text{SNR}}}$, or equivalently, $\Delta^{\text{final}} \propto 10^{-\text{SNR (dB)}/20}$. This relationship explains the linear decay observed in the semilogarithmic plot. As shown in Figure 1(b), ThWF demonstrates superior noise robustness and achieves lower recovery errors compared to the convex lifting method [10, 28] across all SNR regimes.

Finally, we next investigate the sample complexity of the proposed algorithm by characterizing the phase transition behavior in the (M, s)-plane, where for simplicity we set $s = s_h = s_x$. Following the experimental setup, we set $M = N$ and vary $M \in \{100, 140, \cdots, 700\}$ and $s \in \{2, 4, \ldots, 30\}$. A trial is declared successful if the final relative error satisfies $\Delta^{\text{final}} < 10^{-3}$. The step size is set to $\eta = 0.5$. Figure 1(c) presents the phase transition heat map without noise and Figure 1(d) is with $\sigma = 10^{-4}$, where white indicates perfect recovery (success rate = 1) and black indicates failure (success rate = 0). The success rate exhibits a clear transition from complete recovery to complete failure as the sparsity s increases or the number of measurements M decreases.

5.2 Image Deblurring

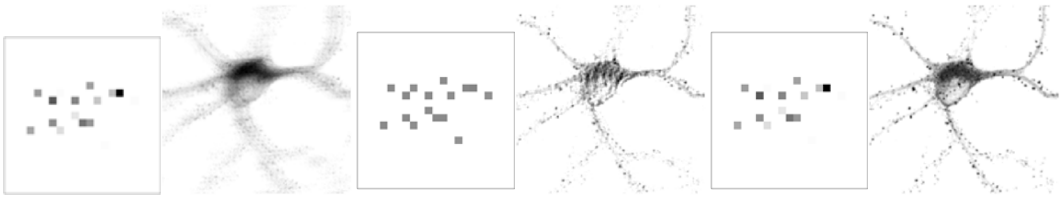
We further evaluate the practical performance of the proposed ThWF algorithm on a biological image deblurring task using a 160×160 ground-truth neuron image (courtesy of the Paul De Koninck Lab



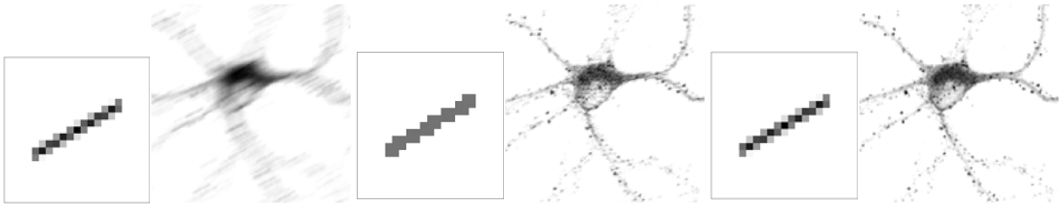
(a) Ground truth $\mathbf{g}^\star$

Method	BG kernel	Motion kernel	Gaussian kernel
CVX [10]	25.82	31.48	28.69
ThWF	**111.10**	**83.40**	**40.18**

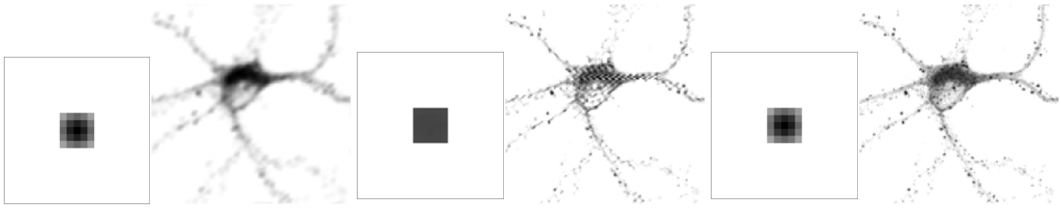
(b) PSNR (dB) comparison for the neuron image.



(c) BG $\mathbf{h}^\star$ (d) Observation $\mathbf{f}$ (e) $\widehat{\mathbf{h}}$ by [10] (f) $\widehat{\mathbf{g}}$ by [10] (g) $\widehat{\mathbf{h}}$ by ThWF (h) $\widehat{\mathbf{g}}$ by ThWF



(i) Motion $\mathbf{h}^\star$ (j) Observation $\mathbf{f}$ (k) $\widehat{\mathbf{h}}$ by [10] (l) $\widehat{\mathbf{g}}$ by [10] (m) $\widehat{\mathbf{h}}$ by ThWF (n) $\widehat{\mathbf{g}}$ by ThWF



(o) Gauss. $\mathbf{h}^\star$ (p) Observation $\mathbf{f}$ (q) $\widehat{\mathbf{h}}$ by [10] (r) $\widehat{\mathbf{g}}$ by [10] (s) $\widehat{\mathbf{h}}$ by ThWF (t) $\widehat{\mathbf{g}}$ by ThWF

Figure 2: Sparse blind deconvolution results for a neuron image.

[8], Figure 2(a)). The image is represented such that white pixels correspond to 0 and black pixels to 1. Since the image exhibits natural sparsity in the pixel domain, we employ the identity dictionary $\mathbf{C} = \mathbf{I}$. The algorithm is tested against three distinct blurring kernels, each originally defined on a 9×15 grid: (i) a Bernoulli-Gaussian (BG) kernel with variance 2 and activation probability 0.1; (ii) a motion kernel; and (iii) a Gaussian kernel that is truncated to its central 5×5 region with the remaining entries set to zero. For simplicity, we assume the kernel support is known a priori, while both the support and the exact number of non-zero entries s_x of the image are unknown. For the ThWF algorithm, the sparsity levels are set to $s_x = 5400, 5500$, and 4700 for the BG, motion, and Gaussian cases, respectively, with corresponding step sizes $\eta \in \{0.1, 0.05, 0.01\}$. We terminate the iterations when the relative change in the objective function satisfies $|f(\mathbf{z}^{k+1}) - f(\mathbf{z}^k)|/f(\mathbf{z}^k) < 10^{-7}$.

The recovery results are displayed in Figure 2. The second row (Figure 2(c)–2(h)) displays the results for the BG kernel, the third row (Figure 2(i)–2(n)) for the motion kernel, and the fourth row (Figure 2(o)–2(t)) for the Gaussian kernel. In each case, we compare the recovery of both the kernel $\widehat{\mathbf{h}}$ and the image $\widehat{\mathbf{g}}$ against the convex lifting method [10]. The quantitative comparison in terms of peak signal-to-noise ratio (PSNR) is summarized in Table 2(b). As observed in Figure 2, the proposed ThWF method consistently yields superior restoration of both the blurring kernels and the latent neuron images compared to the convex approach. It is observed that our method successfully captures fine structural details of the neurons with high fidelity across all tested kernel types. Additional image deblurring results are provided in Appendix F, including comparisons with the deep learning method SelfDeblur in Appendix F.2.

6 Conclusions and Discussions

In this paper, we have studied sparse blind deconvolution using a two-stage algorithm that first performs sparse spectral initialization to estimate the supports of the ground-truth kernel and sparse coefficient vector, and then recovers the unknowns via thresholded Wirtinger flow. Theoretically, we have proved that the proposed initialization achieves exact support recovery, provided that the nonzero entries of the ground-truth kernel and coefficient vector are large. Building on this initialization, we have established that the iterates converge linearly to the ground truth up to the inherent scaling ambiguity, even in the presence of additive noise. To the best of our knowledge, this is the first work to provide an algorithmic convergence guarantee for sparse blind deconvolution. Experiments have validated these theoretical findings.

Several interesting directions remain open. First, the sparse spectral initialization requires a relatively high SNR to avoid noise-induced bias. Developing a debiased initialization could extend the algorithm to lower SNR regimes. Second, our analysis focuses on the case where the dictionary $\mathbf{C}$ has i.i.d. Gaussian entries. Extending the guarantees to structured dictionaries (e.g., random demodulators) is an important next step. Finally, our proof justifies local convergence by reducing the sparse problem to a low-dimensional non-sparse counterpart via support recovery. A significant future direction is to develop a more intrinsic global analysis that bypasses the need to fix the support a priori, potentially by leveraging restricted isometry property arguments to achieve a near-optimal sample complexity.

Acknowledgments and Disclosure of Funding

This work was supported by the National Science and Technology Major Project under Grant 2025ZD1009908.

References

[1] A. Ahmed. Blind deconvolution using modulated inputs. *IEEE Transactions on Signal Processing*, 68:374–387, 2019.

[2] A. Ahmed and L. Demanet. Leveraging diversity and sparsity in blind deconvolution. *IEEE Transactions on Information Theory*, 64(6):3975–4000, 2018.

[3] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

[4] A. S. Bandeira and R. van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

[5] E. J. Candés, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[6] Y. Chi. Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):782–794, 2016.

[7] S. Choudhary and U. Mitra. Fundamental limits of blind deconvolution part I: Ambiguity kernel. *arXiv preprint arXiv:1411.3810*, 2014.

[8] P. De Koninck. Greenspine gallery: Neuron expressing GFP-CaMKII and immunolabelling of excitatory synapses. `https://greenspine.ca/neuron-expressing-gfp-camkii-green-and-immunolabelling-excitatory-synapses`, 2026.

[9] J. Dong and Y. Shi. Nonconvex demixing from bilinear measurements. *IEEE Transactions on Signal Processing*, 66(19):5152–5166, 2018.

[10] A. Flinth. Sparse blind deconvolution and demixing through $\ell_{1,2}$-minimization. *Advances in Computational Mathematics*, 44:1–21, 2018.

[11] A. Flinth, I. Roth, and G. Wunder. Bisparse blind deconvolution through hierarchical sparse recovery. *Advances in Computational Mathematics*, 51(6):58, 2025.

[12] Y. Gao, X. Lin, and M. K. Ng. Blind deconvolution for multiple observed images with missing values. *Pacific Journal of Optimization*, 19(1):69–96, 2023.

[13] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. 2008.

[14] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, 2014.

[15] G. Jagatap and C. Hegde. Sample-efficient algorithms for recovering structured signals from magnitude-only measurements. *IEEE Transactions on Information Theory*, 65(7):4434–4456, 2019.

[16] P. Jung, F. Krahmer, and D. Stöger. Blind demixing and deconvolution at near-optimal rate. *IEEE Transactions on Information Theory*, 64(2):704–727, 2018.

[17] H.-W. Kuo, Y. Lau, Y. Zhang, and J. Wright. Geometry and symmetry in short-and-sparse deconvolution. In *International Conference on Machine Learning*, pages 3570–3580, 2019.

[18] Y. Lau, Q. Qu, H.-W. Kuo, P. Zhou, Y. Zhang, and J. Wright. Short and sparse deconvolution – a geometric approach. In *International Conference on Learning Representations*, 2020.

[19] K. Lee and M. Junge. RIP-like properties in subsampled blind deconvolution. *arXiv preprint arXiv:1511.06146*, 2015.

[20] K. Lee, Y. Li, M. Junge, and Y. Bresler. Stability in blind deconvolution of sparse signals and reconstruction by alternating minimization. In *Proceedings of the IEEE International Conference on Sampling Theory and Applications*, pages 158–162, 2015.

[21] K. Lee, Y. Li, M. Junge, and Y. Bresler. Blind recovery of sparse signals from subsampled convolution. *IEEE Transactions on Information Theory*, 63(2):802–821, 2017.

[22] K. Lee, N. Tian, and J. Romberg. Fast and guaranteed blind multichannel deconvolution under a bilinear system model. *IEEE Transactions on Information Theory*, 64(7):4792–4818, 2018.

[23] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1964–1971, 2009.

[24] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and Computational Harmonic Analysis*, 47(3):893–934, 2019.

[25] Y. Li and Y. Bresler. Global geometry of multichannel sparse blind deconvolution on the sphere. In *Advances in Neural Information Processing Systems*, volume 31, pages 1132–1143, 2018.

[26] Y. Li and Y. Bresler. Multichannel sparse blind deconvolution on the sphere. *IEEE Transactions on Information Theory*, 65(11):7415–7436, 2019.

[27] Y. Li, K. Lee, and Y. Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on Information Theory*, 62(7):4266–4275, 2016.

[28] S. Ling and T. Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Probl.*, 31(11):115002, 2015.

[29] S. Ling and T. Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017.

[30] S. Ling and T. Strohmer. Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing. *Information and Inference: A Journal of the IMA*, 8(1):1–49, 2019.

[31] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 2019.

[32] S. Makni, P. Ciuciu, J. Idier, and J.-B. Poline. Joint detection-estimation of brain activity in functional MRI: a multichannel deconvolution solution. *IEEE Transactions on Signal Processing*, 53(9):3488–3502, 2005.

[33] Q. Qu, X. Li, and Z. Zhu. A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. *Advances in neural information processing systems*, 32, 2019.

[34] Q. Qu, X. Li, and Z. Zhu. Exact recovery of multichannel sparse blind deconvolution via gradient descent. *SIAM Journal on Imaging Sciences*, 13(3):1630–1652, 2020.

[35] D. Ren, K. Zhang, Q. Wang, Q. Hu, and W. Zuo. Neural blind deconvolution using deep priors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3338–3347, 2020.

[36] J. Romberg, N. Tian, and K. Sabra. Multichannel blind deconvolution using low rank recovery. In *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XI*, volume 8750, pages 108–113, 2013.

[37] L. Shi and Y. Chi. Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently. *IEEE Transactions on Information Theory*, 67(7):4784–4811, 2021.

[38] F. Sroubek and J. Flusser. Multichannel blind iterative image restoration. *IEEE Transactions on Image Processing*, 12(9):1094–1106, 2003.

[39] M. Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2005.

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

[41] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[42] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[43] L. Zhang, A. Cichocki, and S. Amari. Multichannel blind deconvolution of nonminimum-phase systems using filter decomposition. *IEEE Transactions on Signal Processing*, 52(5):1430–1442, 2004.

[44] Y. Zhang, H.-W. Kuo, and J. Wright. Structured local minima in sparse blind deconvolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, 2018.

[45] Y. Zhang, Y. Lau, H.-W. Kuo, S. Cheung, A. Pasupathy, and J. Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4902, 2017.

Contents

A Computations **14**

A.1 Expectation of $\mathbf{Z}$ . . . 14

A.2 Gradients . . . 14

B Proof of Theorem 1 **14**

B.1 Exact Support Recovery . . . 14

B.2 Closeness to Ground Truth . . . 16

C Proof of Theorem 2 **19**

C.1 Local geometry . . . 19

C.1.1 Proof of Lemma 7 . . . 20

C.2 Error contraction of ThWF . . . 21

Base Case ($k = 0$): . . . 22

Inductive Step: . . . 23

D Proof of Theorem 3 **23**

D.1 Modified sparse spectral initialization . . . 23

D.2 Modified local geometry . . . 24

Modified proof of Lemma 9: . . . 24

D.3 Modified error contraction . . . 25

Modified proof of Lemma 10: . . . 25

E Supporting Lemmas **26**

E.1 Basics . . . 26

E.2 Concentration inequalities . . . 26

F More Image Deblurring Experiments **32**

F.1 Structured Dictionaries and Comparison with Convex Lifting . . . 32

Haar wavelet matrix case: . . . 32

DCT matrix case: . . . 32

F.2 Comparison with the Deep Learning Method SelfDeblur . . . 36

A Computations

A.1 Expectation of Z

The j-th element of the noisy observation in the Fourier domain is

$$y_j = \mathbf{b}_j^{\mathsf{H}} \mathbf{h}^\star \mathbf{x}^{\star\mathsf{H}} \mathbf{a}_j + e_j.$$

The expectation of $\mathbf{Z}$ is

$$\begin{aligned}\mathbb{E}\left[\mathbf{Z}\right] &= \sum_{j=1}^{M} \mathbb{E}\left[\mathbf{b}_j^{\mathsf{H}} \mathbf{h}^\star \mathbf{x}^{\star\mathsf{H}} \mathbf{a}_j \cdot \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} + e_j \cdot \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}}\right] \\ &= \sum_{j=1}^{M} \mathbb{E}\left[\mathbf{a}_j \mathbf{a}_j^{\mathsf{H}}\right] \mathbf{h}^\star \mathbf{x}^{\star\mathsf{H}} \mathbf{b}_j \mathbf{b}_j^{\mathsf{H}} = \mathbf{h}^\star \mathbf{x}^{\star\mathsf{H}}.\end{aligned}$$

A.2 Gradients

The objective function is

$$f\left(\mathbf{h}, \mathbf{x}\right) = \sum_{j=1}^{M} \left|\mathbf{b}_j^{\mathsf{H}} \mathbf{h} \mathbf{x}^{\mathsf{H}} \mathbf{a}_j - y_j\right|^2.$$

Denote $r_j = \mathbf{b}_j^{\mathsf{H}} \mathbf{h} \mathbf{x}^{\mathsf{H}} \mathbf{a}_j - y_j$. We have $r_j^* = \mathbf{b}_j^{\top} \mathbf{h}^* \mathbf{x}^{\top} \mathbf{a}_j^* - y_j^*$. Then, $f = \sum_{j=1}^{M} r_j^* r_j$. Recall the definition of $\mathbf{z}$ in (17), and for notational simplicity, we write $f\left(\mathbf{z}\right) = f\left(\mathbf{h}, \mathbf{x}\right)$. Since $\mathbf{h}$ and $\mathbf{x}$ are complex-valued, we need to resort to Wirtinger calculus; see Section 6 in [5] for a brief introduction. The Wirtinger gradient of f with respect to $\mathbf{z}$ is

$$\nabla f\left(\mathbf{z}\right) = \left[\frac{\partial f}{\partial \mathbf{z}^*}; \frac{\partial f}{\partial \mathbf{z}}\right] = \left[\frac{\partial f}{\partial \mathbf{h}^*}; \frac{\partial f}{\partial \mathbf{x}^*}; \frac{\partial f}{\partial \mathbf{h}}; \frac{\partial f}{\partial \mathbf{x}}\right].$$

Since f is a real-valued function, the following identities always hold: $\frac{\partial f}{\partial \mathbf{h}^*} = \frac{\partial f^*}{\partial \mathbf{h}}$ and $\frac{\partial f}{\partial \mathbf{x}^*} = \frac{\partial f^*}{\partial \mathbf{x}}$. In light of these observations, one often omits the gradient with respect to the conjugates. For $\frac{\partial f}{\partial \mathbf{h}^*}$, since r_j does not contain $\mathbf{h}^*$ and r_j^* is linear in $\mathbf{h}^*$, we have $\frac{\partial r_j}{\partial \mathbf{h}^*} = \mathbf{0}$ and $\frac{\partial r_j^*}{\partial \mathbf{h}^*} = \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \mathbf{x}$, and thus $\frac{\partial\left(r_j^* r_j\right)}{\partial \mathbf{h}^*} = r_j^* \frac{\partial r_j}{\partial \mathbf{h}^*} + r_j \frac{\partial r_j^*}{\partial \mathbf{h}^*} = \mathbf{0} + r_j \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \mathbf{x}$. Similarly for $\frac{\partial f}{\partial \mathbf{x}^*}$, we have $\frac{\partial r_j^*}{\partial \mathbf{x}^*} = \mathbf{0}$, $\frac{\partial r_j}{\partial \mathbf{x}^*} = \mathbf{a}_j \mathbf{b}_j^{\mathsf{H}} \mathbf{h}$, and $\frac{\partial\left(r_j^* r_j\right)}{\partial \mathbf{x}^*} = r_j^* \frac{\partial r_j}{\partial \mathbf{x}^*} + r_j \frac{\partial r_j^*}{\partial \mathbf{x}^*} = \mathbf{0} + r_j \mathbf{a}_j \mathbf{b}_j^{\mathsf{H}} \mathbf{h}$. Therefore, the gradients are

$$\begin{aligned}\nabla_{\mathbf{h}} f\left(\mathbf{z}\right) = \frac{\partial f}{\partial \mathbf{h}^*} &= \sum_{j=1}^{M} \left(\mathbf{b}_j^{\mathsf{H}} \mathbf{h} \mathbf{x}^{\mathsf{H}} \mathbf{a}_j - y_j\right) \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \mathbf{x}, \\ \nabla_{\mathbf{x}} f\left(\mathbf{z}\right) = \frac{\partial f}{\partial \mathbf{x}^*} &= \sum_{j=1}^{M} \left(\mathbf{b}_j^{\mathsf{H}} \mathbf{h} \mathbf{x}^{\mathsf{H}} \mathbf{a}_j - y_j\right)^* \mathbf{a}_j \mathbf{b}_j^{\mathsf{H}} \mathbf{h}. \end{aligned} \tag{9}$$

Correspondingly, the gradient update rule (6) can be written as

$$\begin{aligned}\mathbf{h}^{k+\frac{1}{2}} &= \mathbf{h}^k - \frac{\eta}{\left\|\mathbf{x}^k\right\|^2} \sum_{j=1}^{M} \left(\mathbf{b}_j^{\mathsf{H}} \mathbf{h}^k \mathbf{x}^{k\mathsf{H}} \mathbf{a}_j - y_j\right) \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \mathbf{x}^k, \\ \mathbf{x}^{k+\frac{1}{2}} &= \mathbf{x}^k - \frac{\eta}{\left\|\mathbf{h}^k\right\|^2} \sum_{j=1}^{M} \left(\mathbf{b}_j^{\mathsf{H}} \mathbf{h}^k \mathbf{x}^{k\mathsf{H}} \mathbf{a}_j - y_j\right)^* \mathbf{a}_j \mathbf{b}_j^{\mathsf{H}} \mathbf{h}^k.\end{aligned}$$

B Proof of Theorem 1

B.1 Exact Support Recovery

The foundational intuition behind the sparse spectral initialization is that the expectation of the adjoint matrix $\mathbf{Z}$ perfectly aligns with the rank-one ground truth: $\mathbb{E}\left[\mathbf{Z}\right] = \sum_{j=1}^{M} \mathbb{E}\left[\mathbf{Z}_j\right] = \mathbf{h}^\star \mathbf{x}^{\star\mathsf{H}}$. We will

give formal proofs in this section that the initialization reliably recovers the true supports and is close to the ground truth with high probability.

Instead of aggregating energies across entire rows or columns, which accumulates noise unnecessarily, we examine the maximum absolute entries. For the l-th index in $\mathbf{h}^\star$ and the i-th index in $\mathbf{x}^\star$, the expected value of the corresponding element in $\mathbf{Z}$ is exactly $\mathbb{E}[Z_{l,i}] = h_l^\star x_i^{\star *}$.

Consequently, if we define the maximum magnitude of the i-th column as $r_i := \max_{l\in[M]} |Z_{l,i}|$, its behavior strictly depends on whether i belongs to the true support $I_x = \mathrm{supp}(\mathbf{x}^\star)$. For an off-support index $i \notin I_x$, we have $x_i^\star = 0$, meaning the entire column consists purely of zero-mean noise. Conversely, for an on-support index $i \in I_x$, the signal component is $h_l^\star x_i^{\star *}$. By taking the maximum over l, we selectively pick up the strongest signal component, which scales with $(\max_l |h_l^\star|)|x_i^\star|$ and prominently stands out from the noise. A symmetric argument applies to the row-wise maximums $c_l := \max_{i\in[N]} |Z_{l,i}|$, which effectively separates the kernel support I_h from the background noise.

Formally, we can write each element as

$$Z_{l,i} = h_l^\star x_i^{\star *} + E_{l,i},$$

where $E_{l,i} = \sum_{j=1}^M \left(\mathbf{b}_j^{\mathsf{H}}\mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\mathbf{a}_j b_{jl} a_{ji}^* - \mathbb{E}[\mathbf{b}_j^{\mathsf{H}}\mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\mathbf{a}_j b_{jl} a_{ji}^*]\right)$ is the perturbation noise. Standard processing of the sub-exponential random variables $(\mathbf{a}_j^{\mathsf{H}}\mathbf{x}^\star)a_{ji}$ dictates that the variance parameter is bounded by $\mathcal{O}(\frac{1}{M}\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2)$ and the sub-exponential norm is bounded by $\mathcal{O}(\frac{\mu}{M}\|\mathbf{h}^\star\|\|\mathbf{x}^\star\|)$. Applying Bernstein's inequality and taking a union bound over all MN elements, we obtain

$$\max_{l\in[M],i\in[N]} |E_{l,i}| \le C\left(\sqrt{\frac{\log(MN)}{M}} + \mu\frac{\log(MN)}{M}\right)\|\mathbf{h}^\star\|\|\mathbf{x}^\star\| := \epsilon,$$

with probability at least $1 - \mathcal{O}((MN)^{-10})$ for some absolute constant $C > 0$.

The support of $\mathbf{x}^\star$ is estimated by finding the largest s_x elements of $r_i = \max_{l\in[M]} |Z_{l,i}|$. For any on-support index $i \in I_x$, the triangle inequality yields

$$r_i \ge \max_{l\in I_h} |Z_{l,i}| \ge \max_{l\in I_h} (|h_l^\star||x_i^\star| - |E_{l,i}|) \ge \left(\min_{l\in I_h} |h_l^\star|\right)\left(\min_{i\in I_x} |x_i^\star|\right) - \epsilon.$$

By Assumption 3, the nonzero elements are uniformly lower-bounded as $\min_{l\in I_h} |h_l^\star| \ge \sqrt{\frac{C_h}{s_h}}\|\mathbf{h}^\star\|$ and $\min_{i\in I_x} |x_i^\star| \ge \sqrt{\frac{C_x}{s_x}}\|\mathbf{x}^\star\|$. Thus,

$$r_i \ge \sqrt{\frac{C_h C_x}{s_h s_x}}\|\mathbf{h}^\star\|\|\mathbf{x}^\star\| - \epsilon, \quad \forall i \in I_x.$$

Conversely, for any off-support index $i \notin I_x$, we have $x_i^\star = 0$, leading to $Z_{l,i} = E_{l,i}$. Consequently,

$$r_i = \max_{l\in[M]} |E_{l,i}| \le \epsilon, \quad \forall i \notin I_x.$$

To achieve exact support recovery, the condition $\min_{i\in I_x} r_i > \max_{i\notin I_x} r_i$ must be strictly satisfied, which is guaranteed if:

$$\sqrt{\frac{C_h C_x}{s_h s_x}} > 2C\left(\sqrt{\frac{\log(MN)}{M}} + \mu\frac{\log(MN)}{M}\right).$$

This inequality holds provided the sample complexity satisfies

$$M \gtrsim s_h s_x \log(MN) + \mu\sqrt{s_h s_x}\log(MN).$$

Given that μ characterizes the spectral flatness and $\mu \le \sqrt{M}$, the first term dominates, reducing the requisite sample complexity for support recovery to $M \gtrsim s_h s_x \log(MN)$. The support of $\mathbf{h}^\star$ is recovered with identical precision via the symmetric row-wise evaluation $c_l = \max_{i\in[N]} |Z_{l,i}|$.

B.2 Closeness to Ground Truth

Denote the estimated supports of $\mathbf{h}^0$ and $\mathbf{x}^0$ as $\hat{I}_h$ and $\hat{I}_x$, respectively. Denote the true support of $\mathbf{z}^\star$ as I_z and the estimated support of $\mathbf{z}^0$ as $\hat{I}_z$. Define $\mathbf{z}^\star_{\hat{I}_z} = \left[\mathbf{h}^{\star\mathsf{H}}_{\hat{I}_h}, \mathbf{x}^{\star\mathsf{H}}_{\hat{I}_x}\right]^{\mathsf{H}}$. If we can prove that $\mathbf{z}^\star \approx \mathbf{z}^\star_{\hat{I}_z}$ and $\mathbf{z}^\star_{\hat{I}_z} \approx \mathbf{z}^0$ up to scaling ambiguity, then we can claim that $\mathbf{z}^0 \approx \mathbf{z}^\star$ up to scaling ambiguity. First, we prove that $\mathbf{z}^\star \approx \mathbf{z}^\star_{\hat{I}_z}$ by the following Lemma.

Lemma 4. *If $\hat{I}_x$ and $\hat{I}_h$ are chosen as in Algorithm 1, with probability greater than $1 - \frac{1}{M}$,*

$$\left\|\mathbf{x}^\star - \mathbf{x}^\star_{\hat{I}_x}\right\| \le \delta_1 \left\|\mathbf{x}^\star\right\|, \quad \left\|\mathbf{h}^\star - \mathbf{h}^\star_{\hat{I}_h}\right\| \le \delta_1 \left\|\mathbf{h}^\star\right\|, \tag{10}$$

if the number of measurements satisfies $M \ge C_{\delta_1}\left(s_h^2 + \mu^4 s_x^2\right)\log M$.

Proof. If $\hat{I}_x$ is chosen such that it corresponds to the top s_x marginals r_i, then it will pick up three parts: I_x^+ corresponding to large marginals $r_i > \Theta$, $I_x^1 = \hat{I}_x \cap I_x^-$ and $I_x^2 = \hat{I}_x \cap I_x^c$ corresponding to small marginals $r_i \le \Theta$. That is,

$$\mathbf{x}^\star_{\hat{I}_x} = \mathbf{x}^\star_{I_x^+} + \mathbf{x}^\star_{I_x^1} + \mathbf{x}^\star_{I_x^2}.$$

By definition $\mathbf{x}^\star_{I_x^c} = \mathbf{0}$ and therefore $\mathbf{x}^\star_{I_x^2} = \mathbf{0}$. We have

$$\left\|\mathbf{x}^\star - \mathbf{x}^\star_{\hat{I}_x}\right\|^2 = \left\|\mathbf{x}^\star - \mathbf{x}^\star_{I_x^+} - \mathbf{x}^\star_{I_x^1}\right\|^2 \le \left\|\mathbf{x}^\star - \mathbf{x}^\star_{I_x^+}\right\|^2 + \left\|\mathbf{x}^\star_{I_x^1}\right\|^2 \le \left\|\mathbf{x}^\star - \mathbf{x}^\star_{I_x^+}\right\|^2 + \left\|\mathbf{x}^\star_{I_x^-}\right\|^2.$$

By construction, $I_x = I_x^- + I_x^+$. So $\mathbf{x}^\star = \mathbf{x}^\star_{I_x^-} + \mathbf{x}^\star_{I_x^+}$ and

$$\left\|\mathbf{x}^\star - \mathbf{x}^\star_{\hat{I}_x}\right\|^2 \le 2\left\|\mathbf{x}^\star_{I_x^-}\right\|^2.$$

By the definition of I_x^-, since there at most s_x elements in I_x^-, we have

$$\left\|\mathbf{x}^\star - \mathbf{x}^\star_{\hat{I}_x}\right\|^2 \le 2\left\|\mathbf{x}^\star_{I_x^-}\right\|^2 \le 26\mu^2 s_x\sqrt{\frac{\log M}{M}}\left\|\mathbf{x}^\star\right\|^2 \le \delta_1^2\left\|\mathbf{x}^\star\right\|^2.$$

Similarly, we can prove that

$$\left\|\mathbf{h}^\star - \mathbf{h}^\star_{\hat{I}_h}\right\|^2 \le 16 s_h\sqrt{\frac{\log M}{M}}\left\|\mathbf{h}^\star\right\|^2 \le \delta_1^2\left\|\mathbf{h}^\star\right\|^2.$$

Thus,

$$\left\|\mathbf{z}^\star - \mathbf{z}^\star_{\hat{I}_z}\right\|^2 = \left\|\mathbf{h}^\star - \mathbf{h}^\star_{\hat{I}_h}\right\|^2 + \left\|\mathbf{x}^\star - \mathbf{x}^\star_{\hat{I}_x}\right\|^2 \le \delta_1^2\left\|\mathbf{z}^\star\right\|^2.$$

This requires sample complexity M to satisfy $\max\left\{26\mu^2 s_x\sqrt{\frac{\log M}{M}}, 16 s_h\sqrt{\frac{\log M}{M}}\right\} \le \delta_1^2$, i.e.,

$$M \ge \max\left\{\frac{26^2}{\delta_1^4}\mu^4 s_x^2\log M, \frac{16^2}{\delta_1^4}s_h^2\log M\right\}.$$

Overall, if the number of measurements satisfies $M \ge C_{\delta_1}\left(s_h^2 + \mu^4 s_x^2\right)\log M$ with $\delta_1 = O(1)$, (10) holds true. □

We have proved that $\mathbf{z}^\star \approx \mathbf{z}^\star_{\hat{I}_z}$. Now we need to prove that $\mathbf{z}^\star_{\hat{I}_z} \approx \mathbf{z}^0$ up to scaling ambiguity, which we do using Lemma 5.

Lemma 5. *Let $\left(\mathbf{x}^0, \mathbf{h}^0\right)$ be constructed from* (5)*. With probability greater than $1 - \frac{1}{M}$,*

$$\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\| + \left\|\alpha\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h}\right\|\right\} \le \xi\left\|\mathbf{x}^\star\right\|\left\|\mathbf{h}^\star\right\|.$$

if the number of measurements M satisfies $M \ge C\mu^2\left(s_h + s_x\right)\log M/\xi^2$.

Proof. Define $I_x^3 = \left(I_x \cup I_x^2\right) \subset \left(I \cup \hat{I}_x\right)$ and $I_h^3 = \left(I_h \cup I_h^2\right) \subset \left(I \cup \hat{I}_h\right)$. Recall that $I_x^2 = \hat{I}_x \cap I_x^c$ so $\mathbf{x}^\star_{I_x^2} = \mathbf{0}$. Then, $\left|I_x^3\right| \leq 2s_x$, $\mathbf{x}^\star_{I_x^3} = \mathbf{x}^\star_{I_x} + \mathbf{x}^\star_{I_x^2} = \mathbf{x}^\star$, and

$$\mathbb{E}\left[\mathbf{Z}_{I_h^3, I_x^3}\right] = \sum_{j=1}^{M} \mathbb{E}\left[y_j \cdot \mathbf{a}_{j,I_x^3}\mathbf{b}^{\mathsf{H}}_{j,I_h^3}\right] = \sum_{j=1}^{M} \mathbb{E}\left[\mathbf{a}_{j,I_x^3}\mathbf{a}^{\mathsf{H}}_{j,I_x^3}\right] \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\mathbf{b}_{j,I_h^3}\mathbf{b}^{\mathsf{H}}_{j,I_h^3} = \mathbf{h}^\star_{I_h^3}\mathbf{x}^{\star\mathsf{H}}_{I_x^3} = \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}.$$

By Lemma 14, we have

$$\left\|\mathbf{Z}_{I_h^3, I_x^3} - \mathbb{E}\left[\mathbf{Z}_{I_h^3, I_x^3}\right]\right\|_2 \leq \xi \left\|\mathbf{x}^\star\right\| \left\|\mathbf{h}^\star\right\|,$$

with probability at least $1 - M^{-1}$, provided that $M \geq C\mu^2 \left(s_h + s_x\right) \log M/\xi^2$ for some universal constant $C > 0$. Now we can use the fact that $\hat{I}_x \subset I_x^3$, so that

$$\left\|\mathbf{Z}_{\hat{I}_h, \hat{I}_x} - \mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]\right\|_2 \leq \left\|\mathbf{Z}_{I_h^3, I_x^3} - \mathbb{E}\left[\mathbf{Z}_{I_h^3, I_x^3}\right]\right\|_2 \leq \xi \left\|\mathbf{x}^\star\right\| \left\|\mathbf{h}^\star\right\|. \tag{11}$$

Since $\mathbf{Z}_{\hat{I}_h, \hat{I}_x}$ can be seen as a perturbation of $\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]$, where the top two singular values of $\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]$ are $\left\|\mathbf{x}^\star\right\| \left\|\mathbf{h}^\star\right\|$ and 0, we can use a variant of Wedin's $\sin\Theta$ theorem [42] to bound the distance between the normalized top-singular vectors $\left(\check{\mathbf{h}}^0, \check{\mathbf{x}}^0\right)$ of $\mathbf{Z}_{\hat{I}_h, \hat{I}_x}$ and $\left(\check{\mathbf{h}}^\star_{\hat{I}_h}, \check{\mathbf{x}}^\star_{\hat{I}_x}\right)$ of $\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]$. That is,

$$\min_{\alpha \in \mathbb{C}, |\alpha| = 1} \left\{\left\|\alpha\check{\mathbf{x}}^0 - \check{\mathbf{x}}^\star_{\hat{I}_x}\right\| + \left\|\alpha\check{\mathbf{h}}^0 - \check{\mathbf{h}}^\star_{\hat{I}_h}\right\|\right\} \leq \frac{c_1 \left\|\mathbf{Z}_{\hat{I}_h, \hat{I}_x} - \mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]\right\|_2}{\sigma_1\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]\right) - \sigma_2\left(\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right)}, \tag{12}$$

for some universal constant $c_1 > 0$. Regarding the denominator of the above inequality, we can take (11) together with Weyl's inequality to demonstrate that

$$\begin{aligned}\sigma_1\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]\right) - \sigma_2\left(\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right) \geq& \sigma_1\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]\right) - \sigma_2\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]\right) - \left\|\mathbf{Z}_{\hat{I}_h, \hat{I}_x} - \mathbb{E}\left[\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right]\right\| \\ \geq& (1 - \xi)\left\|\mathbf{x}^\star\right\| \left\|\mathbf{h}^\star\right\|.\end{aligned}$$

These together with (12) reveal that

$$\min_{\alpha \in \mathbb{C}, |\alpha| = 1} \left\{\left\|\alpha\check{\mathbf{x}}^0 - \check{\mathbf{x}}^\star_{\hat{I}_x}\right\| + \left\|\alpha\check{\mathbf{h}}^0 - \check{\mathbf{h}}^\star_{\hat{I}_h}\right\|\right\} \leq \frac{c_1\xi}{(1 - \xi)} \leq 2c_1\xi, \tag{13}$$

as long as $\xi \leq \frac{1}{2}$.

Now we connect the preceding bound (13) with the scaled singular vectors $\mathbf{h}^0 = \sqrt{\sigma_1\left(\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right)}\check{\mathbf{h}}^0$ and $\mathbf{x}^0 = \sqrt{\sigma_1\left(\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right)}\check{\mathbf{x}}^0$. For any $\alpha \in \mathbb{C}$ with $|\alpha| = 1$, we have

$$\begin{aligned}\left\|\alpha\mathbf{x}^0 - \check{\mathbf{x}}^\star_{\hat{I}_x}\right\| + \left\|\alpha\mathbf{h}^0 - \check{\mathbf{h}}^\star_{\hat{I}_h}\right\| =& \left\|\sigma_1\left(\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right)\left(\alpha\check{\mathbf{x}}^0\right) - \check{\mathbf{x}}^\star_{\hat{I}_x}\right\| \\ &+ \left\|\alpha\sigma_1\left(\mathbf{Z}_{\hat{I}_h, \hat{I}_x}\right)\left(\alpha\check{\mathbf{h}}^0\right) - \check{\mathbf{h}}^\star_{\hat{I}_h}\right\|.\end{aligned}$$

Since $\left(\alpha\check{\mathbf{h}}^0, \alpha\check{\mathbf{x}}^0\right)$, $\left(\check{\mathbf{h}}^\star_{\hat{I}_h}, \check{\mathbf{x}}^\star_{\hat{I}_x}\right)$ are also the leading left and right singular vectors of $\mathbf{Z}_{\hat{I}_h,\hat{I}_x}$ and $\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right]$, respectively, we can invoke Lemma 60 in [31] to get

$$\begin{aligned}\left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\| + \left\|\alpha\mathbf{h}^0 - \mathbf{h}^\star\right\| \leq& \sqrt{\sigma_1\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right]\right)}\left(\left\|\alpha\check{\mathbf{x}}^0 - \check{\mathbf{x}}^\star_{\hat{I}_x}\right\| + \left\|\alpha\check{\mathbf{h}}^0 - \check{\mathbf{h}}^\star_{\hat{I}_h}\right\|\right)\\ &+ \frac{2\left|\sigma_1\left(\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right) - \sigma_1\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right]\right)\right|}{\sqrt{\sigma_1\left(\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right)} + \sqrt{\sigma_1\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right]\right)}}\\ =& \left\|\mathbf{x}^\star\right\|\left\|\mathbf{h}^\star\right\|\left(\left\|\alpha\check{\mathbf{x}}^0 - \check{\mathbf{x}}^\star_{\hat{I}_x}\right\| + \left\|\alpha\check{\mathbf{h}}^0 - \check{\mathbf{h}}^\star_{\hat{I}_h}\right\|\right)\\ &+ \frac{2\left|\sigma_1\left(\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right) - \sigma_1\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right]\right)\right|}{\sqrt{\sigma_1\left(\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right)} + \left\|\mathbf{x}^\star\right\|\left\|\mathbf{h}^\star\right\|}. \end{aligned} \tag{14}$$

In addition, we can apply Weyl's inequality once again to deduce that

$$\left|\sigma_1\left(\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right) - \sigma_1\left(\mathbb{E}\left[\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right]\right)\right| \leq \left\|\mathbf{Z}_{\hat{I}_h,\hat{I}_x} - \mathbb{E}\left[\mathbf{Z}_{\hat{I}_h,\hat{I}_x}\right]\right\|_2 \leq \xi\left\|\mathbf{x}^\star\right\|\left\|\mathbf{h}^\star\right\|.$$

Substitute the above inequality into (14) to obtain

$$\left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\| + \left\|\alpha\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h}\right\| \leq \left\|\mathbf{x}^\star\right\|\left\|\mathbf{h}^\star\right\|\left(\left\|\alpha\check{\mathbf{x}}^0 - \check{\mathbf{x}}^\star_{\hat{I}_x}\right\| + \left\|\alpha\check{\mathbf{h}}^0 - \check{\mathbf{h}}^\star_{\hat{I}_h}\right\|\right) + 2\xi.$$

Taking the minimum over α, one can thus conclude that

$$\begin{aligned}\min_{|\alpha|=1}\left\{\left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\| + \left\|\alpha\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h}\right\|\right\} &\leq \min_{|\alpha|=1}\left\{\left\|\alpha\check{\mathbf{x}}^0 - \check{\mathbf{x}}^\star_{\hat{I}_x}\right\| + \left\|\alpha\check{\mathbf{h}}^0 - \check{\mathbf{h}}^\star_{\hat{I}_h}\right\|\right\}\left\|\mathbf{x}^\star\right\|\left\|\mathbf{h}^\star\right\| + 2\xi\\ &\leq 2c_1\xi\left\|\mathbf{x}^\star\right\|\left\|\mathbf{h}^\star\right\| + 2\xi.\end{aligned}$$

Since ξ is arbitrary, by taking $M/\left(\mu^2\left(s_h+s_x\right)\log M\right)$ to be large enough, we finish the proof. □

Remark 6. The bound stated in Lemma 5 has an additive dimensional dependence $\mu^2(s_h+s_x)$ for spectral estimation after support restriction. This differs from the multiplicative requirement $s_h s_x$ for entrywise support recovery in Theorem 1. The symmetric $\mu^2(s_h+s_x)$ dependence in our matrix concentration argument comes from a common worst-case variance bound; sharpening it toward $s_h+\mu^2 s_x$ remains an open direction. Concentration on the fixed true support must also be distinguished from uniform concentration over adaptively selected supports, which requires additional control.

Define $\mathbf{z}^\star_{\hat{I}_z} = \left[\mathbf{h}^{\star\mathsf{H}}_{\hat{I}_h}, \mathbf{x}^{\star\mathsf{H}}_{\hat{I}_x}\right]^\mathsf{H}$. From the definition of dist $(\cdot,\cdot)$ we immediately have

$$\begin{aligned}\text{dist}\left(\mathbf{z}^0, \mathbf{z}^\star_{\hat{I}_z}\right) &= \min_{\alpha\in\mathbb{C}}\sqrt{\left\|\frac{1}{\alpha^*}\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h}\right\|^2 + \left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\|^2}\\ &\leq \min_{\alpha\in\mathbb{C}}\left\{\left\|\frac{1}{\alpha^*}\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h}\right\| + \left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\|\right\}\\ &\leq \min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\| + \left\|\alpha\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h}\right\|\right\} \lesssim \xi\left\|\mathbf{x}^\star\right\|\left\|\mathbf{h}^\star\right\|,\end{aligned}$$

as long as $M \geq C\mu^2\left(s_h+s_x\right)\log M/\xi^2$ for some sufficiently large constant $C>0$. If ξ is chosen to satisfy $\xi \ll \frac{\|\mathbf{z}^\star\|}{2\|\mathbf{x}^\star\|\|\mathbf{h}^\star\|}$, we have

$$\text{dist}\left(\mathbf{z}^0, \mathbf{z}^\star_{\hat{I}_z}\right) \leq \delta_2\left\|\mathbf{z}^\star\right\| \ll \frac{1}{2}\left\|\mathbf{z}^\star\right\|.$$

Combining Lemmas 4 and 5, we have the final result: If the number of measurements M satisfies $M \gtrsim \left(s_h^2 + \mu^4 s_x^2\right) \log M$, for $0 < \delta_1 + \delta_2 < 1$,

$$\begin{aligned}
\operatorname{dist}\left(\mathbf{z}^0, \mathbf{z}^\star\right) &= \min_{\alpha\in\mathbb{C}} \sqrt{\left\| \frac{1}{\alpha^*}\mathbf{h}^0 - \mathbf{h}^\star \right\|^2 + \left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star\right\|^2} \\
&= \min_{\alpha\in\mathbb{C}} \sqrt{\left\| \frac{1}{\alpha^*}\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h} + \mathbf{h}^\star_{\hat{I}_h} - \mathbf{h}^\star \right\|^2 + \left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x} + \mathbf{x}^\star_{\hat{I}_x} - \mathbf{x}^\star\right\|^2} \\
&\le \min_{\alpha\in\mathbb{C}} \sqrt{\left(\left\| \frac{1}{\alpha^*}\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h} \right\| + \left\| \mathbf{h}^\star_{\hat{I}_h} - \mathbf{h}^\star \right\|\right)^2 + \left(\left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\| + \left\|\mathbf{x}^\star_{\hat{I}_x} - \mathbf{x}^\star\right\|\right)^2} \\
&\le \min_{\alpha\in\mathbb{C}} \left\{ \sqrt{\left\| \frac{1}{\alpha^*}\mathbf{h}^0 - \mathbf{h}^\star_{\hat{I}_h} \right\|^2 + \left\|\alpha\mathbf{x}^0 - \mathbf{x}^\star_{\hat{I}_x}\right\|^2} + \sqrt{\left\| \mathbf{h}^\star_{\hat{I}_h} - \mathbf{h}^\star \right\|^2 + \left\|\mathbf{x}^\star_{\hat{I}_x} - \mathbf{x}^\star\right\|^2} \right\} \\
&= \operatorname{dist}\left(\mathbf{z}^0, \mathbf{z}^\star_{\hat{I}_z}\right) + \left\|\mathbf{z}^\star - \mathbf{z}^\star_{\hat{I}_z}\right\|^2 \le (\delta_1 + \delta_2)\left\|\mathbf{z}^\star\right\|.
\end{aligned} \tag{15}$$

C Proof of Theorem 2

In this section, we establish that under the proposed sparse spectral initialization and suitable sample complexity, the iterates generated by Algorithm 2 converge linearly to the ground truth up to the intrinsic scaling ambiguity. The proof is organized into three main steps:

(1) Local restricted strong convexity and smoothness. In Lemma 7, we define a region of contraction (ROC) wherein the empirical Wirtinger Hessian satisfies uniform restricted strong convexity and smoothness properties.
(2) Error contraction of the gradient descent step. By Lemmas 10 and 11, we demonstrate that within the ROC, a single step of the normalized gradient descent reduces the distance to the ground truth by a constant factor.
(3) Trajectory of ThWF staying in the ROC. With the constrained minimum value condition of the truths, once the iterates are sufficiently close to the true support, the hard-thresholding operator reduces to an orthogonal projection onto the true support and does not disrupt the contraction, and thus the iterates of ThWF stay in the ROC.

Without loss of generality, we assume $\|\mathbf{h}^\star\| = \|\mathbf{x}^\star\| = 1$. We say $\mathbf{z}_1 = \left[\mathbf{h}_1^{\mathsf{H}}, \mathbf{x}_1^{\mathsf{H}}\right]^{\mathsf{H}}$ is aligned with $\mathbf{z}_2 = \left[\mathbf{h}_2^{\mathsf{H}}, \mathbf{x}_2^{\mathsf{H}}\right]^{\mathsf{H}}$, if

$$\|\mathbf{h}_1 - \mathbf{h}_2\|^2 + \|\mathbf{x}_1 - \mathbf{x}_2\|^2 = \min_{\alpha\in\mathbb{C}} \left\{ \left\| \frac{1}{\alpha^*}\mathbf{h}_1 - \mathbf{h}_2 \right\|^2 + \|\alpha\mathbf{x}_1 - \mathbf{x}_2\|^2 \right\}.$$

To simplify notations, define

$$\widetilde{\mathbf{z}}^k = \left[\widetilde{\mathbf{h}}^{k\mathsf{H}}, \widetilde{\mathbf{x}}^{k\mathsf{H}}\right]^{\mathsf{H}} = \left[\frac{1}{\alpha^{*k}}\mathbf{h}^{k\mathsf{H}}, \alpha^k\mathbf{x}^{k\mathsf{H}}\right]^{\mathsf{H}}, \tag{16}$$

where the alignment parameter α^k minimizes the distance (8). We can see that $\widetilde{\mathbf{z}}^k$ is aligned with $\mathbf{z}^\star$ and

$$\operatorname{dist}\left(\mathbf{z}^k, \mathbf{z}^\star\right) = \operatorname{dist}\left(\widetilde{\mathbf{z}}^k, \mathbf{z}^\star\right) = \left\|\widetilde{\mathbf{z}}^k - \mathbf{z}^\star\right\|.$$

C.1 Local geometry

We begin by characterizing the ROC within which the empirical loss function exhibits favorable geometric properties. This is formalized in the following lemma.

Lemma 7. *(Restricted strong convexity and smoothness) Let $c > 0$ be a sufficiently small constant and $\delta = c/\log^2 M$. Suppose the sample size satisfies $M \gtrsim \mu^2 (s_h + s_x) \log^9 M$ for some sufficiently*

large constant $c_0 > 0$. Then with probability $1 - \mathcal{O}\left(M^{-10} + e^{-(s_h+s_x)} \log M\right)$ with $C_0 > 0$, the Wirtinger Hessian $\nabla^2 f(\mathbf{z})$ satisfies

$$\mathbf{u}^{\mathsf{H}} \left[\mathbf{D}\nabla^2 f(\mathbf{z}) + \nabla^2 f(\mathbf{z})\, \mathbf{D}\right] \mathbf{u} \geq \frac{1}{4} \|\mathbf{u}\|^2, \quad \left\|\nabla^2 f(\mathbf{z})\right\| \leq 3,$$

simultaneously for all

$$\mathbf{z} = \begin{bmatrix} \mathbf{h} \\ \mathbf{x} \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} \mathbf{h}_1 - \mathbf{h}_2 \\ \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{h}_1^* - \mathbf{h}_2^* \\ \mathbf{x}_1^* - \mathbf{x}_2^* \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} \gamma_1 \mathbf{I}_N & & & \\ & \gamma_2 \mathbf{I}_N & & \\ & & \gamma_1 \mathbf{I}_N & \\ & & & \gamma_2 \mathbf{I}_N \end{bmatrix}, \tag{17}$$

where $\mathbf{z}$ satisfies

$$\|\mathbf{h}\|_0 \leq s_h, \quad \|\mathbf{x}\|_0 \leq s_x, \quad \max\left\{\|\mathbf{h} - \mathbf{h}^\star\|, \|\mathbf{x} - \mathbf{x}^\star\|\right\} \leq \delta, \tag{18}$$

$$\max_{i \leq j \leq M} \left|\mathbf{a}_j^{\mathsf{H}} (\mathbf{x} - \mathbf{x}^\star)\right| \leq 2C_3 \frac{1}{\log^{3/2} M}, \quad \max_{i \leq j \leq M} \left|\mathbf{b}_j^{\mathsf{H}} \mathbf{h}\right| \leq 2C_4 \frac{\mu}{\sqrt{M}} \log^2 M, \tag{19}$$

$\mathbf{z}_1 = \left[\mathbf{h}_1^{\mathsf{H}}, \mathbf{x}_1^{\mathsf{H}}\right]^{\mathsf{H}}$ is aligned with $\mathbf{z}_2 = \left[\mathbf{h}_2^{\mathsf{H}}, \mathbf{x}_2^{\mathsf{H}}\right]^{\mathsf{H}}$ and they satisfy

$$\begin{aligned} \max\left\{\|\mathbf{h}_1 - \mathbf{h}^\star\|, \|\mathbf{x}_1 - \mathbf{x}^\star\|, \|\mathbf{h}_2 - \mathbf{h}^\star\|, \|\mathbf{x}_2 - \mathbf{x}^\star\|\right\} &\leq \delta, \\ \|\mathbf{h}_1 - \mathbf{h}_2\|_0 \leq 2s_h, \quad \|\mathbf{x}_1 - \mathbf{x}_2\|_0 &\leq 2s_x, \end{aligned}$$

and $\mathbf{D}$ satisfies for $\gamma_1, \gamma_2 \in \mathbb{R}$,

$$\max\left\{|\gamma_1 - 1|, |\gamma_2 - 1|\right\} \leq \delta.$$

C.1.1 Proof of Lemma 7

It is convenient to decompose the Wirtinger Hessian into the expected Wirtinger Hessian at the truth and the perturbation part as follows:

$$\nabla^2 f(\mathbf{z}) = \nabla^2 F(\mathbf{z}^\star) + \left(\nabla^2 f(\mathbf{z}) - \nabla^2 F(\mathbf{z}^\star)\right).$$

The proof then proceeds by establishing two claims: (i) the population Hessian $\nabla^2 F(\mathbf{z}^\star)$ satisfies the restricted strong convexity and smoothness properties as advertised; and (ii) the perturbation $\nabla^2 f(\mathbf{z}) - \nabla^2 F(\mathbf{z}^\star)$ is well-controlled within the ROC.

The population Hessian satisfies the following lemma.

Lemma 8. *[31]Instate the notation and the conditions of Lemma 7. We have*

$$\left\|\nabla^2 F(\mathbf{z}^\star)\right\| = 2, \quad \mathbf{u}^{\mathsf{H}} \left[\mathbf{D}\nabla^2 F(\mathbf{z}^\star) + \nabla^2 F(\mathbf{z}^\star)\, \mathbf{D}\right] \mathbf{u} \geq \|\mathbf{u}\|^2.$$

Proof. The proof of Lemma 8 is similar to the proof of Lemma 26 in [31] which we omit here. □

Next we bound the spectral norm of the perturbation uniformly over the ROC. To this end, we define the set

$$\mathcal{R} := \{\mathbf{z} \mid \mathbf{z} \text{ satisfies conditions in Lemma 7}\}$$

While Lemma 27 in [31] loses its analytical significance when $K = M$, our proposed thresholded Wirtinger flow algorithm overcomes this by leveraging a sparse spectral initialization that achieves exact support recovery. Consequently, the Hard-thresholding operator ensures that the iterates $\{\mathbf{z}^k\}$ remain within a restricted sparse manifold throughout the execution. This allows us to re-characterize the local geometry on the subspaces defined by the union of the estimated and true supports. Specifically, the dimension K in [31] is replaced by the restricted dimensions $2s_h$ and $2s_x$ for $\mathbf{h}$ and $\mathbf{x}$, respectively. Under the restricted subspaces, the concentration of the Hessian perturbation is restored, ensuring a descent toward the ground truth.

Lemma 9. *Suppose the sample complexity satisfies $M \gg \mu^2 (s_h + s_x) \log^9 M$, $c > 0$ is a sufficiently small constant, and $\delta = c/\log^2 M$. Then with probability at least $1 - \mathcal{O}\left(M^{-10} + e^{-(s_h+s_x)} \log M\right)$, one has*

$$\sup_{\mathbf{z} \in \mathcal{R}} \left\|\nabla^2 f(\mathbf{z}) - \nabla^2 F(\mathbf{z}^\star)\right\| \leq \frac{1}{4}.$$

Proof. In view of the expressions of $\nabla^2 f(\mathbf{z})$ and $\nabla^2 F(\mathbf{z}^\star)$ and the triangle inequality, we get

$$\left\|\nabla^2 f(\mathbf{z}) - \nabla^2 F(\mathbf{z}^\star)\right\| \le 2\alpha_1 + 2\alpha_2 + 4\alpha_3 + 4\alpha_4,$$

where the four terms on the right-hand side are defined as

$$\alpha_1 = \left\|\sum_{j=1}^{M} \left|\mathbf{a}_j^{\mathsf{H}}\mathbf{x}\right|^2 \mathbf{b}_j\mathbf{b}_j^{\mathsf{H}} - \mathbf{I}_M\right\|, \quad \alpha_2 = \left\|\sum_{j=1}^{M} \left|\mathbf{b}_j^{\mathsf{H}}\mathbf{h}\right|^2 \mathbf{a}_j\mathbf{a}_j^{\mathsf{H}} - \mathbf{I}_N\right\|,$$

$$\alpha_3 = \left\|\sum_{j=1}^{M} \left(\mathbf{b}_j^{\mathsf{H}}\mathbf{h}\mathbf{x}^{\mathsf{H}}\mathbf{a}_j - y_j\right)\mathbf{b}_j\mathbf{a}_j^{\mathsf{H}}\right\|, \quad \alpha_4 = \left\|\sum_{j=1}^{M} \mathbf{b}_j\mathbf{b}_j^{\mathsf{H}}\mathbf{h}\left(\mathbf{a}_j\mathbf{a}_j^{\mathsf{H}}\mathbf{x}\right)^\top - \mathbf{h}^\star\mathbf{x}^{\star\top}\right\|.$$

A similar proof to that of Lemma 27 in [31] gives that:

(1) $\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{R}} \alpha_1 \le 1/32\right) = 1 - \mathcal{O}\left(M^{-10}\right)$;

(2) $\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{R}} \alpha_2 \le 1/32\right) = 1 - \mathcal{O}\left(M^{-10}\right)$ under the sample complexity $M \gg \mu^2(s_h + s_x)\log^5 M$;

(3) $\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{R}} \alpha_3 \le 1/96\right) = 1 - \mathcal{O}\left(M^{-10} + e^{-(s_h+s_x)}\log M\right)$ if $M \gg \mu^2(s_h + s_x)\log^9 M$, with a RIP-type property proved by Lemma 17;

(4) $\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{R}} \alpha_4 \le 1/48\right) = 1 - \mathcal{O}\left(M^{-10} + e^{-(s_h+s_x)}\log M\right)$.

Overall, $\left\|\nabla^2 f(\mathbf{z}) - \nabla^2 F(\mathbf{z}^\star)\right\| \le \frac{1}{4}$ with probability $1 - \mathcal{O}\left(M^{-10} + e^{-(s_h+s_x)}\log M\right)$, if $M \gg \mu^2(s_h + s_x)\log^9 M$. □

Combining Lemma 8 and Lemma 9, the restricted strong convexity and smoothness within the ROC follow immediately, as in [31]. This completes the proof of Lemma 8.

C.2 Error contraction of ThWF

With the local geometry established in Lemma 8, we now analyze the contraction rate of a single complex gradient descent step. We will demonstrate that, as long as the iterates remain within the ROC, the distance to the ground truth, measured by $\mathrm{dist}(\cdot, \mathbf{z}^\star)$, decreases by a constant factor.

The proof of this contraction consists of three substeps:

(1) Gradient descent contraction (Lemma 10): We show that a single normalized gradient descent step yields linear convergence in the aligned coordinate system.
(2) Convergence of the alignment parameter (Lemma 11): We establish that the scaling factor α^k converges rapidly, ensuring that the raw iterates remain well-balanced.
(3) ThWF iterates staying in the ROC (by induction): We prove by induction that, given that the minimum values of $\mathbf{x}^\star$ and $\mathbf{h}^\star$ have upper bounds, the hard-thresholding operation consistently identifies the true support, reducing to an orthogonal projection that preserves the contraction. Moreover, by the leave-one-out technique, we can prove that the ThWF iterates stay in the ROC.

Lemma 10. *(GD contraction) Suppose the number of measurements satisfies $M \gg \mu^2(s_h + s_x)\log^9 M$ for some sufficiently large constant $C > 0$, and the step size $\eta > 0$ is some sufficiently small constant. There exists an event that does not depend on k and has probability $1 - \mathcal{O}\left(M^{-10} + e^{-(s_h+s_x)}\log M\right)$, such that if*

$$\mathit{dist}\left(\mathbf{z}^k, \mathbf{z}^\star\right) \le \xi, \quad \max_{i\le j\le M} \left|\mathbf{a}_j^{\mathsf{H}}\left(\widetilde{\mathbf{x}}^k - \mathbf{x}^\star\right)\right| \le C_3\frac{1}{\log^{3/2} M}, \quad \max_{i\le j\le M} \left|\mathbf{b}_j^{\mathsf{H}}\widetilde{\mathbf{h}}^k\right| \le C_4\frac{\mu}{\sqrt{M}}\log^2 M,$$

hold for some constants $C_3, C_4 > 0$, one has

$$\textit{dist}\left(\mathbf{z}^{k+\frac{1}{2}}, \mathbf{z}^\star\right) \le \left(1 - \frac{\eta}{16}\right) \textit{dist}\left(\mathbf{z}^k, \mathbf{z}^\star\right).$$

where $\mathbf{z}^{k+\frac{1}{2}} = \left[\mathbf{h}^{k+\frac{1}{2}\mathsf{H}}, \mathbf{x}^{k+\frac{1}{2}\mathsf{H}}\right]^\mathsf{H}$, with $\mathbf{h}^{k+\frac{1}{2}}$ and $\mathbf{x}^{k+\frac{1}{2}}$ are the variables updated by a step of Wirtinger gradient descent in (6)*. Here, $\widetilde{\mathbf{h}}^k$ and $\widetilde{\mathbf{x}}^k$ are defined in* (16)*, and $\xi \le 1/\log^2 M$.*

Proof. The proof is similar to that of Lemma 15 in [31] which we omit here. □

Before establishing the contraction of the iterates, we must address the scaling ambiguity inherent in blind deconvolution. Recall that the distance metric $\text{dist}(\mathbf{z}, \mathbf{z}^\star)$ is defined via an optimal alignment parameter α that minimizes the joint discrepancy. While the gradient update (6) is applied to the raw iterates $\mathbf{h}^k$ and $\mathbf{x}^k$, our geometric analysis in Lemma 7 and the contraction result in Lemma 8 are stated in terms of the aligned iterates $\widetilde{\mathbf{h}}^k$ and $\widetilde{\mathbf{x}}^k$.

To bridge the gap between the algorithmic update of the raw iterates and the theoretical contraction of the aligned iterates, it is essential to control the evolution of the alignment parameter α^k. The following lemma demonstrates that once the iterates are sufficiently close to the ground truth, the scaling factor stabilizes rapidly. Specifically, the relative change in α^k is bounded by the current estimation error. This property is crucial for two reasons: (i) it guarantees that the magnitudes of the estimates $\left\|\mathbf{h}^k\right\|$ and $\left\|\mathbf{x}^k\right\|$ remain balanced and converge without requiring explicit norm regularization; and (ii) it ensures that the hard-thresholding operation commutes with the alignment in the sense that thresholding an aligned iterate is equivalent to aligning a thresholded raw iterate up to negligible error.

Lemma 11. *(convergence of α^k) When $M > 1$ is sufficiently large, the following two claims hold true.*

(1) If $\left|\left|\alpha^k\right| - 1\right| \le \frac{1}{2}$ and $\textit{dist}\left(\mathbf{z}^k, \mathbf{z}^\star\right) \le C_1/\log^2 M$, then $\left|\frac{\alpha^{k+1}}{\alpha^k} - 1\right| \le c\textit{dist}\left(\mathbf{z}^k, \mathbf{z}^\star\right) \le cC_1/\log^2 M$ for some absolute constant $c > 0$;

(2) If $\left|\left|\alpha^0\right| - 1\right| \le 1/4$ and $\textit{dist}\left(\mathbf{z}^s, \mathbf{z}^\star\right) \le C_1\left(1 - \eta/16\right)^s/\log^2 M$ for all $0 \le s \le k$, then $\left|\left|\alpha^{s+1}\right| - 1\right| \le \frac{1}{2}$.

Proof. The proof is similar to that of Lemma 16 in [31] which we omit here. □

Finally, we establish the linear convergence of ThWF via induction. Let $\mathcal{E}_k$ be the event that the k-th iterate $\mathbf{z}^k$ resides in the ROC as specified in Lemma 7, and additionally satisfies the exact support recovery with high probability:

$$\text{supp}\left(\mathbf{x}^k\right) = I_x, \quad \text{supp}\left(\mathbf{h}^k\right) = I_h, \tag{20}$$

$$\text{dist}\left(\mathbf{z}^k, \mathbf{z}^\star\right) \le \left(1 - \frac{\eta}{16}\right)^k \text{dist}\left(\mathbf{z}^0, \mathbf{z}^\star\right). \tag{21}$$

Base Case ($k = 0$): Under the assumption on the minimum magnitude of the ground truth, and the fact that the spectral initialization satisfies $\text{dist}(\widetilde{\mathbf{z}}^0, \mathbf{z}^\star) \le \lambda\|\mathbf{z}^\star\|$ with probability at least $1 - O(M^{-10})$, the entry-wise error bound $\|\widetilde{\mathbf{x}}^0 - \mathbf{x}^\star\|_\infty \le \|\widetilde{\mathbf{x}}^0 - \mathbf{x}^\star\| < \left\|\widetilde{\mathbf{z}}^0 - \mathbf{z}^\star\right\| < \frac{1}{2}\min_{i\in I_x}|x_i^\star|$ holds, if we choose $\lambda = \delta_1 + \delta_2 < \frac{1}{2\|\mathbf{z}^\star\|}\min\left\{\left|\min_{i\in I_x} x_i^\star\right|, \left|\min_{l\in I_h} h_l^\star\right|\right\} \lesssim \frac{1}{2\|\mathbf{z}^\star\|}\min\left\{\frac{1}{\sqrt{s_x}}\|\mathbf{x}^\star\|, \frac{1}{\sqrt{s_h}}\|\mathbf{h}^\star\|\right\}$. This ensures that the magnitude of any signal element on the true support I_x strictly exceeds that of any element off the support. Consequently, the hard-thresholding operator $\mathcal{T}_{s_x}$ identifies the true support with high probability, i.e., $\text{supp}(\mathbf{x}^0) = I_x$. A symmetric argument applies to $\mathbf{h}^0$. The incoherence conditions (19) then follow by invoking Lemmas 19-21 in [31], where the effective dimension is restricted to $s = s_h + s_x$ instead of K.

Inductive Step: Suppose $\mathcal{E}_k$ holds. Since $\text{supp}(\mathbf{z}^k) = I_z = I_h \cup I_x$, the gradient update $\mathbf{z}^{k+\frac{1}{2}}$ effectively operates within the fixed $(s_h + s_x)$-dimensional subspace $\mathcal{S}_{I_z}$. Within this subspace, the restricted local geometry is well-behaved. Specifically, the dimension K in the analysis of [31] is replaced by s_h, and since $M \gtrsim (s_h + s_x)\log^9 M$, the subspace Hessian satisfies the restricted strong convexity and smoothness. By Lemma 10, we have:

$$\text{dist}(\mathbf{z}^{k+\frac{1}{2}}, \mathbf{z}^\star) \le \rho \cdot \text{dist}(\mathbf{z}^k, \mathbf{z}^\star), \quad 0 < \rho < 1.$$

Due to this contraction, the temporary variable $\mathbf{z}^{k+\frac{1}{2}}$ remains closer to $\mathbf{z}^\star$ than $\mathbf{z}^0$ was. Therefore,

$$\begin{aligned}&\max\left\{\left\|\widetilde{\mathbf{x}}^{k+\frac{1}{2}} - \mathbf{x}^\star\right\|, \left\|\widetilde{\mathbf{h}}^{k+\frac{1}{2}} - \mathbf{h}^\star\right\|\right\} \\ &< \text{dist}(\mathbf{z}^{k+\frac{1}{2}}, \mathbf{z}^\star) \le \rho^{k+1}\lambda\left\|\mathbf{z}^\star\right\| < \frac{1}{2}\min\left\{\left|\min_{i\in I_x} x_i^\star\right|, \left|\min_{l\in I_h} h_l^\star\right|\right\}.\end{aligned}$$

Thus, the subsequent hard-thresholding step $\mathcal{T}_s$ again selects the true support I_z with high probability. Therefore,

$$\text{dist}\left(\mathbf{z}^{k+1}, \mathbf{z}^\star\right) \le \left\|\widetilde{\mathbf{z}}^{k+\frac{1}{2}} - \mathbf{z}^\star\right\| \le \rho \cdot \text{dist}(\mathbf{z}^k, \mathbf{z}^\star), \quad 0 < \rho < 1.$$

Now we consider the incoherence conditions (19). As demonstrated by the assumptions in Lemma 10, the key is to show that the whole trajectory lies in the ROC. The difficulty lies in the statistical dependency between the iterates $\{\mathbf{z}^k\}$ and the measurement vectors $\{\mathbf{a}_j\}$. Conditioned on the event of exact support recovery, the ThWF updates operate on the true supports I_x and I_h with high probability. Thus, the iterates are effectively generated by standard gradient descent in $\mathbb{C}^{s_x} \times \mathbb{C}^{s_h}$. We therefore adapt the leave-one-out sequence $\left\{\mathbf{h}_{I_h}^{k,(l)}, \mathbf{x}_{I_x}^{k,(l)}\right\}_{k\ge 0}$ for $1 \le l \le M$ from [31] restricted exclusively to the index sets I_x and I_h. Specifically, the ThWF sequence is from solving the loss function

$$f^{(l)}\left(\mathbf{x}_{I_x}, \mathbf{h}_{I_h}\right) = \sum_{j\ne l}^{M}\left|\mathbf{b}_{j,I_h}^{\mathsf{H}}\mathbf{h}_{I_h}\mathbf{x}_{I_x}^{\mathsf{H}}\mathbf{a}_{j,I_x} - y_j\right|^2.$$

We denote $\mathbf{z}_{I_z}^{k,(l)} = \left[\mathbf{h}_{I_h}^{k,(l)\mathsf{H}}, \mathbf{x}_{I_x}^{k,(l)\mathsf{H}}\right]^{\mathsf{H}}$. Since the thresholding operator preserves the true support at iteration $k+1$ with high probability, the update is mathematically identical to a standard gradient descent step operating in the low-dimensional subspace $\mathbb{C}^{s_x} \times \mathbb{C}^{s_h}$. Consequently, the leave-one-out induction arguments established in [31] apply directly to our reduced-dimension iterates $\mathbf{z}_{I_z}^{k+1,(l)}$ and $\mathbf{z}_{I_z}^k$, which guarantees that the incoherence hypotheses hold for iteration $k+1$. The key idea is to prove: (1) proximity between the original and the leave-one-out iterates, (2) incoherence of the leave-one-out iterate $\mathbf{x}_{I_x}^{k+1,(l)}$ w.r.t. $\mathbf{a}_{l,I_x}$, (3) combining to show incoherence of $\mathbf{x}_{I_x}^{k+1}$ w.r.t. $\mathbf{a}_{l,I_x}$, (4) incoherence of $\mathbf{h}_{I_h}^{k+1}$ w.r.t. $\mathbf{b}_{l,I_h}$.

D Proof of Theorem 3

D.1 Modified sparse spectral initialization

In the noisy case, the frequency-domain measurements are $\tilde{y}_j = y_j + e_j$, where $e_j \sim \mathcal{CN}(0, \sigma^2)$. The constructed adjoint matrix is therefore $\tilde{\mathbf{Z}} = \mathbf{Z} + \mathbf{E}_{noise}$, where the noise matrix is defined as $\mathbf{E}_{noise} = \sum_{j=1}^{M} e_j \mathbf{b}_j \mathbf{a}_j^H$.

Element-wise, the matrix entry is given by

$$\tilde{Z}_{l,i} = h_l^\star x_i^{\star *} + E_{l,i} + E'_{l,i},$$

where $E_{l,i}$ is the perturbation from the sensing matrices in the noiseless case, and $E'_{l,i} = \sum_{j=1}^{M} e_j b_{jl} a_{ji}^*$ is the additive noise perturbation.

From the noiseless analysis in Appendix B, we already know that with probability at least $1 - \mathcal{O}((MN)^{-10})$,

$$\max_{l,i}|E_{l,i}| \le C\left(\sqrt{\frac{\log(MN)}{M}} + \mu\frac{\log(MN)}{M}\right)\|\mathbf{h}^\star\|\|\mathbf{x}^\star\| := \epsilon_{clean}.$$

We now bound the additive noise term $E'_{l,i}$. Notice that $E'_{l,i}$ is a sum of independent, mean-zero random variables $X_j = e_j b_{jl} a^*_{ji}$. Since $e_j \sim \mathcal{CN}(0, \sigma^2)$ and $a_{ji} \sim \mathcal{CN}(0, 1)$ are sub-Gaussian, their product is sub-exponential. Specifically, the sub-exponential norm is bounded by $\|X_j\|_{\psi_1} \lesssim \|e_j\|_{\psi_2} \|b_{jl} a^*_{ji}\|_{\psi_2} \lesssim \sigma \frac{1}{\sqrt{M}} \cdot 1 = \frac{\sigma}{\sqrt{M}}$. The variance parameter for the sum is

$$\sum_{j=1}^{M} \mathbb{E}[|X_j|^2] = \sum_{j=1}^{M} \mathbb{E}[|e_j|^2] |b_{jl}|^2 \mathbb{E}[|a_{ji}|^2] = \sum_{j=1}^{M} \sigma^2 \frac{1}{M} \cdot 1 = \sigma^2.$$

Applying Bernstein's inequality for sub-exponential variables, for any $t > 0$,

$$\mathbb{P}\left(|E'_{l,i}| \geq t\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{\sigma^2}, \frac{t}{\sigma/\sqrt{M}}\right)\right).$$

Setting $t = C' \sigma \sqrt{\log(MN)}$ for a sufficiently large constant $C' > 0$. Since the sample complexity requires $M \gtrsim \log(MN)$, the sub-Gaussian tail $\frac{t^2}{\sigma^2}$ dominates. Taking a union bound over all MN elements, we obtain

$$\max_{l \in [M], i \in [N]} |E'_{l,i}| \leq C' \sigma \sqrt{\log(MN)} := \epsilon_{noise},$$

with probability at least $1 - \mathcal{O}((MN)^{-10})$.

Let the total perturbation bound be $\Delta_{max} = \epsilon_{clean} + \epsilon_{noise}$. Following the support estimation procedure, we compute the maximum marginals $r_i = \max_l |\tilde{Z}_{l,i}|$. For the true support $i \in I_x$, the triangle inequality implies:

$$r_i \geq \min_{l \in I_h} |h^\star_l| \min_{i \in I_x} |x^\star_i| - \Delta_{max} \geq \sqrt{\frac{C_h C_x}{s_h s_x}} \|\mathbf{h}^\star\| \|\mathbf{x}^\star\| - \Delta_{max}.$$

For the off-support indices $i \notin I_x$, the true signal is zero, meaning $r_i \leq \Delta_{max}$.

To guarantee exact support separation, we require $\sqrt{\frac{C_h C_x}{s_h s_x}} \|\mathbf{h}^\star\| \|\mathbf{x}^\star\| > 2\Delta_{max}$. This inequality holds if the signal strictly dominates both the clean perturbation and the noise perturbation. Dominating the clean perturbation ϵ_{clean} yields the same sample complexity as the noiseless case: $M \gtrsim s_h s_x \log(MN)$. Dominating the noise perturbation ϵ_{noise} requires:

$$\min_{l \in I_h} |h^\star_l| \min_{i \in I_x} |x^\star_i| \gtrsim \sigma \sqrt{\log(MN)},$$

which is precisely satisfied by the modified Assumption 3. Thus, the hard-thresholding operation successfully and exactly identifies the true supports I_x and I_h even in the presence of noise.

D.2 Modified local geometry

Lemma 7 does not change. We need to add a bound on the noise, i.e., $\sigma \lesssim \frac{1}{\sqrt{M \log M}}$, such that the noise will not be too big to disturb the RSC and smoothness proved by Lemma 8.

Modified proof of Lemma 9: In the noisy case, α_3 contains a noisy term

$$\alpha_3 \leq \left\| \sum_{j=1}^{M} \mathbf{b}_j^{\mathsf{H}} \left(\mathbf{h}\mathbf{x}^{\mathsf{H}} - \mathbf{h}^\star \mathbf{x}^{\star \mathsf{H}}\right) \mathbf{a}_j \right\| + \left\| \sum_{j=1}^{M} e_j \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \right\|,$$

where the first term is upper bounded by $C/\log M$ for $C > 0$ with high probability, proved in [31]. The noisy term by Lemma 15 is upper bounded by

$$\left\| \sum_{j=1}^{M} e_j \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \right\| \lesssim \sigma \left(\sqrt{M} + \sqrt{N \log M}\right),$$

with probability at least $1 - \mathcal{O}\left(e^{-N} + M^{-10}\right)$. Therefore,

$$\alpha_3 \lesssim \frac{1}{\log M} + \sigma \sqrt{N \log M} \leq 1/96,$$

with high probability, if $\sigma \lesssim \frac{1}{\sqrt{M \log M}}$. Here, we assume $N \asymp M$.

D.3 Modified error contraction

Modified proof of Lemma 10: In the noisy case, the Wirtinger gradient at the ground truth is $\nabla f\left(\mathbf{z}^\star\right)=\left[\begin{array}{c}-\sum_{j=1}^M e_j\mathbf{b}_j\mathbf{a}_j^{\mathsf{H}}\mathbf{x}^\star\\ -\sum_{j=1}^M e_j\mathbf{a}_j\mathbf{h}^{\star\mathsf{H}}\mathbf{b}_j\end{array}\right]\neq\mathbf{0}$. Equation (207) in [31] becomes

$$\left[\begin{array}{c}\nabla_{\mathbf{h}}f\left(\widetilde{\mathbf{z}}^k\right)\\ \nabla_{\mathbf{x}}f\left(\widetilde{\mathbf{z}}^k\right)\\ \nabla_{\mathbf{h}^*}f\left(\widetilde{\mathbf{z}}^k\right)\\ \nabla_{\mathbf{x}^*}f\left(\widetilde{\mathbf{z}}^k\right)\end{array}\right]=\underbrace{\int_0^1\nabla^2 f\left(\mathbf{z}\left(\tau\right)\right)d\tau}_{:=\mathbf{C}}\left[\begin{array}{c}\widetilde{\mathbf{h}}^k-\mathbf{h}^\star\\ \widetilde{\mathbf{x}}^k-\mathbf{x}^\star\\ (\widetilde{\mathbf{h}}^k-\mathbf{h}^\star)^*\\ (\widetilde{\mathbf{x}}^k-\mathbf{x}^\star)^*\end{array}\right]+\left[\begin{array}{c}\nabla_{\mathbf{h}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{x}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{h}^*}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{x}^*}f\left(\mathbf{z}^\star\right)\end{array}\right],$$

where $\mathbf{z}\left(\tau\right):=\mathbf{z}^\star+\tau\left(\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\right)$. The modified equation (208) in [31] is thus

$$\left[\begin{array}{c}\widehat{\mathbf{z}}^{k+1}-\mathbf{z}^\star\\ (\widehat{\mathbf{z}}^{k+1}-\mathbf{z}^\star)^*\end{array}\right]=(\mathbf{I}-\eta\mathbf{D}\mathbf{C})\left[\begin{array}{c}\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\\ (\widetilde{\mathbf{z}}^k-\mathbf{z}^\star)^*\end{array}\right]-\eta\mathbf{D}\left[\begin{array}{c}\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{z}^*}f\left(\mathbf{z}^\star\right)\end{array}\right],$$

where $\mathbf{D}=\operatorname{diag}\left(\left[\begin{array}{cccc}\frac{\mathbf{I}_N}{\|\widetilde{\mathbf{x}}^k\|^2} & \frac{\mathbf{I}_N}{\|\widetilde{\mathbf{h}}^k\|^2} & \frac{\mathbf{I}_N}{\|\widetilde{\mathbf{x}}^k\|^2} & \frac{\mathbf{I}_N}{\|\widetilde{\mathbf{h}}^k\|^2}\end{array}\right]\right)$ is a block diagonal matrix. Take the squared Euclidean norm of both sides of the above equation to reach

$$\begin{aligned}\left\|\widehat{\mathbf{z}}^{k+1}-\mathbf{z}^\star\right\|^2=&\frac{1}{2}\left[\begin{array}{c}\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\\ (\widetilde{\mathbf{z}}^k-\mathbf{z}^\star)^*\end{array}\right]^{\mathsf{H}}\left(\mathbf{I}+\eta^2\mathbf{C}\mathbf{D}^2\mathbf{C}-\eta\left(\mathbf{D}\mathbf{C}+\mathbf{C}\mathbf{D}\right)\right)\left[\begin{array}{c}\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\\ (\widetilde{\mathbf{z}}^k-\mathbf{z}^\star)^*\end{array}\right]\\ &-\operatorname{Re}\left(\left[\begin{array}{c}\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{z}^*}f\left(\mathbf{z}^\star\right)\end{array}\right]^{\mathsf{H}}\left(\eta\mathbf{D}-\eta^2\mathbf{D}^2\mathbf{C}\right)\left[\begin{array}{c}\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\\ (\widetilde{\mathbf{z}}^k-\mathbf{z}^\star)^*\end{array}\right]\right)\\ &+\frac{1}{2}\left[\begin{array}{c}\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{z}^*}f\left(\mathbf{z}^\star\right)\end{array}\right]^{\mathsf{H}}\eta^2\mathbf{D}^2\left[\begin{array}{c}\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{z}^*}f\left(\mathbf{z}^\star\right)\end{array}\right],\end{aligned}\tag{22}$$

where$\|\mathbf{C}\|\leq 3$ and $\|\mathbf{D}\|\leq 1+\delta$, invoked by Lemma 7. The first term is the same as that in (209) in [31] and implies that

$$\left\|\widehat{\mathbf{z}}^{k+1}-\mathbf{z}^\star\right\|^2\leq\left(1+16\eta^2-\frac{\eta}{4}\right)\left\|\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\right\|^2.$$

We derive the upper bound for the second term. By the Young's inequality, for any $\epsilon>0$,

$$\begin{aligned}-\eta\operatorname{Re}\left(\left[\begin{array}{c}\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{z}^*}f\left(\mathbf{z}^\star\right)\end{array}\right]^{\mathsf{H}}\mathbf{D}\left[\begin{array}{c}\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\\ (\widetilde{\mathbf{z}}^k-\mathbf{z}^\star)^*\end{array}\right]\right)&\leq 2\eta\left\|\mathbf{D}\right\|\left\|\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\right\|\left\|\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\right\|\\ &\leq\eta\left\|\mathbf{D}\right\|\left(\epsilon\left\|\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\right\|^2+\frac{1}{\epsilon}\left\|\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\right\|^2\right),\end{aligned}$$

and by the Cauchy–Schwarz inequality,

$$\begin{aligned}\eta^2\operatorname{Re}\left(\left[\begin{array}{c}\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{z}^*}f\left(\mathbf{z}^\star\right)\end{array}\right]^{\mathsf{H}}\mathbf{D}^2\mathbf{C}\left[\begin{array}{c}\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\\ (\widetilde{\mathbf{z}}^k-\mathbf{z}^\star)^*\end{array}\right]\right)&\leq 2\eta^2\left\|\mathbf{D}\right\|^2\left\|\mathbf{C}\right\|\left\|\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\right\|\left\|\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\right\|\\ &\leq\eta^2\left\|\mathbf{D}\right\|^2\left\|\mathbf{C}\right\|\left(\left\|\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\right\|^2+\left\|\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\right\|^2\right).\end{aligned}$$

For the third term,

$$\frac{1}{2}\left[\begin{array}{c}\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{z}^*}f\left(\mathbf{z}^\star\right)\end{array}\right]^{\mathsf{H}}\eta^2\mathbf{D}^2\left[\begin{array}{c}\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\\ \nabla_{\mathbf{z}^*}f\left(\mathbf{z}^\star\right)\end{array}\right]\leq\eta^2\left\|\mathbf{D}\right\|^2\left\|\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\right\|^2\leq 2\eta^2\left\|\nabla_{\mathbf{z}}f\left(\mathbf{z}^\star\right)\right\|^2.$$

Substitution into (22) indicates that

$$\begin{aligned}\left\|\widehat{\mathbf{z}}^{k+1}-\mathbf{z}^\star\right\|^2\leq&\underbrace{\left(1+\left(16+\left\|\mathbf{D}\right\|^2\left\|\mathbf{C}\right\|\right)\eta^2-\left(\frac{1}{4}-\left\|\mathbf{D}\right\|\epsilon\right)\eta\right)}_{:=\rho_\epsilon}\left\|\widetilde{\mathbf{z}}^k-\mathbf{z}^\star\right\|^2\\ &+\left(\eta^2\left\|\mathbf{D}\right\|^2\left\|\mathbf{C}\right\|+\eta\frac{\left\|\mathbf{D}\right\|}{\epsilon}\right)\left\|\nabla f\left(\mathbf{z}^\star\right)\right\|^2.\end{aligned}$$

If we take $0 < \epsilon < \frac{1}{4}$, we have $\frac{1}{4} - \|\mathbf{D}\|\epsilon > 0$, which makes $\rho_\epsilon < 1$ possible. Since $\|\mathbf{C}\| \le 3$ and $\|\mathbf{D}\| \le 1 + \delta < 2$, if we take $\epsilon = \frac{1}{16}$ and $0 < \eta \le \frac{1}{128}$,

$$\rho_\epsilon = 1 - \frac{\eta}{8} + \mathcal{O}\left(\eta^2\right) \le 1 - \frac{\eta}{16}, \quad \eta^2 \|\mathbf{D}\|^2 \|\mathbf{C}\| + \eta\frac{\|\mathbf{D}\|}{\epsilon} \le 32\eta + \mathcal{O}\left(\eta^2\right).$$

By Lemma 16, we have

$$\left\|\widetilde{\mathbf{z}}^{k+1} - \mathbf{z}^\star\right\|^2 \le \left\|\widehat{\mathbf{z}}^{k+1} - \mathbf{z}^\star\right\|^2 \le \left(1 - \frac{\eta}{16}\right)\left\|\widetilde{\mathbf{z}}^k - \mathbf{z}^\star\right\|^2 + C\eta\sigma^2(M+N).$$

For $M \asymp N$,

$$\mathrm{dist}^2\left(\mathbf{z}^{k+1}, \mathbf{z}^\star\right) \le (1 - \eta/16)\,\mathrm{dist}^2\left(\mathbf{z}^k, \mathbf{z}^\star\right) + \mathcal{O}\left(\eta\sigma^2 M\right).$$

E Supporting Lemmas

E.1 Basics

The standard Gaussian concentration inequality and the union bound give

$$\max_{1\le j\le M}\left|\mathbf{a}_j^{\mathsf{H}}\mathbf{x}^\star\right| \le 5\sqrt{\log M}\,\|\mathbf{x}^\star\|, \tag{23}$$

$$\max_{1\le j\le M}\sup_{\|\mathbf{x}\|_0\le s}\left|\mathbf{a}_j^{\mathsf{H}}\left(\mathbf{x} - \mathbf{x}^\star\right)\right| \le 5\left(\sqrt{\log M} + \sqrt{s\log\left(N/s\right)}\right)\|\mathbf{x} - \mathbf{x}^\star\|, \tag{24}$$

$$\max_{1\le j\le M}|a_j| \le 5\sqrt{\log M}, \tag{25}$$

with probability at least $1 - \mathcal{O}\left(M^{-10}\right)$. In addition, with probability exceeding $1 - CM\exp\left(-cN\right)$ for some constants $c, C > 0$,

$$\max_{1\le j\le M}\|\mathbf{a}_j\| \le 3\sqrt{N}. \tag{26}$$

Lemma 12. *Suppose $\mathbf{a} \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{I}_N\right)$ and $\mathbf{x}^\star$ is a constant vector. Then, (1) $\mathbb{E}\left[\left|\mathbf{a}^{\mathsf{H}}\mathbf{x}^\star\right|^2\|\mathbf{a}\|^2\right] = (N+1)\|\mathbf{x}^\star\|^2$, (2) $\mathbb{E}\left[\left|\mathbf{a}^{\mathsf{H}}\mathbf{x}^\star\right|^2|a_i|^2\right] = \|\mathbf{x}^\star\|^2 + |x_i^\star|^2$.*

E.2 Concentration inequalities

Lemma 13. *[40, 41](Bernstein inequality) Given M zero-mean, independent random variables $X_j \in \mathbb{C}^N$. Let $\sigma^2 = \sum_{j=1}^M \mathbb{E}|X_j|^2$ and $|X_j| \le R$. For any $t > 0$, the following inequality holds:*

$$\mathbb{P}\left(\left|\sum_{j=1}^M X_j\right| \ge t\right) \le N\exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).$$

Based on Lemma 13, we can prove the following matrix concentration inequality.

Lemma 14. *Let $J_x \subset [N]$ and $J_h \subset [M]$ be any fixed subsets with $|J_x| \le 2s_x$ and $|J_h| \le 2s_h$. With probability at least $1 - M^{-\gamma}$, the following uniform bound holds simultaneously for all such subset pairs:*

$$\|\mathbf{Z}_{J_h,J_x} - \mathbb{E}\left[\mathbf{Z}_{J_h,J_x}\right]\| \le \xi\|\mathbf{x}^\star\|\|\mathbf{h}^\star\|,$$

if the number of measurements satisfies $M \ge C_\gamma\mu^2\left(s_h + s_x\right)\left(s_h\log M + s_x\log N\right)/\xi^2$.

Consequently, this bound applies to the specific data-dependent subsets $I_x^3 = I_x \cup (\hat{I}_x \cap I_x^c)$ and $I_h^3 = I_h \cup (\hat{I}_h \cap I_h^c)$, since $|I_x^3| \le 2s_x$ and $|I_h^3| \le 2s_h$.

Proof. For any fixed pair of subsets J_h and J_x, we have $\mathbb{E}\left[\mathbf{Z}_{J_h,J_x}\right] = \mathbf{h}_{J_h}^\star\mathbf{x}_{J_x}^{\star\mathsf{H}}$. We will use the matrix Bernstein inequality to bound the deviation

$$\left\|\mathbf{Z}_{I_h^3,I_x^3} - \mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\right\| \le \xi\|\mathbf{x}^\star\|\|\mathbf{h}^\star\|.$$

By definition of $\mathcal{A}^*$,

$$\begin{aligned}\mathbf{Z}_{J_h,J_x}-\mathbf{h}^\star_{J_h}\mathbf{x}^{\star\mathsf{H}}_{J_x}&=\sum_{j=1}^{M}\mathbf{b}_j^{\mathsf{H}}\mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\mathbf{a}_j\cdot\mathbf{a}_{j,J_x}\mathbf{b}^{\mathsf{H}}_{j,J_h}-\mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\\&=\sum_{j=1}^{M}\left(\mathbf{a}_{j,J_x}\mathbf{a}^{\mathsf{H}}_{j,J_x}-\mathbf{I}_{|J_x|}\right)\mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\mathbf{b}_{j,J_h}\mathbf{b}^{\mathsf{H}}_{j,J_h}.\end{aligned}$$

Define $\mathbf{Z}_j=\left(\mathbf{a}_{j,J_x}\mathbf{a}^{\mathsf{H}}_{j,J_x}-\mathbf{I}_{|J_x|}\right)\mathbf{h}^\star\mathbf{x}^{\star\mathsf{H}}\mathbf{b}_{j,J_h}\mathbf{b}^{\mathsf{H}}_{j,J_h}$. In order to apply the Bernstein inequality, we need to estimate both the exponential norm $\|\mathbf{Z}_j\|_{\psi_1}$ and the variance $\sigma_0^2=\max\left\{\left\|\mathbb{E}\left[\sum_{j=1}^M\mathbf{Z}_j^{\mathsf{H}}\mathbf{Z}_j\right]\right\|,\left\|\mathbb{E}\left[\sum_{j=1}^M\mathbf{Z}_j\mathbf{Z}_j^{\mathsf{H}}\right]\right\|\right\}$.

$$\begin{aligned}\|\mathbf{Z}_j\|_{\psi_1}&\lesssim\left|\mathbf{h}^{\star\mathsf{H}}\mathbf{b}_{j,J_h}\right|\|\mathbf{b}_{j,J_h}\|\left\|\left(\mathbf{a}_{j,J_x}\mathbf{a}^{\mathsf{H}}_{j,J_x}-\mathbf{I}_{2s_x}\right)\mathbf{x}^\star\right\|_{\psi_1}\\&\lesssim\frac{\mu\|\mathbf{h}^\star\|}{\sqrt{M}}\sqrt{\frac{s_h}{M}}\left\|\left(\mathbf{a}_{j,J_x}\mathbf{a}^{\mathsf{H}}_{j,J_x}-\mathbf{I}_{2s_x}\right)\mathbf{x}^\star\right\|_{\psi_1},\end{aligned}$$

where Lemma 6.4 in [24] gives

$$\left\|\left(\mathbf{a}_{j,J_x}\mathbf{a}^{\mathsf{H}}_{j,J_x}-\mathbf{I}_{2s_x}\right)\mathbf{x}^\star\right\|_{\psi_1}\lesssim\|\mathbf{x}^\star\|\sqrt{s_x}.$$

So

$$\|\mathbf{Z}_j\|_{\psi_1}\lesssim\frac{\mu\sqrt{s_xs_h}}{M}\|\mathbf{x}^\star\|\|\mathbf{h}^\star\|.$$

For the variance terms, we calculate

$$\begin{aligned}\left\|\sum_{j=1}^M\mathbb{E}\left[\mathbf{Z}_j^{\mathsf{H}}\mathbf{Z}_j\right]\right\|\leq&\left\|\sum_{j=1}^M\mathbb{E}\left[\|\mathbf{a}_{j,J_x}\|^2\left|\mathbf{a}^{\mathsf{H}}_{j,J_x}\mathbf{x}^\star\right|^2\right]\left|\mathbf{h}^{\star\mathsf{H}}\mathbf{b}_{j,J_h}\right|^2\mathbf{b}_{j,J_h}\mathbf{b}^{\mathsf{H}}_{j,J_h}\right\|+\frac{1}{M}\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2\\&+\frac{1}{M}\left\|\sum_{j=1}^M\mathbb{E}\left[\left|\mathbf{a}^{\mathsf{H}}_{j,J_x}\mathbf{x}^\star\right|^2\right]\left(\mathbf{b}^{\mathsf{H}}_{j,J_h}\mathbf{h}^\star\mathbf{b}_{j,J_h}\mathbf{h}^{\star\mathsf{H}}+\mathbf{h}^{\star\mathsf{H}}\mathbf{b}_{j,J_h}\mathbf{h}^\star\mathbf{b}^{\mathsf{H}}_{j,J_h}\right)\right\|\\\leq&\left(|J_x|+1\right)\|\mathbf{x}^\star\|^2\cdot\frac{\mu^2}{M}\|\mathbf{h}^\star\|^2+\frac{1}{M}\|\mathbf{x}^\star\|^2\cdot2\|\mathbf{h}^\star\|^2+\frac{1}{M}\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2\\\leq&\frac{\mu^2\left(2s_x+1\right)+3}{M}\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2,\end{aligned}$$

where the second inequality is due to Lemma 12. Moreover, we have

$$\begin{aligned}\left\|\sum_{j=1}^M\mathbb{E}\left[\mathbf{Z}_j\mathbf{Z}_j^{\mathsf{H}}\right]\right\|\leq&\left\|\sum_{j=1}^M\|\mathbf{b}_{j,J_h}\|^2\left|\mathbf{h}^{\star\mathsf{H}}\mathbf{b}_{j,J_h}\right|^2\mathbb{E}\left[\left|\mathbf{a}^{\mathsf{H}}_{j,J_x}\mathbf{x}^\star\right|^2\mathbf{a}_{j,J_x}\mathbf{a}^{\mathsf{H}}_{j,J_x}\right]\right\|+\frac{1}{M}\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2\\&+\frac{1}{M}\left\|\sum_{j=1}^M\left|\mathbf{h}^{\star\mathsf{H}}\mathbf{b}_{j,J_h}\right|^2\mathbb{E}\left[\mathbf{a}^{\mathsf{H}}_{j,J_x}\mathbf{x}^\star\mathbf{a}_{j,J_x}\mathbf{x}^{\star\mathsf{H}}+\mathbf{x}^{\star\mathsf{H}}\mathbf{a}_{j,J_x}\mathbf{x}^\star\mathbf{a}^{\mathsf{H}}_{j,J_x}\right]\right\|\\\leq&\frac{2\mu^2s_h}{M}\|\mathbf{h}^\star\|^2\left\|\|\mathbf{x}^\star\|^2\mathbf{I}_{|J_x|}+\mathbf{x}^\star_{J_x}\mathbf{x}^{\star\top}_{J_x}\right\|+\frac{1}{M}\|\mathbf{h}^\star\|^2\left\|2\mathbf{x}^\star\mathbf{x}^{\star\mathsf{H}}\right\|+\frac{1}{M}\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2\\=&\frac{2\mu^2s_h+3}{M}\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2.\end{aligned}$$

Therefore, $\sigma_0^2\lesssim\frac{2\mu^2(s_h+s_x)}{M}\|\mathbf{h}^\star\|^2\|\mathbf{x}^\star\|^2$. We apply the Theorem 6.2 in [24] by choosing $t=\gamma\log M$, and obtain

$$\begin{aligned}\left\|\sum_{j=1}^MZ_j\right\|\leq&C_0\|\mathbf{x}^\star\|\|\mathbf{h}^\star\|\max\{\sqrt{\frac{\mu^2\left(s_h+s_x\right)}{M}\left(\gamma\log M+\log\left(s_h+s_x\right)\right)},\\&\frac{\sqrt{\mu^2s_xs_h}}{M}\log\left(\sqrt{\frac{s_xs_h}{\left(s_h+s_x\right)}}\right)\left(\gamma\log M+\log\left(s_h+s_x\right)\right)\}\\\leq&\xi\|\mathbf{x}^\star\|\|\mathbf{h}^\star\|,\end{aligned}$$

with probability at least $1-M^{-\gamma}$ if $M \geq C_\gamma \mu^2 (s_h + s_x) \log M/\xi^2$. To resolve the data dependency of the estimated supports, we apply a union bound over all possible subset pairs J_h and J_x bounded by sizes $2s_h$ and $2s_x$. The total number of such subset combinations is bounded by:

$$\sum_{k=1}^{2s_h} \binom{M}{k} \sum_{l=1}^{2s_x} \binom{N}{l} \leq M^{2s_h} N^{2s_x} = \exp\left(2s_h \log M + 2s_x \log N\right).$$

Applying Theorem 6.2 in [24] for a fixed pair, we choose $t = (\gamma+2)(s_h \log M + s_x \log N)$. By the union bound, the uniform failure probability is bounded by $\exp(2s_h \log M + 2s_x \log N)\exp(-t) \leq M^{-\gamma}$. The uniform deviation satisfies:

$$\begin{aligned}\left\|\sum_{j=1}^{M} \mathbf{Z}_j\right\| \leq & C_0 \|\mathbf{x}^\star\| \|\mathbf{h}^\star\| \max\left\{\sqrt{\frac{\mu^2(s_h+s_x)}{M}}t,\ \frac{\sqrt{\mu^2 s_x s_h}}{M}\log\left(\sqrt{\frac{s_x s_h}{s_h+s_x}}\right)t\right\} \\ \leq & \xi \|\mathbf{x}^\star\| \|\mathbf{h}^\star\|,\end{aligned}$$

which holds true provided the sample complexity scales as $M \geq C_\gamma \mu^2 (s_h+s_x)(s_h \log M + s_x \log N)/\xi^2$. □

Lemma 15. *In the noisy case, the following holds*

$$\left\|\sum_{j=1}^{M} e_j \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}}\right\| \leq \sigma\left(\sqrt{M} + \sqrt{N \log M}\right),$$

with probability at least $1 - \mathcal{O}\left(e^{-N} + M^{-10}\right)$*, for some constant* $C > 0$*.*

Proof. Define the diagonal matrix $\mathbf{N} = \mathrm{diag}(e_1, \ldots, e_M)$ and let $\mathbf{A} \in \mathbb{C}^{M\times N}$ be a standard complex Gaussian matrix with rows $\mathbf{a}_j^{\mathsf{H}}$. Let $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_M]$ denote the unitary matrix of the basis. The sum can be expressed compactly as:

$$\mathbf{E} = \sum_{j=1}^{M} e_j \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} = \mathbf{BNA}. \tag{27}$$

By the unitary invariance of the operator norm, we have $\|\mathbf{E}\| = \|\mathbf{BNA}\| = \|\mathbf{NA}\|$.

Conditioning on the noise vector $\mathbf{n}$ (and thus on $\mathbf{N}$), the matrix $\mathbf{X} = \mathbf{NA}$ has independent complex Gaussian entries $X_{ji} \sim \mathcal{CN}(0, |e_j|^2)$. By the Bandeira-van Handel bound [4] for the expected operator norm of Gaussian matrices with independent entries, there exists a universal constant $C > 0$ such that:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{NA}\| \mid \mathbf{N}\right] &\leq C\left(\max_i \sqrt{\sum_{j=1}^{M} \mathbb{E}\left[|X_{ji}|^2 \mid \mathbf{N}\right]} + \max_j \sqrt{\sum_{i=1}^{N} \mathbb{E}\left[|X_{ji}|^2 \mid \mathbf{N}\right]}\right) \\ &= C\left(\|\mathbf{e}\|_2 + \sqrt{N}\|\mathbf{e}\|_\infty\right).\end{aligned} \tag{28}$$

For a fixed $\mathbf{N}$, the mapping $\mathbf{A} \mapsto \|\mathbf{NA}\|$ is Lipschitz continuous with respect to the Frobenius norm with Lipschitz constant $L = \|\mathbf{N}\| = \|\mathbf{e}\|_\infty$. By the Gaussian concentration inequality, for any $t > 0$, we have:

$$\mathbb{P}\left(\|\mathbf{NA}\| \geq \mathbb{E}\left[\|\mathbf{NA}\| \mid \mathbf{N}\right] + \|\mathbf{e}\|_\infty \sqrt{2t} \;\middle|\; \mathbf{N}\right) \leq e^{-t}. \tag{29}$$

We now control the norms of $\mathbf{e} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_M)$. Standard concentration for the ℓ_2-norm of a Gaussian vector implies that $\|\mathbf{e}\|_2 \leq \sigma(\sqrt{M} + \sqrt{t})$ with probability at least $1 - e^{-t}$. Furthermore, by a union bound over the M components, the ℓ_∞-norm satisfies $\|\mathbf{e}\|_\infty \leq \sigma\sqrt{\log M + t}$ with probability at least $1 - e^{-t}$.

By applying a union bound over the concentration events for $\mathbf{A}$ and $\mathbf{w}$, and setting t proportional to $\max\{N, \log M\}$, we obtain that with high probability:

$$\|\mathbf{E}\| \lesssim \sigma\left(\sqrt{M} + \sqrt{N \log M}\right). \tag{30}$$

□

Lemma 16. *In the noisy case, the squared norm of $\nabla f(\mathbf{z}^\star)$ satisfies*

$$\|\nabla f(\mathbf{z}^\star)\|^2 \le C\sigma^2(M+N),$$

with probability at least $1-\mathcal{O}(\exp(-M)+\exp(-N))$, for some constant $C>0$.

Proof. Without loss of generality, we assume $\|\mathbf{h}^\star\| = \|\mathbf{x}^\star\| = 1$. Recall that the Wirtinger gradient of the loss function evaluated at the ground truth is

$$\nabla f(\mathbf{z}^\star) = \begin{bmatrix} -\sum_{j=1}^M e_j \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \mathbf{x}^\star \\ -\sum_{j=1}^M e_j \mathbf{a}_j \mathbf{h}^{\star\mathsf{H}} \mathbf{b}_j \end{bmatrix},$$

where $n_j \sim \mathcal{CN}(0,\sigma^2)$, $\mathbf{a}_j \sim \mathcal{CN}(\mathbf{0},\mathbf{I}_N)$, and $\mathbf{b}_j$'s satisfy $\|\mathbf{b}_j\|^2 = 1$ and $\sum_{j=1}^m \mathbf{b}_j\mathbf{b}_j^{\mathsf{H}} = \mathbf{I}_M$. We aim to bound

$$\|\nabla f(\mathbf{z}^\star)\|^2 = \|\nabla_{\mathbf{h}} f(\mathbf{z}^\star)\|^2 + \|\nabla_{\mathbf{x}} f(\mathbf{z}^\star)\|^2.$$

We first establish a high-probability upper bound for $\|\nabla_{\mathbf{h}} f(\mathbf{z}^\star)\| = \left\| -\sum_{j=1}^M e_j \mathbf{b}_j \mathbf{a}_j^{\mathsf{H}} \mathbf{x}^\star \right\|$. Let us condition on the randomness of $\{\mathbf{a}_j\}_{j=1}^M$. Define the random variables $\xi_j = \mathbf{a}_j^{\mathsf{H}}\mathbf{x}^\star \sim \mathcal{CN}(0,1)$. Conditional on $\boldsymbol{\xi} = [\xi_1,\dots,\xi_M]^\top$, $\nabla_{\mathbf{h}} f(\mathbf{z}^\star)$ is a zero-mean complex Gaussian random vector. Specifically, define $w_j = e_j \cdot \mathbf{a}_j^{\mathsf{H}}\mathbf{x}^\star$, we have $w_j \mid \xi_j \sim \mathcal{CN}\left(0,\sigma^2|\xi_j|^2\right)$. Consequently, the conditional covariance matrix of $\nabla_{\mathbf{h}} f(\mathbf{z}^\star)$ conditional on $\boldsymbol{\xi}$ is given by

$$\boldsymbol{\Sigma}_{\nabla_{\mathbf{h}} f|\boldsymbol{\xi}} = \sigma^2 \sum_{j=1}^M |\xi_j|^2 \mathbf{b}_j\mathbf{b}_j^{\mathsf{H}}.$$

Taking expectation with respect to $\boldsymbol{\xi}$ yields

$$\mathbb{E}\left[\boldsymbol{\Sigma}_{\nabla_{\mathbf{h}} f|\boldsymbol{\xi}}\right] = \sigma^2 \sum_{j=1}^M \mathbb{E}\left[|\xi_j|^2\right] \mathbf{b}_j\mathbf{b}_j^{\mathsf{H}} = \sigma^2 \sum_{j=1}^M \mathbf{b}_j\mathbf{b}_j^{\mathsf{H}} = \sigma^2\mathbf{I}_M.$$

This shows that $\nabla_{\mathbf{h}} f(\mathbf{z}^\star)$ is an isotropic random vector with covariance on the order of $\sigma^2\mathbf{I}_M$. Standard concentration results for Gaussian mixture random vectors then imply that its squared norm concentrates sharply around its mean. In particular, there exist universal constants $c, C > 0$ such that

$$\|\nabla_{\mathbf{h}} f(\mathbf{z}^\star)\|^2 \le C\sigma^2 M,$$

holds with probability at least $1-\exp(-cM)$.

We next bound $\|\nabla_{\mathbf{x}} f(\mathbf{z}^\star)\| = \left\| -\sum_{j=1}^M e_j \mathbf{a}_j \mathbf{h}^{\star\mathsf{H}} \mathbf{b}_j \right\|$. Let us condition on the randomness of the noise $\mathbf{n}$. Define the deterministic scalars $c_j = \mathbf{h}^{\star\mathsf{H}}\mathbf{b}_j$. Due to the unitary property of $\mathbf{B}$ and our assumption, we have $\sum_{j=1}^M |c_j|^2 = \|\mathbf{h}^\star\|^2 = 1$. Conditional on $\mathbf{n}$ (and hence $\mathbf{e} = [e_1,\dots,e_M]^\top$), $\nabla_{\mathbf{x}} f(\mathbf{z}^\star) = -\sum_{j=1}^M e_j c_j \mathbf{a}_j$ is a zero-mean complex Gaussian random vector, because it is a linear combination of the independent Gaussian vectors $\mathbf{a}_j \sim \mathcal{CN}(\mathbf{0},\mathbf{I}_N)$. Consequently, the conditional covariance matrix of $\nabla_{\mathbf{x}} f(\mathbf{z}^\star)$ is given by

$$\boldsymbol{\Sigma}_{\nabla_{\mathbf{x}} f|\mathbf{n}} = \sum_{j=1}^M |e_j c_j|^2 \mathbf{I}_N = \left(\sum_{j=1}^M |e_j|^2 |c_j|^2\right) \mathbf{I}_N.$$

Taking the expectation with respect to $\mathbf{n}$ yields

$$\mathbb{E}\left[\boldsymbol{\Sigma}_{\nabla_{\mathbf{x}} f|\mathbf{n}}\right] = \sum_{j=1}^M \mathbb{E}\left[|e_j|^2\right] |c_j|^2 \mathbf{I}_N = \sigma^2 \sum_{j=1}^M |c_j|^2 \mathbf{I}_N = \sigma^2\mathbf{I}_N.$$

Similar to the first part, this shows that unconditionally, $\nabla_{\mathbf{x}} f(\mathbf{z}^\star)$ is an isotropic random vector with covariance on the order of $\sigma^2\mathbf{I}_N$. Standard concentration results dictate that its squared norm concentrates sharply around its mean. Thus, there exists a universal constant $C > 0$ such that

$$\|\nabla_{\mathbf{x}} f(\mathbf{z}^\star)\|^2 \le C\sigma^2 N,$$

holds with probability at least $1-\exp\left(-cN\right)$.

Combining the high-probability bounds for $\left\|\nabla_{\mathbf{h}} f\left(\mathbf{z}^{\star}\right)\right\|^2$ and $\left\|\nabla_{\mathbf{x}} f\left(\mathbf{z}^{\star}\right)\right\|^2$, and taking a union bound, we have

$$\left\|\nabla f\left(\mathbf{z}^{\star}\right)\right\|^2=\left\|\nabla_{\mathbf{h}} f\left(\mathbf{z}^{\star}\right)\right\|^2+\left\|\nabla_{\mathbf{x}} f\left(\mathbf{z}^{\star}\right)\right\|^2 \leq C\sigma^2(M+N),$$

with probability at least $1-\exp\left(-cM\right)-\exp\left(-cN\right)$. □

Lemma 17. *(RIP of $\mathcal{A}$ on sparse, rank-one matrices) There exist absolute numerical constants $C>0$ and $\eta<1$ such that the following holds. Let $\mathbf{C}\in\mathbb{C}^{M\times N}$ be a random matrix whose entries are i.i.d. following $\mathcal{CN}\left(0,1/M\right)$. Define a linear operator $\mathcal{A}:\mathbb{C}^{N\times M}\to\mathbb{C}^{M}$ such that $\mathcal{A}(\mathbf{x}\mathbf{h}^{\mathsf{H}})=\mathbf{y}$. If $M\gg C_\delta\left(\mu^2 s_x\log N+s_h\log^2 M\right)$, then, with probability $1-\mathcal{O}\left(M^{-10}\right)$,*

$$\left|\left\langle\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}},\left(\mathcal{A}^{*}\mathcal{A}-\mathbf{I}\right)\mathbf{x}\mathbf{h}^{\mathsf{H}}\right\rangle\right|\leq\delta\left\|\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}}\right\|_F\left\|\mathbf{x}\mathbf{h}^{\mathsf{H}}\right\|_F,$$

for all s_x-sparse $\mathbf{x}^{\star},\mathbf{x}$ and s_h-sparse, μ-spectrally flat $\mathbf{h}^{\star},\mathbf{h}$. Here, $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$.

Proof. Define $\mathbf{R}_{x,h}\in\mathbb{C}^{N\times MN}$ and $\boldsymbol{\xi}\in\mathbb{C}^{MN}$ by

$$\mathbf{R}_{x,h}:=\mathbf{x}^{\top}\otimes\mathrm{Diag}\left(\mathbf{B}^{\mathsf{H}}\mathbf{h}\right),\quad\boldsymbol{\xi}:=\sqrt{M}\left(\mathbf{I}_N\otimes\mathbf{B}^{\mathsf{H}}\right)\mathrm{vec}\left(\mathbf{C}\right).$$

Then, we have

$$\begin{aligned}\mathbf{R}_{x,h}\boldsymbol{\xi}&=\left[x_1\mathrm{Diag}\left(\mathbf{B}^{\mathsf{H}}\mathbf{h}\right),\ldots,x_N\mathrm{Diag}\left(\mathbf{B}^{\mathsf{H}}\mathbf{h}\right)\right]\cdot\sqrt{M}\mathrm{vec}\left(\mathbf{B}^{\mathsf{H}}\mathbf{C}\right)\\&=\sqrt{M}\sum_{i=1}^{N}x_i\mathrm{Diag}\left(\mathbf{B}^{\mathsf{H}}\mathbf{h}\right)\left(\mathbf{B}^{\mathsf{H}}\mathbf{C}\right)_i=\mathcal{A}\left(\mathbf{x}\mathbf{h}^{\mathsf{H}}\right).\end{aligned}$$

Therefore, we have

$$\left\langle\mathcal{A}\left(\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}}\right),\mathcal{A}\left(\mathbf{x}\mathbf{h}^{\mathsf{H}}\right)\right\rangle=\left\langle\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}},\mathcal{A}^{*}\mathcal{A}\left(\mathbf{x}\mathbf{h}^{\mathsf{H}}\right)\right\rangle=\left\langle\mathbf{R}_{x^{\star},h^{\star}}\boldsymbol{\xi},\mathbf{R}_{x,h}\boldsymbol{\xi}\right\rangle,$$

where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$. Moreover, we have

$$\mathbb{E}_{\mathbf{C}}\left[\left\langle\mathbf{R}_{x^{\star},h^{\star}}\boldsymbol{\xi},\mathbf{R}_{x,h}\boldsymbol{\xi}\right\rangle\right]=\left\langle\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}},\mathbb{E}_{\mathbf{C}}\mathcal{A}^{*}\mathcal{A}\left(\mathbf{x}\mathbf{h}^{\mathsf{H}}\right)\right\rangle=\left\langle\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}},\mathbf{h}\mathbf{x}^{\mathsf{H}}\right\rangle.$$

By the triangle inequality, we have

$$\begin{aligned}&\left|\left\langle\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}},\left(\mathcal{A}^{*}\mathcal{A}-\mathbf{I}\right)\left(\mathbf{x}\mathbf{h}^{\mathsf{H}}\right)\right\rangle\right|\\&\leq\left|\left\langle\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}},\left(\mathcal{A}^{*}\mathcal{A}-\mathbb{E}_{\mathbf{C}}\left(\mathcal{A}^{*}\mathcal{A}\right)\right)\left(\mathbf{x}\mathbf{h}^{\mathsf{H}}\right)\right\rangle\right|+\left|\left\langle\mathbf{x}^{\star}\mathbf{h}^{\star\mathsf{H}},\left(\mathbb{E}_{\mathbf{C}}\left[\mathcal{A}^{*}\mathcal{A}\right]-\mathbf{I}\right)\left(\mathbf{x}\mathbf{h}^{\mathsf{H}}\right)\right\rangle\right|\\&=\left|\left\langle\mathbf{R}_{x^{\star},h^{\star}}\boldsymbol{\xi},\mathbf{R}_{x,h}\boldsymbol{\xi}\right\rangle-\mathbb{E}_{\mathbf{C}}\left[\left\langle\mathbf{R}_{x^{\star},h^{\star}}\boldsymbol{\xi},\mathbf{R}_{x,h}\boldsymbol{\xi}\right\rangle\right]\right|+\left|\mathbf{x}^{\star\mathsf{H}}\mathbf{x}\mathbf{h}^{\mathsf{H}}\mathbf{h}^{\star}\right|.\end{aligned}$$

Therefore, it suffices to show

$$\sup_{\mathbf{M}',\mathbf{M}\in\Delta}\left|\left\langle\mathbf{M}'\boldsymbol{\xi},\mathbf{M}\boldsymbol{\xi}\right\rangle-\mathbb{E}_{\mathbf{C}}\left\langle\mathbf{M}'\boldsymbol{\xi},\mathbf{M}\boldsymbol{\xi}\right\rangle\right|\leq\delta,\tag{31}$$

where

$$\Delta:=\left\{\mathbf{R}_{x,h}\mid\|\mathbf{x}\|=1,\|\mathbf{h}\|=1,\mathbf{h}\in\mathcal{C}_{\mu}\right\},$$

where $\mathcal{C}_\mu$ is the set of μ-spectrally flat $\mathbf{h}$. Since $\mathbf{I}\otimes\mathbf{h}$ is a unitary transform, $\boldsymbol{\xi}$ is a Gaussian vector satisfying $\mathbb{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{H}}\right]=\mathbf{I}_{MN}$. The desired concentration of the subgaussian bilinear form in (31) is then derived using Theorem 2.3 in [19]. To apply Theorem 2.3, we need upper bounds on $d_F\left(\Delta\right)$, $d_{2\to2}\left(\Delta\right)$, and $\gamma_2\left(\Delta,\|\cdot\|_{2\to2}\right)$.
The Frobenious norm of $\mathbf{R}_{x,h}\in\Delta$ is upper-bounded by

$$\left\|\mathbf{R}_{x,h}\right\|_F\leq\|\mathbf{x}\|\left\|\mathbf{B}\mathrm{Diag}\left(\mathbf{B}^{\mathsf{H}}\mathbf{h}\right)\right\|_F=\left\|\mathbf{B}\mathrm{Diag}\left(\mathbf{B}^{\mathsf{H}}\mathbf{h}\right)\right\|_F=\left\|\mathbf{B}^{\mathsf{H}}\mathbf{h}\right\|=1.$$

The spectral norm of $\mathbf{R}_{x,h}$ is upper-bounded by

$$\left\|\mathbf{R}_{x,h}\right\|_{2\to2}\leq\|\mathbf{x}\|\left\|\mathrm{Diag}\left(\mathbf{B}^{\mathsf{H}}\mathbf{h}\right)\right\|_{2\to2}=\left\|\mathrm{Diag}\left(\mathbf{B}^{\mathsf{H}}\mathbf{h}\right)\right\|_{2\to2}=\left\|\mathbf{B}^{\mathsf{H}}\mathbf{h}\right\|_{\infty}\leq\frac{\mu}{\sqrt{M}},$$

where the last step follows from $\mathbf{h}\in\mathcal{C}_{\mu_1}$. Therefore, we deduce

$$d_F\left(\Delta\right)\leq1,\quad d_{2\to2}\left(\Delta\right)\leq\frac{\mu}{\sqrt{M}}.$$

Next, by Dudley-type integral [39], $\gamma_2\left(\Delta, \|\cdot\|_{2\to 2}\right)$ is upper-bounded by

$$\gamma_2\left(\Delta, \|\cdot\|_{2\to 2}\right) \lesssim \int_0^\infty \sqrt{\log N\left(\Delta, \epsilon B_{S_\infty}\right)} d\epsilon,$$

where B_{S_∞} denotes the unit ball in the Schatten class S_∞ with $\|\cdot\|_{2\to 2}$ and the covering number $N\left(\Delta, \epsilon B_{S_\infty}\right)$ is given by

$$N\left(\Delta, \epsilon B_{S_\infty}\right) = \inf\left\{k \in N \mid \exists \left(u_i\right)_{i=1}^k, \text{ s.t. } \Delta \subset \cup_i u_i + \epsilon B_{S_\infty}\right\}.$$

We have shown that $d_{2\to 2}\left(\Delta\right) \leq \frac{\mu}{\sqrt{M}}$, which implies that $N\left(\Delta, \epsilon B_{S_\infty}\right) = 1, \forall \epsilon \geq \frac{\mu}{\sqrt{M}}$. Therefore, the integral reduces to

$$\int_0^\infty \sqrt{\log N\left(\Delta, \epsilon B_{S_\infty}\right)} d\epsilon = \int_0^{\frac{\mu}{\sqrt{M}}} \sqrt{\log N\left(\Delta, \epsilon B_{S_\infty}\right)} d\epsilon.$$

To compute the above integral, we first compute an estimate for the difference. For $\mathbf{R}_{x,h}, \mathbf{R}_{x',h'} \in \Delta$, we have

$$\begin{aligned}
\left\|\mathbf{R}_{x,h} - \mathbf{R}_{x',h'}\right\|_{2\to 2} &\leq \left\|\mathbf{R}_{x-x',h'} + \mathbf{R}_{x,h-h'}\right\|_{2\to 2} \\
&\leq \left\|\mathbf{R}_{x-x',h'}\right\|_{2\to 2} + \left\|\mathbf{R}_{x,h-h'}\right\|_{2\to 2} \\
&\leq \left\|\mathbf{x} - \mathbf{x}'\right\|_2 \left\|\mathbf{B}^{\mathsf{H}}\mathbf{h}\right\|_\infty + \left\|\mathbf{B}^{\mathsf{H}}\left(\mathbf{h} - \mathbf{h}'\right)\right\|_\infty \\
&\leq \frac{\mu}{\sqrt{M}} \left\|\mathbf{x} - \mathbf{x}'\right\|_2 + \left\|\mathbf{B}^{\mathsf{H}}\left(\mathbf{h} - \mathbf{h}'\right)\right\|_\infty.
\end{aligned}$$

Therefore, we have

$$N\left(\Delta, \epsilon B_{S_\infty}\right) \leq N_x\left(\Gamma_{s_x} \cap B_2^N, \frac{\epsilon}{2}\frac{\sqrt{M}}{\mu} B_2^N\right) N_h\left(\mathbf{B}^{\mathsf{H}}\left(\Gamma_{s_h} \cap B_2^M \cap C_\mu\right), \frac{\epsilon}{2} B_\infty^M\right),$$

where the covering numbers N_x and N_h are defined in ℓ_2^N and ℓ_∞^M. Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we deduce with a change of variable that

$$\begin{aligned}
\int_0^{\frac{\mu}{\sqrt{M}}} \sqrt{\log N\left(\Delta, \epsilon B_{S_\infty}\right)} d\epsilon \leq & \int_0^{\frac{\mu}{\sqrt{M}}} \sqrt{\log N_x\left(\Gamma_{s_x} \cap B_2^N, \frac{\epsilon}{2}\frac{\sqrt{M}}{\mu} B_2^N\right)} d\epsilon \\
& + \int_0^{\frac{\mu}{\sqrt{M}}} \sqrt{\log N_h\left(\mathbf{B}^{\mathsf{H}}\left(\Gamma_{s_h} \cap B_2^M \cap C_\mu\right), \frac{\epsilon}{2} B_\infty^M\right)} d\epsilon \\
\leq & \underbrace{2\frac{\mu}{\sqrt{M}} \int_0^{\frac{1}{2}} \sqrt{\log N_x\left(\Gamma_{s_x} \cap B_2^N, \epsilon B_2^N\right)} d\epsilon}_{(*)} \\
& + \underbrace{2\int_0^{\frac{\mu}{2\sqrt{M}}} \sqrt{\log N_h\left(\mathbf{B}^{\mathsf{H}}\left(\Gamma_{s_h} \cap B_2^M\right), \epsilon B_\infty^M\right)} d\epsilon}_{(**)}.
\end{aligned}$$

By the volumetric argument $N_x\left(\Gamma_{s_x} \cap B_2^N, \epsilon B_2^N\right) \leq \binom{N}{s_x} \left(1 + \frac{2}{\epsilon}\right)^{s_x} \leq \left(\frac{eN}{s_x}\right)^{s_x} \left(1 + \frac{2}{\epsilon}\right)^{s_x}$,

$$\begin{aligned}
(*) &\leq 2\frac{\mu}{\sqrt{M}} \int_0^{\frac{1}{2}} \sqrt{s_x \log\left(\frac{eN}{s_x}\right) + s_x \log\left(1 + \frac{2}{\epsilon}\right)} d\epsilon \\
&= 2\mu\sqrt{\frac{s_x}{M}} \left(\sqrt{\log\left(\frac{eN}{s_x}\right)} + \int_0^{\frac{1}{2}} \sqrt{\log\left(1 + \frac{2}{\epsilon}\right)} d\epsilon\right) \lesssim \mu\sqrt{\frac{s_x}{M} \log\left(\frac{eN}{s_x}\right)}.
\end{aligned}$$

By (14) in [19] and $\|\mathbf{B}\|_{1\to\infty} = \frac{1}{\sqrt{M}}$,

$$(**) \lesssim \|\mathbf{B}\|_{1\to\infty} \sqrt{\frac{s_h}{M}} \log s_h \log M \lesssim \frac{s_h}{M} \log^2 M.$$

So we have

$$\gamma_2\left(\Delta, \|\cdot\|_{2\to 2}\right) \lesssim \mu\sqrt{\frac{s_x}{M}\log\left(\frac{eN}{s_x}\right)} + \frac{\sqrt{s_h}}{M}\log^2 M \lesssim \delta.$$

Combining upper bounds on $K_{1,2,3}$ in Theorem 2.3 in [19],

$$K_1 := \gamma_2\left(\Delta, \|\cdot\|_{2\to 2}\right)\left[\gamma_2\left(\Delta, \|\cdot\|_{2\to 2}\right) + d_F\left(\Delta\right)\right] + d_F\left(\Delta\right) d_{2\to 2}\left(\Delta\right) \leq \delta^2 + \delta + \frac{\mu}{\sqrt{M}},$$

$$K_2 := d_{2\to 2}\left(\Delta\right)\left[\gamma_2\left(\Delta, \|\cdot\|_{2\to 2}\right) + d_F\left(\Delta\right)\right] \leq \frac{\mu}{\sqrt{M}}\left[\delta + 1\right],$$

$$K_3 := d^2_{2\to 2}\left(\Delta\right) \leq \frac{\mu^2}{M},$$

we note that there exist $C > 0$, $0 < c_1 < \frac{1}{4}$ so that $M \geq \max\left\{\mu^2 s_x \log\left(\frac{N}{s_x}\right)/\delta^2, s_h \log^2 M/\delta\right\}$ implies (31) has upper bound $c_1 K_1 + t$ with probability less than $8\exp\left(-c_2\frac{t^2}{K_2^2}\right) \lesssim \exp\left(-c_2\frac{M}{\mu}\right) \lesssim M^{-10}$. □

F More Image Deblurring Experiments

F.1 Structured Dictionaries and Comparison with Convex Lifting

We provide additional image deblurring experiments to demonstrate the performance of ThWF with structured dictionaries. These experiments follow the framework established in Section 5.2. We use 160×160 ground-truth images and evaluate performance under three blurring kernels (BG, motion, and Gaussian). The images are represented such that black pixels correspond to 0 and white pixels to 1. In all cases, we assume the kernel support is known, while the sparse support of the image coefficients $\mathbf{x}$ is unknown. We fix the sparsity level for ThWF at $s_x = 0.2M$ and use the relative objective change of no more than 10^{-7} as the stopping criterion.

Haar wavelet matrix case: We first examine two distinct images that exhibit sparsity in the Haar wavelet domain: the Sydney Opera House[2] (Figure 3(a)) and a protein structure[3] (Figure 4(a)). For these experiments, the dictionary $\mathbf{C}$ is set as the Haar matrix. For the Sydney Opera House image, the step sizes η are $\{0.1, 0.06, 0.05\}$ for the respective kernels. For the protein image, we set η to be $\{0.1, 0.05, 0.05\}$. As illustrated in Figures 3 and 4, ThWF consistently and significantly outperforms the convex lifting baseline. These results across both natural and scientific images confirm that ThWF effectively handles structured sparsity in transform domains, achieving superior kernel estimation and reconstruction quality. The quantitative comparison in terms of PSNR is summarized in Table 3(b) and Table 4(b), respectively.

DCT matrix case: We further evaluate the algorithm using a desert image[4] (Figure 5(a)). To test a different structural prior, we employ the discrete cosine transform (DCT) as the dictionary $\mathbf{C}$. For this experiment, we set η to be $\{0.1, 0.05, 0.05\}$. The results, summarized in Figure 5, reinforce our previous findings: ThWF maintains robust performance across different transform bases, effectively recovering sharp edges and fine textures where convex methods often struggle with blur residual. The PSNR comparison is in Table 5(b).

[2]Source: `https://commons.wikimedia.org/wiki/File:Sydney_Opera_House_._(8114385650).jpg`

[3]Source: `https://commons.wikimedia.org/wiki/File:PBB_Protein_RET_image.jpg`

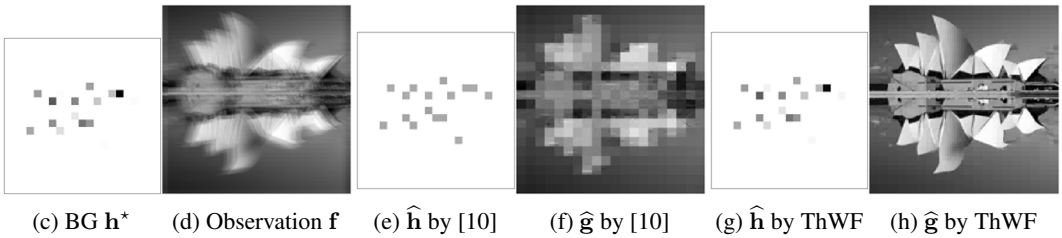
[4]Source: `https://commons.wikimedia.org/wiki/File:2025-04-30_Great_Sand_Dunes_National_Park_Frost_and_Ice_-_081.jpg`



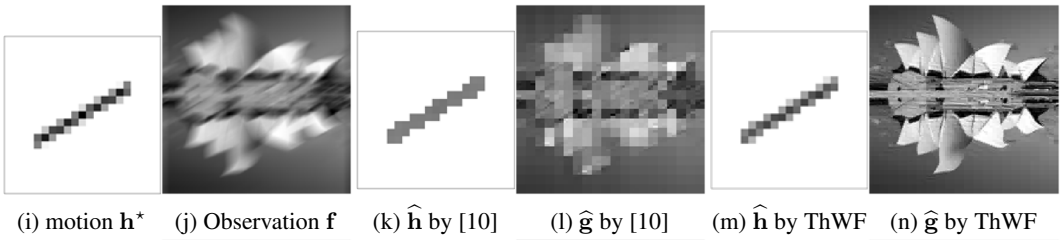
(a) Ground truth $\mathbf{g}^\star$

Method	BG kernel	Motion kernel	Gaussian kernel
CVX [10]	17.61	16.25	20.11
ThWF	**37.82**	**25.48**	**30.08**

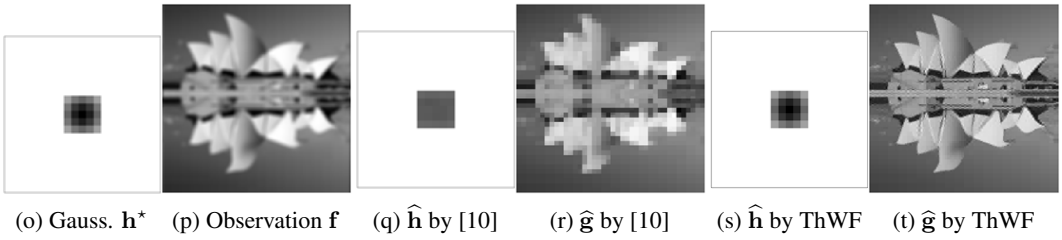
(b) PSNR (dB) comparison.



(c) BG $\mathbf{h}^\star$ (d) Observation $\mathbf{f}$ (e) $\widehat{\mathbf{h}}$ by [10] (f) $\widehat{\mathbf{g}}$ by [10] (g) $\widehat{\mathbf{h}}$ by ThWF (h) $\widehat{\mathbf{g}}$ by ThWF

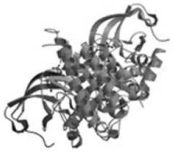


(i) motion $\mathbf{h}^\star$ (j) Observation $\mathbf{f}$ (k) $\widehat{\mathbf{h}}$ by [10] (l) $\widehat{\mathbf{g}}$ by [10] (m) $\widehat{\mathbf{h}}$ by ThWF (n) $\widehat{\mathbf{g}}$ by ThWF



(o) Gauss. $\mathbf{h}^\star$ (p) Observation $\mathbf{f}$ (q) $\widehat{\mathbf{h}}$ by [10] (r) $\widehat{\mathbf{g}}$ by [10] (s) $\widehat{\mathbf{h}}$ by ThWF (t) $\widehat{\mathbf{g}}$ by ThWF

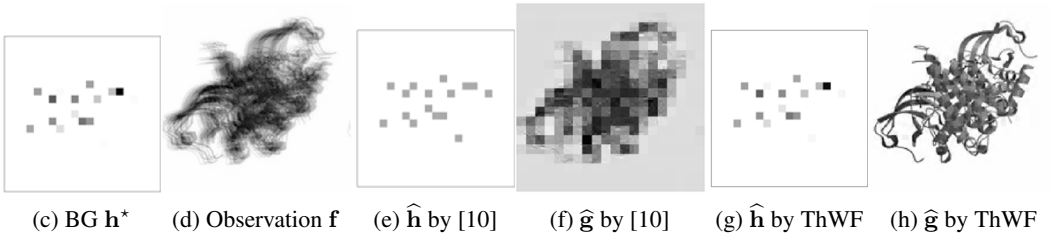
Figure 3: Sparse blind deconvolution results for the Sydney Opera House image.



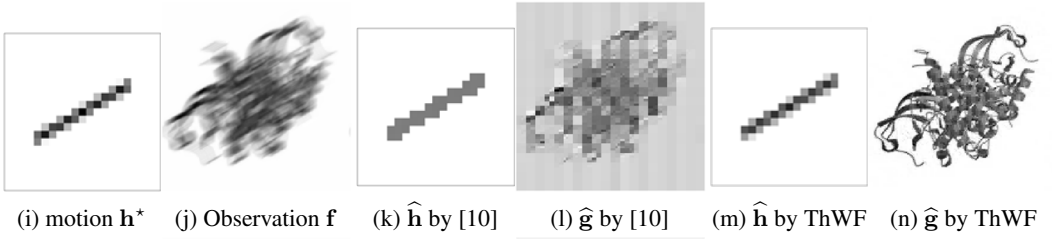
(a) Ground truth $\mathbf{g}^\star$

Method	BG kernel	Motion kernel	Gaussian kernel
CVX [10]	15.94	14.60	18.16
ThWF	**33.12**	**28.71**	**27.43**

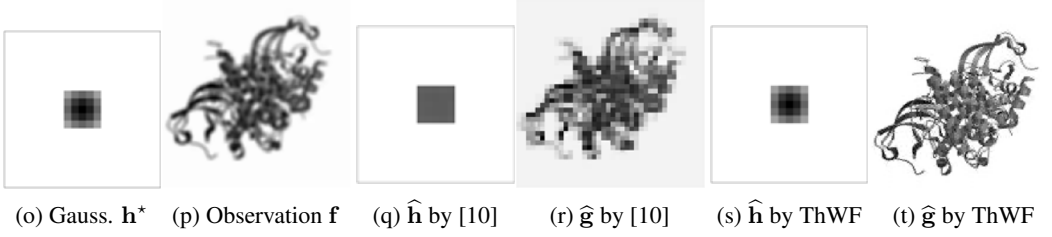
(b) PSNR (dB) comparison.



(c) BG $\mathbf{h}^\star$ (d) Observation $\mathbf{f}$ (e) $\widehat{\mathbf{h}}$ by [10] (f) $\widehat{\mathbf{g}}$ by [10] (g) $\widehat{\mathbf{h}}$ by ThWF (h) $\widehat{\mathbf{g}}$ by ThWF

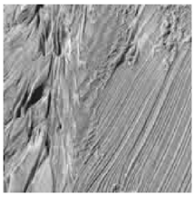


(i) motion $\mathbf{h}^\star$ (j) Observation $\mathbf{f}$ (k) $\widehat{\mathbf{h}}$ by [10] (l) $\widehat{\mathbf{g}}$ by [10] (m) $\widehat{\mathbf{h}}$ by ThWF (n) $\widehat{\mathbf{g}}$ by ThWF



(o) Gauss. $\mathbf{h}^\star$ (p) Observation $\mathbf{f}$ (q) $\widehat{\mathbf{h}}$ by [10] (r) $\widehat{\mathbf{g}}$ by [10] (s) $\widehat{\mathbf{h}}$ by ThWF (t) $\widehat{\mathbf{g}}$ by ThWF

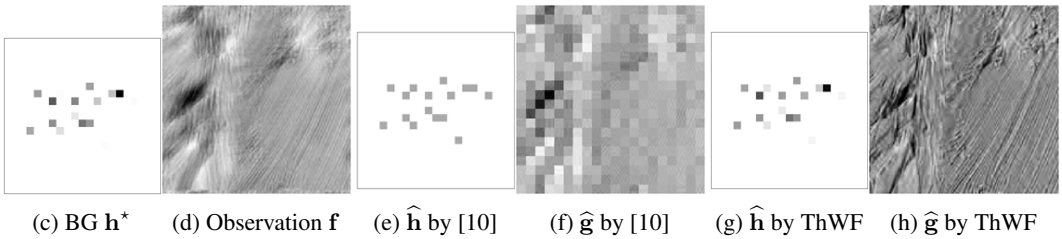
Figure 4: Sparse blind deconvolution results for the protein image.



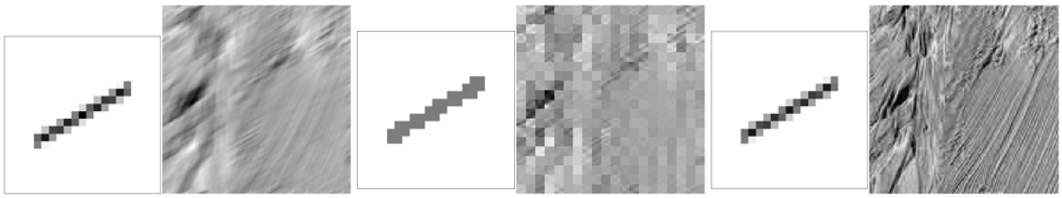
(a) Ground truth $\mathbf{g}^\star$

Method	BG kernel	Motion kernel	Gaussian kernel
CVX [10]	20.02	19.25	20.55
ThWF	**27.73**	**29.50**	**25.36**

(b) PSNR (dB) comparison.



(c) BG $\mathbf{h}^\star$ (d) Observation $\mathbf{f}$ (e) $\widehat{\mathbf{h}}$ by [10] (f) $\widehat{\mathbf{g}}$ by [10] (g) $\widehat{\mathbf{h}}$ by ThWF (h) $\widehat{\mathbf{g}}$ by ThWF



(i) motion $\mathbf{h}^\star$ (j) Observation $\mathbf{f}$ (k) $\widehat{\mathbf{h}}$ by [10] (l) $\widehat{\mathbf{g}}$ by [10] (m) $\widehat{\mathbf{h}}$ by ThWF (n) $\widehat{\mathbf{g}}$ by ThWF



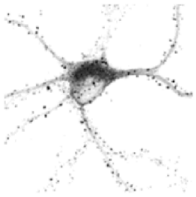
(o) Gauss. $\mathbf{h}^\star$ (p) Observation $\mathbf{f}$ (q) $\widehat{\mathbf{h}}$ by [10] (r) $\widehat{\mathbf{g}}$ by [10] (s) $\widehat{\mathbf{h}}$ by ThWF (t) $\widehat{\mathbf{g}}$ by ThWF

Figure 5: Sparse blind deconvolution results for the desert image.

F.2 Comparison with the Deep Learning Method SelfDeblur

We compare ThWF with SelfDeblur [35], a blind deblurring method based on deep image and kernel priors, on the neuron and Sydney Opera House images under the BG, motion, and Gaussian kernels. We report both the original SelfDeblur and a variant denoted by "SelfDeblur w/supp," which is provided with the true kernel support while still estimating all kernel amplitudes. ThWF also uses the true kernel support, as in the preceding experiments. The ThWF dictionaries and parameters are those used for the corresponding images in Section 5.2 and Appendix F.1.

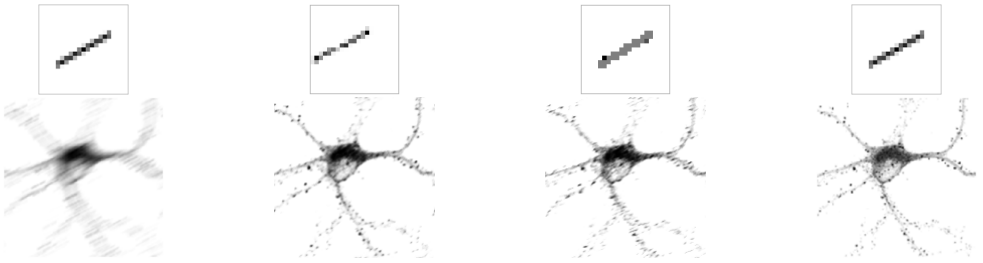
Figures 6 and 7 present the recovered images and kernels together with the image PSNR values. ThWF achieves higher PSNR than both SelfDeblur variants in all six tested cases. Providing the true kernel support improves SelfDeblur in five of the six cases, but a gap remains relative to ThWF. These comparisons illustrate the benefit of the explicit sparsity model for these particular images and blur kernels.



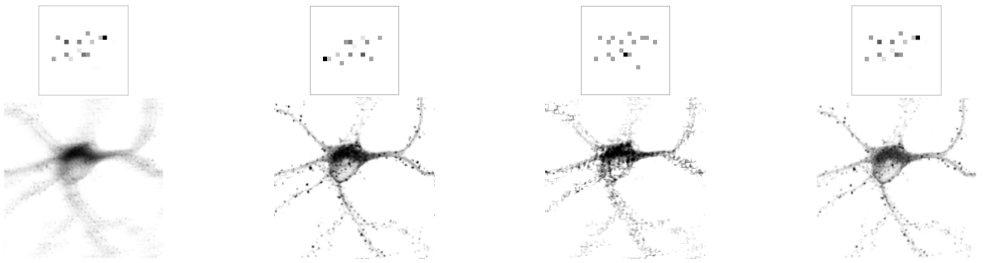
(a) Ground truth $\mathbf{g}^\star$

Method	BG	Motion	Gaussian
SelfDeblur	22.88	21.89	26.25
SelfDeblur w/supp	20.99	27.11	27.14
ThWF	**111.10**	**83.40**	**40.18**

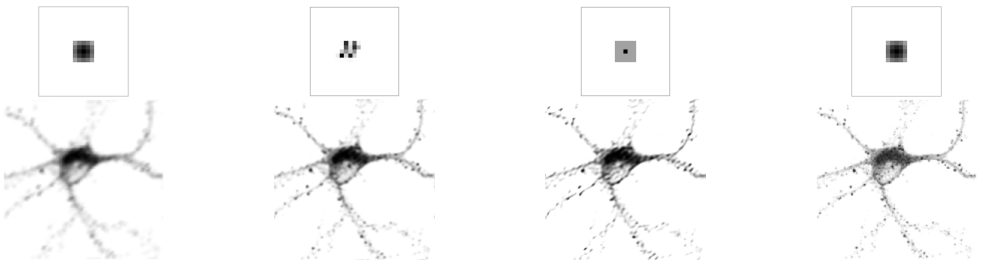
(b) PSNR (dB) comparison.



(c) Observation (d) SelfDeblur (e) SelfDeblur w/supp (f) ThWF



(g) Observation (h) SelfDeblur (i) SelfDeblur w/supp (j) ThWF



(k) Observation (l) SelfDeblur (m) SelfDeblur w/supp (n) ThWF

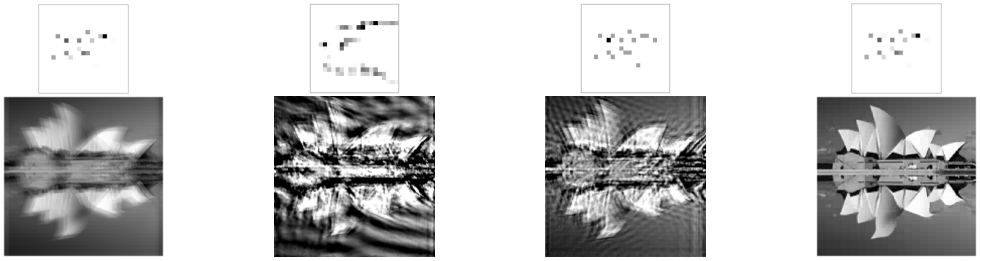
Figure 6: Comparison with SelfDeblur on the neuron image. For SelfDeblur w/supp, the true kernel support is provided while all kernel amplitudes remain estimated.



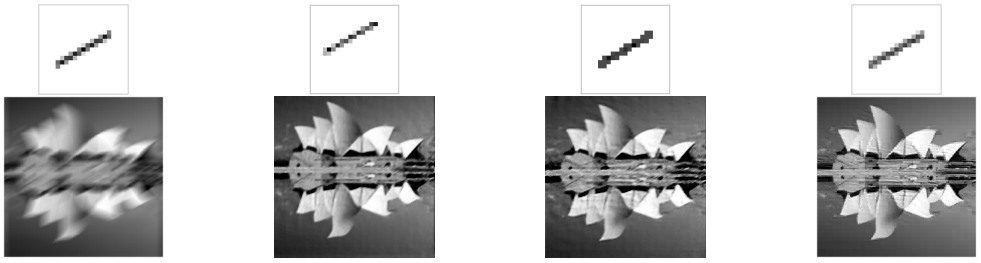
(a) Ground truth $\mathbf{g}^\star$

Method	BG	Motion	Gaussian
SelfDeblur	10.97	15.13	13.73
SelfDeblur w/supp	13.85	23.35	26.56
ThWF	**37.82**	**25.48**	**30.08**

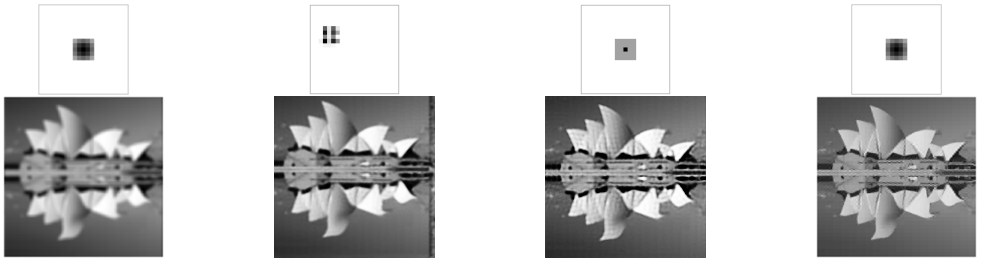
(b) PSNR (dB) comparison.



(c) Observation (d) SelfDeblur (e) SelfDeblur w/supp (f) ThWF



(g) Observation (h) SelfDeblur (i) SelfDeblur w/supp (j) ThWF



(k) Observation (l) SelfDeblur (m) SelfDeblur w/supp (n) ThWF

Figure 7: Comparison with SelfDeblur on the Sydney Opera House image. For SelfDeblur w/supp, the true kernel support is provided while the kernel amplitudes are still estimated.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline our contributions: a new algorithm (ThWF), a sparse spectral initialization scheme, and the first linear convergence guarantee for sparse blind deconvolution. These claims are supported by formal proofs in Theorems 1, 2, and 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Section 6 (Conclusions and Discussions), specifically regarding the high SNR requirement for the current initialization and the reliance on i.i.d. Gaussian assumptions for the dictionary $\mathbf{C}$.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions are explicitly stated in Section 4. Complete proofs for Theorems Theorems 1, 2, and 3 are provided in the Supplemental Material (Appendices B, C, and D respectively).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 5 detail the experimental setup, including problem dimensions (N, M), sparsity levels (s_h, s_x), and hyperparameters like step size (η) and noise levels (σ).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the source code in the supplemental material and have included an anonymous GitHub link in the paper. The repository includes scripts to reproduce the synthetic experiments (Figure 1) and image deblurring tasks (Figures 2 and 3).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameter choices (η) and stopping criteria are specified in Section 5.1 for synthetic data and Section 5.2 for image deblurring.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 1(a) includes shaded regions representing the standard deviation across 100 independent Monte Carlo trials. Figures 1(c) and 1(d) also report mean success rates and Frobenius errors over 100 trials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The hardware specifications (CPU, RAM, GPU) are detailed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research on mathematical optimization for signal processing does not involve human subjects or sensitive data and adheres to all ethical guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Our work provides a foundational algorithm for sparse blind deconvolution, which has potential positive impacts in medical imaging (e.g., clarifying neuron structures) and astronomical observations. While the algorithm is a general-purpose mathematical tool, it could theoretically be misused for unauthorized enhancement of surveillance images, although such applications would still require significant domain-specific adaptation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper focuses on foundational optimization algorithms and poses no high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

- **Software:** For convex baseline implementation, we used the CVX toolbox (Grant and Boyd, 2014), which is properly cited in our references and distributed under the GNU General Public License (GPL).
- **Data:** The neuron image was obtained from the Greenspine Gallery (Paul De Koninck Lab); we have credited the original owner and included the source URL in our references. Images shown in Appendix F were sourced from Wikimedia Commons and are in the public domain (CC0), ensuring no copyright restrictions for their use in this research.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We do not introduce new datasets or models beyond the provided algorithm code.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve any crowdsourcing or research conducted with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research is restricted to mathematical optimization and algorithmic analysis on synthetic and existing open-access image datasets; no human subjects were involved, thus IRB approval is not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were only used for writing and editing assistance and do not impact the core scientific methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.