Blind Deconvolution using Convex Programming

Ali Ahmed, Benjamin Recht, and Justin Romberg*

DRAFT: 11:48pm, November 18, 2012

Abstract

We consider the problem of recovering two unknown vectors, $\boldsymbol{w}$ and $\boldsymbol{x}$, of length L from their circular convolution. We make the structural assumption that the two vectors are members known subspaces, one with dimension N and the other with dimension K. Although the observed convolution is nonlinear in both $\boldsymbol{w}$ and $\boldsymbol{x}$, it is linear in the rank-1 matrix formed by their outer product $\boldsymbol{w}\boldsymbol{x}^*$. This observation allows us to recast the deconvolution problem as low-rank matrix recovery problem from linear measurements, whose natural convex relaxation is a nuclear norm minimization program.

We prove the effectiveness of this relaxation by showing that for "generic" signals, the program can deconvolve $\boldsymbol{w}$ and $\boldsymbol{x}$ exactly when the maximum of N and K is almost on the order of L. That is, we show that if $\boldsymbol{x}$ is drawn from a random subspace of dimension N, and $\boldsymbol{w}$ is a vector in a subspace of dimension K whose basis vectors are "spread out" in the frequency domain, then nuclear norm minimization recovers $\boldsymbol{w}\boldsymbol{x}^*$ without error.

We discuss this result in the context of blind channel estimation in communications. If we have a message of length N which we code using a random $L \times N$ coding matrix, and the encoded message travels through an unknown linear time-invariant channel of maximum length K, then the receiver can recover both the channel response and the message when $L \gtrsim N + K$, to within constant and log factors.

1 Introduction

This paper considers a fundamental problem in signal processing and communications: we observe the convolution of two unknown signals, $\boldsymbol{w}$ and $\boldsymbol{x}$, and want to separate them. We will show that this problem can be naturally relaxed as a semidefinite program (SDP), in particular, a nuclear norm minimization program. We then use this fact in conjunction with recent results on recovering low-rank matrices from underdetermined linear observations to provide conditions under which $\boldsymbol{w}$ and $\boldsymbol{x}$ can be deconvolved exactly. Qualitatively, these results say that if both $\boldsymbol{w}$ and $\boldsymbol{x}$ have length L, $\boldsymbol{w}$ lives in a fixed subspace of dimension K and is spread out in the frequency domain, and $\boldsymbol{x}$ lives in a "generic" subspace chosen at random, then $\boldsymbol{w}$ and $\boldsymbol{x}$ are separable with high probability.

*A. A. and J. R. are with the School of Electrical and Computer Engineering at Georgia Tech in Atlanta, GA. Email: {alikhan,jrom}@ece.gatech.edu. B. R. is with the Department of Computer Science at the University of Wisconsin. Email: brecht@cs.wisc.edu. A. A. and J. R. are supported by ONR grant N00014-11-1-0459 and a grant from the Packard Foundation. B. R. is generously supported by ONR award N00014-11-1-0723 and NSF awards CCF-1139953 and CCF-1148243. A.A. and J.R. would like to thank William Mantzel and M. Salman Asif for discussions related to this paper.

The general statement of the problem is as follows. We will assume that the length L signals live in known subspaces of $\mathbb{R}^L$ whose dimensions are K and N. That is, we can write

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{B}\boldsymbol{h}, \quad \boldsymbol{h} \in \mathbb{R}^K \\ \boldsymbol{x} &= \boldsymbol{C}\boldsymbol{m}, \quad \boldsymbol{m} \in \mathbb{R}^N \end{aligned}$$

for some $L \times K$ matrix $\boldsymbol{B}$ and $L \times N$ matrix $\boldsymbol{C}$. The columns of these matrices provide bases for the subspaces in which $\boldsymbol{w}$ and $\boldsymbol{x}$ live; recovering $\boldsymbol{h}$ and $\boldsymbol{m}$, then, is equivalent to recovering $\boldsymbol{w}$ and $\boldsymbol{x}$.

We observe the circular convolution of $\boldsymbol{w}$ and $\boldsymbol{x}$:

$$\boldsymbol{y} = \boldsymbol{w} * \boldsymbol{x}, \quad \text{or} \quad y[\ell] = \sum_{\ell'=1}^{L} w[\ell']x[\ell - \ell' + 1], \tag{1}$$

where the index $\ell - \ell' + 1$ in the sum above is understood to be modulo $\{1, \ldots, L\}$. It is clear that without structural assumptions on $\boldsymbol{w}$ and $\boldsymbol{x}$, there will not be a unique separation given the observations $\boldsymbol{y}$. But we will see that once we account for our knowledge that $\boldsymbol{w}$ and $\boldsymbol{x}$ lie in the span of the columns of $\boldsymbol{B}$ and $\boldsymbol{C}$, respectively, they can be uniquely separated in many situations. Detailing one such set of conditions under which this separation is unique and can be computed by solving a tractable convex program is the topic of this paper.

1.1 Matrix observations

We can break apart the convolution in (1) by expanding $\boldsymbol{x}$ as a linear combination of the columns $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_N$ of $\boldsymbol{C}$,

$$\begin{aligned} \boldsymbol{y} &= m(1)\boldsymbol{w} * \boldsymbol{C}_1 + m(2)\boldsymbol{w} * \boldsymbol{C}_2 + \cdots + m(N)\boldsymbol{w} * \boldsymbol{C}_N \\ &= \begin{bmatrix} \operatorname{circ}(\boldsymbol{C}_1) & \operatorname{circ}(\boldsymbol{C}_2) & \cdots & \operatorname{circ}(\boldsymbol{C}_N) \end{bmatrix} \begin{bmatrix} m(1)\boldsymbol{w} \\ m(2)\boldsymbol{w} \\ \vdots \\ m(N)\boldsymbol{w} \end{bmatrix}, \end{aligned}$$

where $\operatorname{circ}(\boldsymbol{C}_n)$ corresponds to the $L \times L$ circulant matrix whose action corresponds to circular convolution with the vector $\boldsymbol{C}_n$. Expanding $\boldsymbol{w}$ as a linear combination of the columns of $\boldsymbol{B}$, this becomes

$$\boldsymbol{y} = \begin{bmatrix} \operatorname{circ}(\boldsymbol{C}_1)\boldsymbol{B} & \operatorname{circ}(\boldsymbol{C}_2)\boldsymbol{B} & \cdots & \operatorname{circ}(\boldsymbol{C}_N)\boldsymbol{B} \end{bmatrix} \begin{bmatrix} m(1)\boldsymbol{h} \\ m(2)\boldsymbol{h} \\ \vdots \\ m(N)\boldsymbol{h} \end{bmatrix}. \tag{2}$$

We will find it convenient to write (2) in the Fourier domain. Let $\boldsymbol{F}$ be the L-point normalized discrete Fourier transform (DFT) matrix

$$\boldsymbol{F}(\omega, \ell) = \frac{1}{\sqrt{L}} \mathrm{e}^{-\mathrm{j}2\pi(\omega-1)(\ell-1)/L}, \qquad 1 \leq \omega, \ell \leq L.$$

We will use $\hat{\boldsymbol{C}} = \boldsymbol{F}\boldsymbol{C}$ for the $\boldsymbol{C}$-basis transformed into the Fourier domain, and also $\hat{\boldsymbol{B}} = \boldsymbol{F}\boldsymbol{B}$. Then $\text{circ}(\boldsymbol{C}_n) = \boldsymbol{F}^*\Delta_n\boldsymbol{F}$, where Δ_n is a diagonal matrix constructed from the nth column of $\hat{\boldsymbol{C}}$, $\Delta_n = \text{diag}(\sqrt{L}\hat{\boldsymbol{C}})$, and (2) becomes

$$\hat{\boldsymbol{y}} = \boldsymbol{F}\boldsymbol{y} = \begin{bmatrix}\Delta_1\hat{\boldsymbol{B}} & \Delta_2\hat{\boldsymbol{B}} & \cdots & \Delta_N\hat{\boldsymbol{B}}\end{bmatrix}\begin{bmatrix} m(1)\boldsymbol{h} \\ m(2)\boldsymbol{h} \\ \vdots \\ m(N)\boldsymbol{h}\end{bmatrix}. \tag{3}$$

Clearly, recovering $\hat{\boldsymbol{y}}$ is the same as recovering $\boldsymbol{y}$.

The expansions (2) and (3) make it clear that while $\boldsymbol{y}$ is a nonlinear combination of the coefficients $\boldsymbol{h}$ and $\boldsymbol{m}$, it is a *linear* combination of the entries of their outer product $\boldsymbol{X}_0 = \boldsymbol{h}\boldsymbol{m}^*$. We can pose the blind deconvolution problem as a linear inverse problem where we want to recover a $K \times N$ matrix from observations

$$\hat{\boldsymbol{y}} = \mathcal{A}(\boldsymbol{X}_0), \tag{4}$$

through a linear operator $\mathcal{A}$ which maps $K \times N$ matrices to $\mathbb{R}^L$. For $\mathcal{A}$ to be invertible over all matrices, we need at least as many observations as unknowns, $L \geq NK$. But since we know $\boldsymbol{X}_0$ has special structure, namely that its rank is 1, we will be able to recover it from $L \ll NK$ under certain conditions on $\mathcal{A}$.

As each entry of $\hat{\boldsymbol{y}}$ is a linear combination of the entries in $\boldsymbol{h}\boldsymbol{m}^*$, we can write them as trace inner products of different $K \times N$ matrices against $\boldsymbol{h}\boldsymbol{m}^*$. Using $\boldsymbol{b}_\ell \in \mathbb{C}^K$ for the ℓth column of $\hat{\boldsymbol{B}}^*$ and $\boldsymbol{c}_\ell \in \mathbb{C}^N$ as the ℓth row of $\sqrt{L}\hat{\boldsymbol{C}}$, we can translate one entry in (3) as[1]

$$\begin{aligned}\hat{y}(\ell) &= c_\ell(1)m(1)\langle \boldsymbol{h}, \boldsymbol{b}_\ell\rangle + c_\ell(2)m(2)\langle \boldsymbol{h}, \boldsymbol{b}_\ell\rangle + \cdots + c_\ell(N)m(N)\langle \boldsymbol{h}, \boldsymbol{b}_\ell\rangle \\ &= \langle \boldsymbol{c}_\ell, \boldsymbol{m}\rangle\langle \boldsymbol{h}, \boldsymbol{b}_\ell\rangle \\ &= \text{trace}(\boldsymbol{A}_\ell^*(\boldsymbol{h}\boldsymbol{m}^*)), \quad \text{where} \quad \boldsymbol{A}_\ell = \boldsymbol{b}_\ell\boldsymbol{c}_\ell^*. \end{aligned} \tag{5}$$

Now that we have seen that separating two signals given their convolution can be recast as a matrix recovery problem, we we turn our attention to a method for solving it. In the next section, we argue that a natural way to recover the expansion coefficients $\boldsymbol{m}$ and $\boldsymbol{h}$ from measurements of the form (3) is using nuclear norm minimization.

1.2 Convex relaxation

The previous section demonstrated how the blind deconvolution problem can be recast as a linear inverse problem over the (nonconvex) set of rank-1 matrices. A common heuristic to convexify the problem is to use the *nuclear norm*, the sum of the singular values of a matrix, as a proxy for rank [1]. In this section, we show how this heuristic provides a natural convex relaxation.

Given $\hat{\boldsymbol{y}} \in \mathbb{C}^L$, our goal is to find $\boldsymbol{h} \in \mathbb{R}^K$ and $\boldsymbol{m} \in \mathbb{R}^N$ that are consistent with the observations in (3). Making no assumptions about either of these vectors other than the dimension, the natural way to choose between multiple feasible points is using least-squares. We want to solve

$$\min_{\boldsymbol{u},\boldsymbol{v}} \ \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2 \quad \text{subject to} \quad \hat{\boldsymbol{y}}(\ell) = \langle \boldsymbol{c}_\ell, \boldsymbol{u}\rangle\langle \boldsymbol{h}, \boldsymbol{b}_\ell\rangle, \quad \ell = 1, \ldots, L. \tag{6}$$

[1] As we are now manipulating complex numbers in the frequency domain, we will need to take a little bit of care with definitions. Here and below, we use $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{v}^*\boldsymbol{u} = \text{trace}(\boldsymbol{u}\boldsymbol{v}^*)$ for complex vectors $\boldsymbol{u}$ and $\boldsymbol{v}$.

This is a non-convex quadratic optimization problem. The cost function is convex, but the quadratic equality constraints mean that the feasible set is non-convex. A standard approach to solving such quadratically constrained quadratic programs is to use duality (see for example [2]). A standard calculation shows that the dual of (6) is the semi-definite program (SDP)

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \operatorname{Re}\langle \hat{\boldsymbol{y}}, \boldsymbol{\lambda} \rangle \\ \text{subject to} \quad & \begin{bmatrix} \boldsymbol{I} & \sum_{\ell=1}^{L} \lambda(\ell) \boldsymbol{A}_\ell \\ \sum_{\ell=1}^{L} \lambda(\ell)^* \boldsymbol{A}_\ell^* & \boldsymbol{I} \end{bmatrix} \succeq 0, \end{aligned} \tag{7}$$

with the $\boldsymbol{A}_\ell = \boldsymbol{b}_\ell \boldsymbol{c}_\ell^*$ defined as in the previous section. Taking the dual again will give us a convex program which is in some sense as close to (6) as possible. The dual SDP of (7) is [3]

$$\begin{aligned} \min_{\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{X}} \quad & \tfrac{1}{2}\operatorname{trace}(\boldsymbol{W}_1) + \tfrac{1}{2}\operatorname{trace}(\boldsymbol{W}_2) \\ \text{subject to} \quad & \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{X} \\ \boldsymbol{X}^* & \boldsymbol{W}_2 \end{bmatrix} \succeq 0 \\ & \hat{\boldsymbol{y}} = \mathcal{A}(\boldsymbol{X}), \end{aligned} \tag{8}$$

which is completely equivalent to

$$\begin{array}{ll} \min & \|\boldsymbol{X}\|_* \\ \text{subject to} & \hat{\boldsymbol{y}} = \mathcal{A}(\boldsymbol{X}) \end{array}. \tag{9}$$

That is, the nuclear norm heuristic is the "dual-dual" relaxation of the intuitive but non-convex least-squares estimation problem (6).

Our technique for untangling $\boldsymbol{w}$ and $\boldsymbol{x}$ from their convolution, then, is to take the Fourier transform of the observation $\boldsymbol{y} = \boldsymbol{w} * \boldsymbol{x}$ and use it as constraints in the program (9). That (9) is the natural relaxation is fortunate, as an entire body of literature in the field of *low-rank recovery* has arisen in the past five years that is devoted to analyzing problems of the form (9). We will build on some of the techniques from this area in establishing the theoretical guarantees for when (9) is provably effective presented in the next section.

There have also been tremendous advances in algorithms for computing the solution to optimization problems of both types (6) and (9). In Section 2.1, we will briefly detail one such technique we used to solve (6) on a relatively large scale for a series of numerical experiments in Sections 2.2–2.4.

1.3 Main results

We can guarantee the effectiveness of (9) for relatively large subspace dimensions K and N when $\boldsymbol{B}$ is incoherent in the Fourier domain, and when $\boldsymbol{C}$ is generic. Before presenting our main analytical result, Theorem 1 below, we will carefully specify our models for $\boldsymbol{B}$ and $\boldsymbol{C}$, giving a concrete definition to the terms 'incoherent' and 'generic' in the process.

We will assume, without loss of generality, that the matrix $\boldsymbol{B}$ is an arbitrary $L \times K$ matrix with orthonormal columns:

$$\boldsymbol{B}^*\boldsymbol{B} = \hat{\boldsymbol{B}}^*\hat{\boldsymbol{B}} = \sum_{\ell=1}^{L} \boldsymbol{b}_\ell \boldsymbol{b}_\ell^* = \boldsymbol{I}, \tag{10}$$

where the $\boldsymbol{b}_\ell$ are the columns of $\hat{\boldsymbol{B}}^*$, as in (5). Our results will be most powerful when $\boldsymbol{B}$ is diffuse in the Fourier domain, meaning that the $\boldsymbol{b}_\ell$ all have similar norms. We will use the (in)coherence parameter μ_1 to quantify the degree to which the columns of $\boldsymbol{B}$ are jointly concentrated in the Fourier domain:

$$\mu_1^2 = \frac{L}{K} \max_{1\leq \ell \leq L} \|\boldsymbol{b}_\ell\|_2^2. \tag{11}$$

From (10), we know that the total energy in the rows of $\hat{\boldsymbol{B}}$ is $\sum_{\ell=1}^L \|\boldsymbol{b}_\ell\|_2^2 = K$, and that $\|\boldsymbol{b}_\ell\|_2^2 \leq 1$. Thus $1 \leq \mu_1^2 \leq L/K$, with the coherence taking its minimum value when the energy in $\hat{\boldsymbol{B}}$ is evenly distributed throughout its rows, and its maximum value when the energy is completely concentrated on K of the L rows. Our results will also depends on the minimum of these norms

$$\mu_L^2 = \frac{L}{K} \min_{1\leq \ell \leq L} \|\boldsymbol{b}_\ell\|_2^2. \tag{12}$$

We will always have $0 \leq \mu_L^2 \leq 1$ and $\mu_L^2 \leq \mu_1^2$. An example of a maximally incoherent $\boldsymbol{B}$, where $\mu_1^2 = \mu_L^2 = 1$, is

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{I}_K \\ \boldsymbol{0} \end{bmatrix}, \tag{13}$$

where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix. In this case, the range of $\boldsymbol{B}$ consists of "short" signals whose first K terms may be non-zero. The matrix $\hat{\boldsymbol{B}}$ is simply the first K columns of the discrete Fourier matrix, and so every entry has the same magnitude.

Our analytic results also depend on how diffuse the particular signal we are trying to recover $\boldsymbol{w} = \boldsymbol{B}\boldsymbol{h}$ is in the Fourier domain. With $\hat{\boldsymbol{w}} = \boldsymbol{F}\boldsymbol{w} = \hat{\boldsymbol{B}}\boldsymbol{h}$, we define

$$\mu_h^2 = L \max_{1\leq \ell \leq L} |\hat{w}(\ell)|^2 = L \cdot \max_{1\leq \ell \leq L} |\langle \boldsymbol{h}, \boldsymbol{b}_\ell\rangle|^2. \tag{14}$$

If the signal $\boldsymbol{w}$ is more or less "flat" in the frequency domain, then μ_h^2 will be a small constant. Note that it is always the case that $1 \leq \mu_h^2 \leq \mu_1^2 K$.

With the subspace in which $\boldsymbol{w}$ resides fixed, we will show that separating $\boldsymbol{w}$ and $\boldsymbol{x} = \boldsymbol{C}\boldsymbol{m}$ will be possible for "most" choices of the subspace $\boldsymbol{C}$ of a certain dimension N — we do this by choosing the subspace at random from an isotropic distribution, and show that (9) is successful with high probability. For the remainder of the paper, we will take the entries of $\boldsymbol{C}$ to be independent and identically distributed random variables,

$$C[\ell, n] \ \sim \ \text{Normal}(0, L^{-1}).$$

In the Fourier domain, the entries of $\hat{\boldsymbol{C}}$ will be complex Gaussian, and its columns will have conjugate symmetry (since the columns of $\boldsymbol{C}$ are real). Specifically, the rows of $\hat{\boldsymbol{C}}$ will be distributed as[2]

$$\begin{aligned} \boldsymbol{c}_\ell &\sim \begin{cases} \text{Normal}(0, \boldsymbol{I}) & \ell = 1 \\ \text{Normal}(0, 2^{-1/2}\boldsymbol{I}) + \text{j}\text{Normal}(0, 2^{-1/2}\boldsymbol{I}) & \ell = 2, \ldots, L/2+1 \end{cases}, \\ \boldsymbol{c}_\ell &= \boldsymbol{c}_{L-\ell+2}, \quad \text{for} \ \ \ell = L/2+2, \ldots, L. \end{aligned} \tag{15}$$

Similar results to those we present here most likely hold for other models for $\boldsymbol{C}$. The key property that our analysis hinges critically on is the rows $\boldsymbol{c}_\ell$ of $\hat{\boldsymbol{C}}$ are *independent* — this allows us to apply

[2]We are assuming here that L is even; the argument is straightforward to adapt to odd L.

recently developed tools for estimating the spectral norm of a sum of independent random linear operators.

We now state our main result:

Theorem 1. *Suppose the bases* $\boldsymbol{B}, \boldsymbol{C}$ *and expansion coefficients* $\boldsymbol{h}, \boldsymbol{m}$ *satisfy the conditions* (10), (11), (14), *and* (15) *above. Fix* $\alpha \geq 1$. *Then there exists a constant* $C_\alpha = O(\alpha)$ *depending only on* α, *such that if*

$$L \;\geq\; C_\alpha \,\max\left(\mu_1^2 K,\; \mu_h^2 N\right)\,\log^3(KN), \tag{16}$$

then $\boldsymbol{X}_0 = \boldsymbol{h}\boldsymbol{m}^*$ *is the unique solution to* (9) *with probability* $1 - O(L(NK)^{-\alpha})$.

When the coherences are low, meaning that μ_1 and μ_h are on the order of a constant, then (16) is coming within a logarithmic factor of the inherent number of degrees of freedom in the problem, as it takes $K + N$ variables to specify both $\boldsymbol{h}$ and $\boldsymbol{m}$.

While Theorem 1 establishes theoretical guarantees for specific types of subspaces specified by $\boldsymbol{B}$ and $\boldsymbol{C}$, we have found that treating blind deconvolution as a linear inverse problem with a rank constraint leads to surprisingly good results in many situations; see, for example, the image deblurring experiments in Section 2.4.

The recovery can also be made stable in the presence of noise, as described by our second theorem:

Theorem 2. *Let* $\boldsymbol{X}_0 = \boldsymbol{h}\boldsymbol{m}^*$ *and* $\mathcal{A}$ *as in* (4), *and suppose we observe*

$$\hat{\boldsymbol{y}} = \mathcal{A}(\boldsymbol{X}_0) + \boldsymbol{z},$$

where $\boldsymbol{z} \in \mathbb{R}^L$ *is an unknown noise vector with* $\|\boldsymbol{z}\|_2 \leq \delta$. *If* L *obeys* (16) *and also* $L \leq \mu_L^2 NK(2\sqrt{2}\beta\log(NK))^{-2}$ *for some* $\beta > 0$, *then with probability* $1 - O(L(NK)^{-\beta})$ *the solution* $\tilde{\boldsymbol{X}}$ *to*

$$\begin{array}{ll} \min & \|\boldsymbol{X}\|_* \\ \text{subject to} & \|\hat{\boldsymbol{y}} - \mathcal{A}(\boldsymbol{X})\|_2 \leq \delta \end{array} \tag{17}$$

will obey

$$\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_F \;\leq\; C\,\frac{\mu_1}{\mu_L}\sqrt{\min(K,N)}\,\delta,$$

for a fixed constant C.

The program in (17) is also convex, and is solved with numerical techniques similar to the equality constrained program in (9).

In the end, we are interested in how well we recover $\boldsymbol{x}$ and $\boldsymbol{w}$. The stability result for $\boldsymbol{X}_0$ can easily be extended to a guarantee for the two unknown vectors.

Corollary 1. *Let* $\tilde{\sigma}_1\tilde{\boldsymbol{u}}_1\tilde{\boldsymbol{v}}_1$ *be the best rank-1 approximation to* $\tilde{\boldsymbol{X}}$, *and set* $\tilde{\boldsymbol{h}} = \sqrt{\tilde{\sigma}_1}\tilde{\boldsymbol{u}}_1$ *and* $\tilde{\boldsymbol{m}} = \sqrt{\tilde{\sigma}_1}\tilde{\boldsymbol{v}}_1$. *Set* $\tilde{\delta} = \|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_F$. *Then there exists a constant* C *such that*

$$\|\boldsymbol{h} - \alpha\tilde{\boldsymbol{h}}\|_2 \;\leq\; C\min\left(\tilde{\delta}/\|\boldsymbol{h}\|_2, \|\boldsymbol{h}\|_2\right), \qquad \|\boldsymbol{m} - \alpha^{-1}\tilde{\boldsymbol{m}}\|_2 \;\leq\; C\min\left(\tilde{\delta}/\|\boldsymbol{m}\|_2, \|\boldsymbol{m}\|_2\right).$$

for some scalar multiple α.

Proof of this corollary follows the exact same line of reasoning as the later part of Theorem 1.2 in [4].

1.4 Relationship to phase retrieval and other quadratic problems

Blind deconvolution of $\boldsymbol{w} * \boldsymbol{x}$, as is apparent from (1), is equivalent to solving a system of quadratic equations in the entries of $\boldsymbol{w}$ and $\boldsymbol{x}$. The discussion in Section 1.1 shows how this system of quadratic equations can be recast as a linear set of equations with a rank constraint. In fact, this same recasting can be used for any system of quadratic equations in $\boldsymbol{w}$ and $\boldsymbol{x}$. The reason is simple: taking the outer product of the concatenation of $\boldsymbol{w}$ and $\boldsymbol{x}$ produces a rank-1 matrix that contains all the different combinations of entries of $\boldsymbol{w}$ multiplied with each other and multiplied by entries in $\boldsymbol{x}$:

$$\begin{bmatrix}\boldsymbol{w}\\ \boldsymbol{x}\end{bmatrix}\begin{bmatrix}\boldsymbol{w}^* & \boldsymbol{x}^*\end{bmatrix} = \left[\begin{array}{cccc|cccc} w[1]^2 & w[1]w[2] & \cdots & w[1]w[L] & w[1]x[1] & w[1]x[2] & \cdots & w[1]x[L] \\ w[2]w[1] & w[2]^2 & \cdots & w[2]w[L] & w[2]x[1] & w[2]x[2] & \cdots & w[2]x[L] \\ \vdots & & & & \vdots & & \vdots & \\ w[L]w[1] & w[L]w[2] & \cdots & w[L]^2 & w[L]x[1] & w[L]x[2] & \cdots & w[L]x[L] \\ \hline x[1]w[1] & x[1]w[2] & \cdots & x[1]w[L] & x[1]^2 & x[1]x[2] & \cdots & x[1]x[L] \\ x[2]w[1] & x[2]w[2] & \cdots & x[2]w[L] & x[2]x[1] & x[2]^2 & \cdots & x[2]x[L] \\ \vdots & & & & \vdots & & \vdots & \\ x[L]w[1] & x[L]w[2] & \cdots & x[L]w[L] & x[L]x[1] & x[L]x[2] & \cdots & x[L]^2 \end{array}\right]. \tag{18}$$

Then any quadratic equation can be written as a linear combination of the entries in this matrix, and any system of equations can be written as a linear operator acting on this matrix. For the particular problem of blind deconvolution, we are observing sums along the skew-diagonals of the matrix in the upper right-hand (or lower left-hand) quadrant. Incorporating the subspace constraints allows us to work with the smaller $K \times N$ matrix $\boldsymbol{h}\boldsymbol{m}^*$, but this could also be interpreted as adding additional linear constraints on the matrix in (18).

Recent work on *phase retrieval* [4] has used this same methodology of "lifting" a quadratic problem into a linear problem with a rank constraint to show that a vector $\boldsymbol{w} \in \mathbb{R}^N$ can be recovered from $O(N \log N)$ measurements of the form $|\langle \boldsymbol{w}, \boldsymbol{a}_n\rangle|^2$ for $\boldsymbol{a}_n$ selected uniformly at random from the unit sphere. In this case, the measurements are being made entirely in the upper left-hand (or lower-right hand) quadrant in (18), and the measurements in (5) have the form $\boldsymbol{A}_n = \boldsymbol{a}_n\boldsymbol{a}_n^*$. In fact, another way to interpret the results in [4] is that if a signal of length L is known to live in a generic subspace of dimension $\sim L/\log L$, then it can be recovered from an observation of a convolution with itself.

In the current work, we are considering a non-symmetric rank-1 matrix being measured by matrices $\boldsymbol{b}_\ell\boldsymbol{c}_\ell^*$ formed by the outer product of two different vectors, one of which is random, and one of which is fixed. Another way to cast the problem, which perhaps brings these differences into sharper relief, is that we are measuring the symmetric matrix in (18) by taking inner products against rank-two matrices $\frac{1}{2}\left(\begin{bmatrix}\boldsymbol{b}_\ell\\ \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{c}_\ell^*\end{bmatrix} + \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{c}_\ell\end{bmatrix}\begin{bmatrix}\boldsymbol{b}_\ell^* & \boldsymbol{0}\end{bmatrix}\right)$. These seemingly subtle differences lead to a much different mathematical treatment.

1.5 Application: Multipath channel protection using random codes

The results in Section 1.3 have a direct application in the context of channel coding for transmitting a message over an unknown multipath channel. The problem is illustrated in Figure 1. A message vector $\boldsymbol{m} \in \mathbb{R}^N$ is encoded through an $L \times N$ encoding matrix $\boldsymbol{C}$. The protected message $\boldsymbol{x} = \boldsymbol{C}\boldsymbol{m}$ travels through a channel whose impulse response is $\boldsymbol{w}$. The receiver observes $\boldsymbol{y} = \boldsymbol{w} * \boldsymbol{x}$, and from this would like to jointly estimate the channel and determine the message that was sent.

In this case, a reasonable model for the channel response $\boldsymbol{w}$ is that it is nonzero in relatively small number of known locations. Each of these entries corresponds to a different path over which the encoded message traveled; we are assuming that we know the timing delays for each of these paths, but not the fading coefficients. The matrix $\boldsymbol{B}$ in this case is a subset of columns from the identity, and the $\boldsymbol{b}_\ell$ are partial Fourier vectors. This means that the coherence μ_1 in (11) takes its minimal value of $\mu_1^2 = 1$, and the coherence μ_h^2 in (14) has a direct interpretation as the peak-value of the (normalized) frequency response of the unknown channel. The resulting linear operator $\mathcal{A}$ corresponds to a matrix comprised of N $L \times K$ random Toeplitz matrices, as shown in Figure 2. The first column of each of these matrices corresponds to a columns of $\boldsymbol{C}$. The formulation of this problem as a low-rank matrix recovery program was proposed in [5], which presented some first numerical experiments.

In this context, Theorem 1 tell us that a length N message can be protected against a channel with K *reflections* that is relatively flat in the frequency domain with a random code of length $L \gtrsim (K + N)\log^3(KN)$. Essentially, we have a theoretical guarantee that we can estimate the channel without knowledge of the message from a single transmitted codeword.

It is instructive to draw a comparison in to previous work which connected error correction to structured solutions to underdetermined systems of equations. In [6,7], it was shown that a message of length N could be protected against corruption in K unknown locations with a code of length $L \gtrsim N + K\log(N/K)$ using a random codebook. This result was established by showing how the decoding problem can be recast as a sparse estimation problem to which results from the field of compressed sensing can be applied.

For multipath protection, we have a very different type of corruption: rather than individual entries of the transmitted vector being tampered with, instead we observe overlapping copies of the transmission. We show that with the same type of codebook (i.e. entries chosen independently at random) can protect against K reflections during transmission, where the timing of these bounces is known (or can be reasonably estimated) but the fading coefficients (amplitude and phase change associated with each reflection) are not.

1.6 Other related work

As it is a ubiquitous problem, many different approaches for blind deconvolution have been proposed in the past, each using different statistical or deterministic models tailored to particular applications. A general overview for blind deconvolution techniques in imaging (including methods based on parametric modeling of the inputs and incorporating spatial constraints) can be found in [8]. An example of a more modern method can be found in [9], where it is demonstrated how an image, which is expected to have small total-variation with respect to its energy, can be effectively deconvolved from an unknown kernel with known compact support. In wireless communications, knowledge of the modulation scheme [10] or an estimate of the statistics of the source signal [11] have been used for blind channel identification; these methods are overviewed in the review papers [12–15]. An effective scheme based on a deterministic model was put forth in [16], where fundamental conditions for being able to identify multichannel responses from cross-correlations

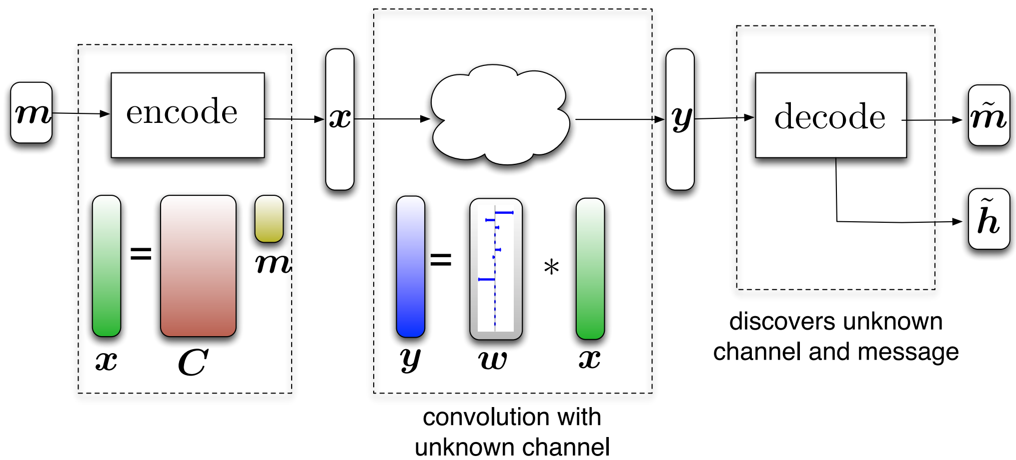


Figure 1: *Overview of the channel protection problem. A message $\boldsymbol{m}$ is encoded by applying a tall matrix $\boldsymbol{C}$; the receiver observes the encoded message convolved with an unknown channel response $\boldsymbol{w} = \boldsymbol{B}\boldsymbol{h}$, where $\boldsymbol{B}$ is a subset of columns from the identity matrix. The decoder is faced with the task of separating the message and channel response from this convolution, which is a nonlinear combination of $\boldsymbol{h}$ and $\boldsymbol{m}$.*

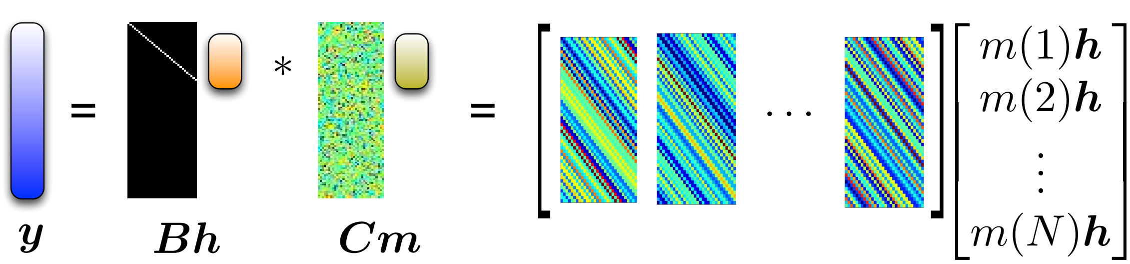


Figure 2: *The multi-toeplitz matrix corresponding to the multipath channel protection problem in Section 1.5. In this case, the columns of $\boldsymbol{B}$ are sampled from the identity, the entries of $\boldsymbol{C}$ are chosen to be iid Gaussian random variables, and the corresponding linear operator $\mathcal{A}$ is formed by concatenating N $L \times K$ random Toeplitz matrices, each of which is generated by a column of $\boldsymbol{C}$.*

are presented. The work in this paper differs from this previous work in that it relies only on a single observation of two convolved signals, the model for these signals is that they lie in known (but arbitrary) subspaces rather than have a prescribed length, and we give a concrete relationship between the dimensions of these subspaces and the length of the observation sufficient for perfect recovery.

Recasting the quadratic problem in (1) as the linear problem with a rank constraint in (5) is appealing since it puts the problem in a form for which we have recently acquired a tremendous amount of understanding. Recovering a $N \times K$ rank-R matrix from a set of linear observations has primarily been considered in two scenarios. In the case where the observations come through a random projection, where either the $\boldsymbol{A}_\ell$ are filled with independent Gaussian random variables or $\mathcal{A}$ is an orthoprojection onto a randomly chosen subspace, the nuclear norm minimization program in (9) is successful with high probability when [3, 17]

$$L \;\geq\; \text{Const} \cdot R \max(K, N).$$

When the observations are randomly chosen entries in the matrix, then subject to incoherence conditions on the singular vectors of the matrix being measured, the number of samples sufficient for recovery, again with high probability, is [18–21]

$$L \;\geq\; \text{Const} \cdot R \max(K, N) \log^2(\max(K, N)).$$

Our main result in Theorem 1 uses a completely different kind measurement system which exhibits a type of *structured randomness*; for example, when $\boldsymbol{B}$ has the form (13), $\mathcal{A}$ has the concatenated Toeplitz structure shown in Figure 2. In this paper, we will only be concerned with how well this type of operator can recover rank-1 matrices, ongoing work has shown that it also effectively recover general low-rank matrices [22].

While this paper is only concerned with recovery by nuclear norm minimization, other types of recovery techniques have proven effective both in theory and in practice; see for example [23–25]. It is possible that the guarantees given in this paper could be extended to these other algorithms.

As we will see below, our mathematical analysis has mostly to do how matrices of the form in (2) act on rank-2 matrices in a certain subspace. Matrices of this type have been considered in the context of sparse recovery in the compressed sensing literature for applications including multiple-input multiple-output channel estimation [26], multi-user detection [27], and multiplexing of spectrally sparse signals [28].

2 Numerical Simulations

In this section, we illustrate the effectiveness of the reconstruction algorithm for the blind deconvolution of vectors $\boldsymbol{x}$ and $\boldsymbol{w}$ with numerical experiments[3]. In particular, we study phase diagrams, which demonstrate the empirical probability of success over a range of dimensions N and K for a fixed L; an image deblurring experiment, where the task is to recover an image blurred by an unknown blur kernel; a channel protection experiment, where we show the robustness of our algorithm in the presence of additive noise.

[3] MATLAB code that reproduces all of the experiments in this section is available at `http://www.aliahmed.org`.

Some of the numerical experiments presented below are "large scale", with thousands (and even 10s of thousands) of unknown variables. Recent advances in SDP solvers, which we discuss in the following subsection, make the solution of such problems computationally feasible.

2.1 Large-scale solvers

To solve the semidefinite program (8) on instances where K and M are of practical size, we rely on the heuristic solver developed by Burer and Monteiro [29]. To implement this solver, we perform the variable substitution

$$\begin{bmatrix} \boldsymbol{H} \\ \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{H} \\ \boldsymbol{M} \end{bmatrix}^* = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{X} \\ \boldsymbol{X}^* & \boldsymbol{W}_2 \end{bmatrix}$$

where $\boldsymbol{H}$ is $K \times r$ and $\boldsymbol{M}$ is $N \times r$ for $r > 1$. Under this substitution, the semidefinite constraint is always satisfied and we are left with the nonlinear program:

$$\min_{\boldsymbol{M},\boldsymbol{H}} \ \|\boldsymbol{M}\|_F^2 + \|\boldsymbol{H}\|_F^2 \quad \text{subject to} \quad \hat{\boldsymbol{y}} = \mathcal{A}(\boldsymbol{H}\boldsymbol{M}^*), \quad \ell = 1, \ldots, L. \tag{19}$$

When $r = 1$, this reformulated problem is equivalent to (6). Burer and Monteiro showed that provided r is bigger than the rank of the optimal solution of (8), all of the local minima of (19) were global minima of (8) [30]. Since we expect a rank one solution, we can work with $r = 2$, declaring recovery when a rank deficient $\boldsymbol{M}$ or $\boldsymbol{H}$ is obtained. Thus, by doubling the size of the decision variable, we can avoid the non-global local solutions of (6). Burer and Monteiro's algorithm has had notable success in matrix completion problems, enabling some of the fastest solvers for nuclear-norm-based matrix completion [31, 32].

To solve (19), we implement the method of multipliers strategy initially suggested by Burer and Monteiro. Indeed, this algorithm is explained in detail by Recht *et al* in the context of solving problem (9) [3]. The inner operation of minimizing the augmented Lagrangian term is performed using LBFGS as implemented by the Matlab solver minfunc [33]. This solver requires only being able to apply $\mathcal{A}$ and $\mathcal{A}^*$ quickly, both of which can be done in time $O(r \min\{N \log N, K \log K\})$. The parameters of the augmented Lagrangian are updated according to the schedule proposed by Burer and Monteiro [29]. This code allows us to solver problems where N and K are in the tens of thousands in seconds on a laptop.

2.2 Phase transitions

Our first set of numerical experiments delineates the boundary, in terms of values for K, N and L, for when (9) is effective on generic instances of four different types of problems. For a fixed value of L, we vary the subspace dimensions N and K and run 100 experiments, with different random instances of $\boldsymbol{w}$ and $\boldsymbol{x}$ for each experiment. Figures 3 and 4 show the collected frequencies of success for four different probabilistic models. We classify a recovery a success if its relative error is less than 2%[4], meaning that if $\hat{\boldsymbol{X}}$ is the solution to (9), then

$$\frac{\|\hat{\boldsymbol{X}} - \boldsymbol{w}\boldsymbol{x}^*\|_F}{\|\boldsymbol{w}\boldsymbol{x}^*\|_F} < 0.02. \tag{20}$$

[4] The diagrams in Figures 3 and 4 do not change significantly if a smaller threshold, say on the order of 10^{-6}, is chosen.

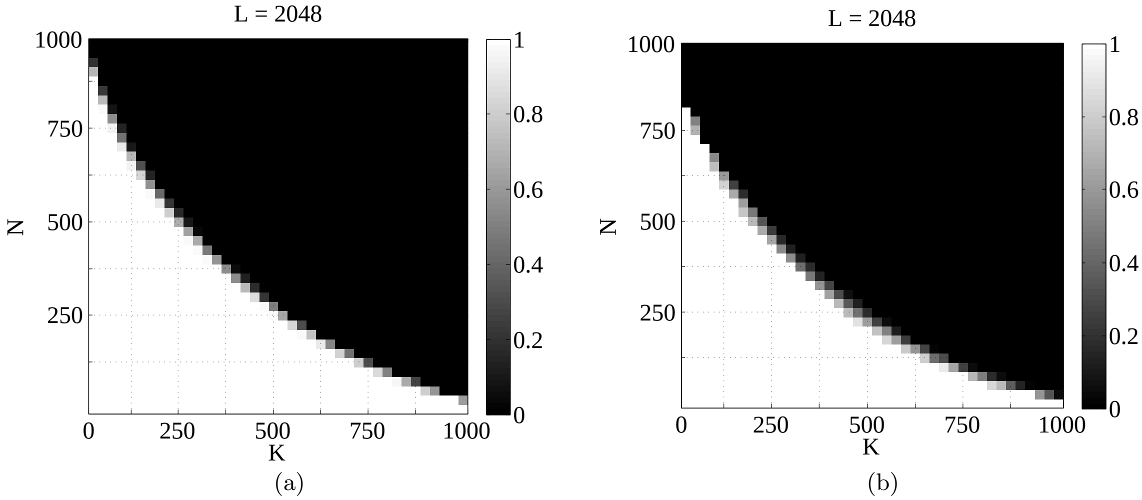


Figure 3: *Empirical success rate for the deconvolution of two vectors $\boldsymbol{x}$ and $\boldsymbol{w}$. In these experiments, $\boldsymbol{x}$ is a random vector in the subspace spanned by the columns of an $L \times N$ matrix whose entries are independent and identically distributed Gaussian random variables. In part (a), $\boldsymbol{w}$ is a generic sparse vector, with support and nonzero entries chosen randomly. In part (b) $\boldsymbol{w}$ is a generic short vector whose first K terms are nonzero and chosen randomly.*

Our first set of experiments mimics the channel protection problem from Section 1.5 and Figure 1. Figure 3 shows the empirical rate of success when $\boldsymbol{C}$ is taken as a dense $L \times N$ Gaussian random matrix. We fix $L = 2048$ and vary N and K from 25 to 1000. In Figure 3(a), we take $\boldsymbol{w}$ to be sparse with known support; we form $\boldsymbol{B}$ by randomly selecting K columns from the $L \times L$ identity matrix. For Figure 3(b), we take $\boldsymbol{w}$ to be "short", forming $\boldsymbol{B}$ from the first K columns of the identity. In both cases, the basis expansion coefficient were drawn to be iid Gaussian random vectors. In both cases, we are able to deconvolve this signals with a high rate of success when $L \gtrsim 2.7(K + N)$.

Figure 4 shows the results of a similar experiment, only here both $\boldsymbol{w}$ and $\boldsymbol{x}$ are randomly generated sparse vectors. We take L to be much larger than the previous experiment, $L = 32,768$, and vary N and K from 1000 to $16,000$. In Figure 4(a), we generate both $\boldsymbol{B}$ and $\boldsymbol{C}$ by randomly selecting columns of the identity — despite the difference in the model for $\boldsymbol{x}$ (sparse instead of randomly oriented) the resulting performance curve in this case is very similar to that in Figure 3(a). In Figure 4(b), we use the same model for $\boldsymbol{C}$ and $\boldsymbol{x}$, but use a "short" $\boldsymbol{w}$ (first K terms are non-zero). Again, despite the difference in the model for $\boldsymbol{x}$, the recovery curve looks almost identical to that in Figure 3(b).

2.3 Recovery in the presence of noise

Figure 5 demonstrates the robustness of the deconvolution algorithm in the presence of noise. We use the same basic experimental setup as in Figure 3(a), with $L = 2048$, $N = 500$ and $K = 250$, but instead of making a clean observation of $\boldsymbol{w} * \boldsymbol{x}$, we add a noise vector $\boldsymbol{z}$ whose entires are iid Gaussian with zero mean and variance σ^2. We solve the program (17) with $\delta = (L + \sqrt{4L})^{1/2}\sigma$, a value chosen since it will be an upper bound for $\|\boldsymbol{z}\|_2$ with high probability.

Figure 5(a) shows how the relative error of the recovery changes with the noise level σ. On a log-log scale, the recovery error (show as $10\log_{10}$ (relative error squared)) is linear in the signal-to-

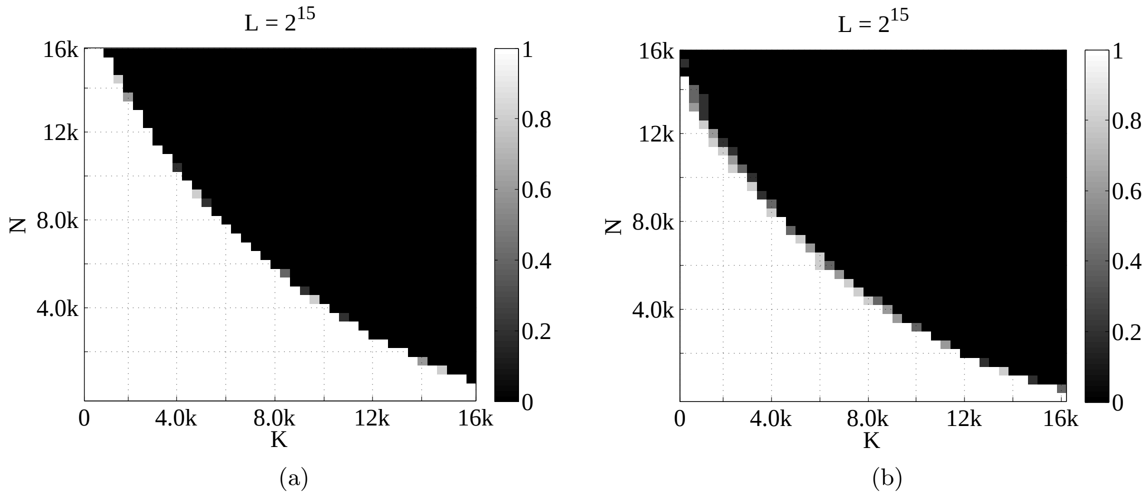


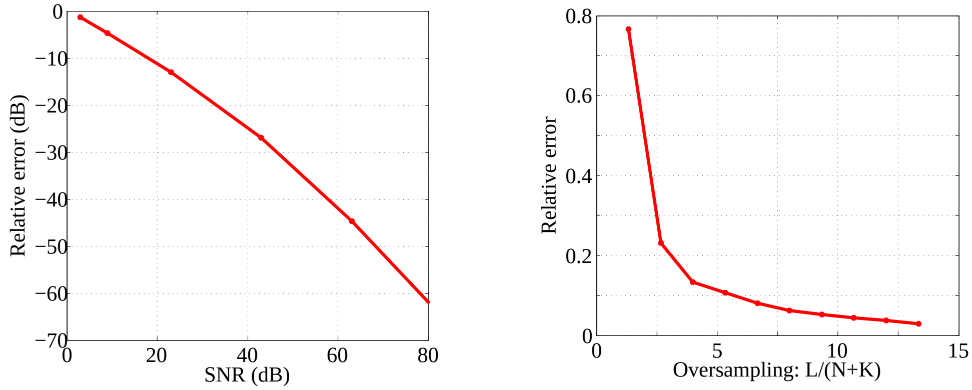
Figure 4: *Empirical success rate for the deconvolution of two vectors* $\boldsymbol{x}$ *and* $\boldsymbol{w}$*. In these experiments,* $\boldsymbol{x}$ *is a random sparse vector whose support and* N *non-zero values on that support are chosen at random. In part (a),* $\boldsymbol{w}$ *is a generic sparse vector, with support and* K *nonzero entries chosen randomly. In part (b)* $\boldsymbol{w}$ *is a generic short vector whose first* K *terms are nonzero and chosen randomly.*

noise ratio (defined as SNR$= 10\log_{10}(\|\boldsymbol{w}\boldsymbol{x}^*\|_F^2/\|\boldsymbol{z}\|_2^2)$. For each SNR level, we calculate the average relative error squared over 100 iterations, each time using independent set of signals, coding matrix, and noise. Figure 5(b) shows how the recovery error is affected by the "oversampling ratio"; as L is made larger relative to $N+K$, the recovery error decreases. As before, each point is averaged over 100 independent iterations.

2.4 Image deblurring

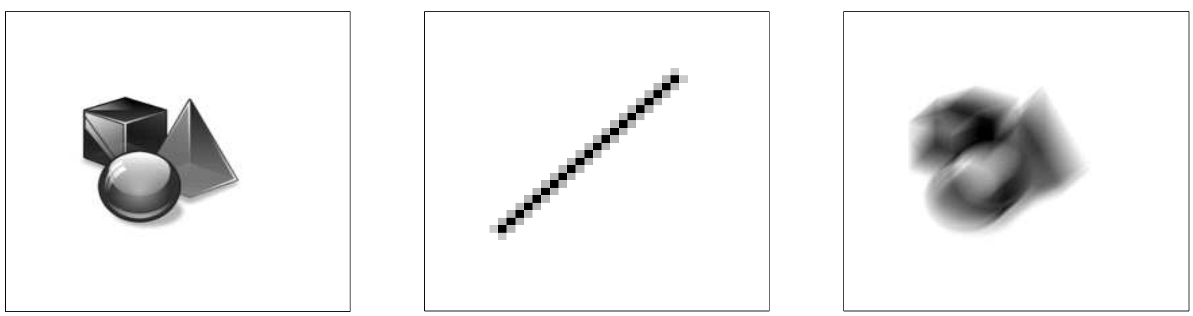
Figure 6, 7, and 8 illustrate an application of our blind deconvolution technique to two image deblurring problems. In the first problem, we assume that we have oracle knowledge of a low-dimensional subspace in which the image to be recovered lies. We observe a convolution of the 65,536 pixel Shapes image shown in Figure 6(a) with the motion blurring kernel shown in Figure 6(b); the observation is shown in Figure 6(c). The Shapes image can be very closely approximated using only $N = 5000$ terms in a Haar wavelet expansion, which capture 99.9% of the energy in the image. We start by assuming (perhaps unrealistically) that we know the indices for these most significant wavelet coefficients; the corresponding wavelet basis functions are taken as columns of $\boldsymbol{B}$. We will also assume that we know the support of the blurring kernel, which consists of $K = 65$ connected pixels; the corresponding columns of the identity constitute $\boldsymbol{C}$. The image and blur kernel recovered by solving (9) are shown in Figure 7.

Figure 8 shows a more realistic example where the support of the image in the wavelet domain is unknown. We take the blurred image shown in Figure 6(c) and, as before, we assume we know the support of the blurring kernel shown in Figure 6(b), with $K = 65$ non-zero elements, but here we use the blurred image to estimate the support in the wavelet domain — we take the Haar wavelet transform of the image in Figure 6(c), and select the indices of the $N = 9000$ largest wavelet coefficients as a proxy for the support of the significant coefficients of the original image. The wavelet coefficients of the original image at this estimated support capture 98.5% of the energy in



(a) (b)

Figure 5: *Performance of the blind deconvolution program in the presence of noise. In all of the experiments,* $L = 2048$, $N = 500$, $K = 250$, $\boldsymbol{B}$ *is a random selection of columns from the identity, and* $\boldsymbol{C}$ *is an iid Gaussian matrix. (a) Relative error vs. SNR on a log-log scale. (b) Oversampling rate vs. relative error for a fixed SNR of* $20dB$



(a) (b) (c)

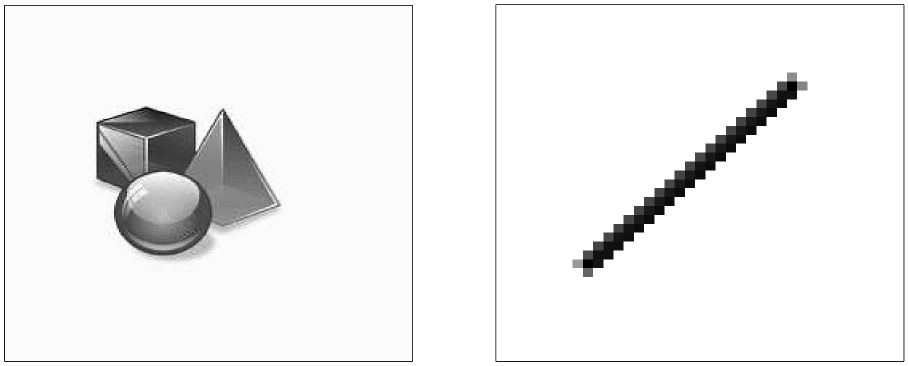
Figure 6: *Shapes image for deblurring experiment. (a) Original* 256×256 *Shapes image* $\boldsymbol{x}$*. (b) Blurring kernel* $\boldsymbol{w}$ *with a support size of 65 pixels, the locations of which are assumed to be known. (c) Convolution of (a) and (b).*

the blurred image. The recovery using (9) run with these linear models is shown in Figure 8(a) and Figure 8(b). Despite not knowing the linear model explicitly, we are able to estimate it well enough from the observed data to get a reasonable reconstruction.

3 Proof of main theorems

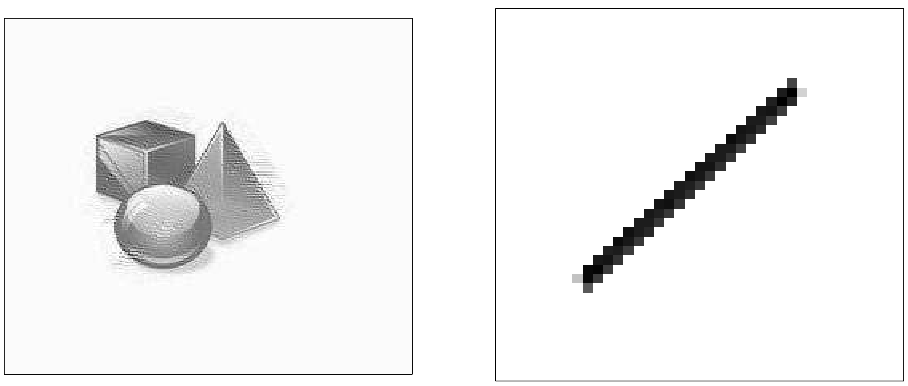
In this section, we will prove Theorems 1 and 2 by establishing a set of standard sufficient conditions for $\boldsymbol{X}_0$ to be the unique minimizer of (9). At a high level, the argument follows previous literature [21, 34] on low-rank matrix recovery by constructing a valid *dual certificate* for the rank-1 matrix $\boldsymbol{X}_0 = \boldsymbol{h}\boldsymbol{m}^*$. The main mathematical innovation in proving these results comes in Lemmas 1, 2, 3 and 4, which control the behavior of the random operator $\mathcal{A}$.

We will work through the main argument in this section, leaving the technical details (including



(a) (b)

Figure 7: *An oracle assisted image deblurring experiment; we assume that we know the support of the* 5000 *most significant wavelet coefficients of the original image. These wavelet coefficients capture 99.9% of the energy in the original image. We obtain from the solution of* (9)*: (a) Deconvolved image* $\hat{\boldsymbol{x}}$ *obtained from the solution of* (9)*, with relative error of* $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2 = 1.6 \times 10^{-2}$*. (b) Estimated blur kernel* $\hat{\boldsymbol{w}}$ *with relative error of* $\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_2/\|\boldsymbol{w}\|_2 = 5.4 \times 10^{-1}$.



(a) (b)

Figure 8: *Image recovery without oracle information. Take the support of the 9000 most-significant coefficients of Haar wavelet transform of the blurred image as our estimate of the subspace in which original image lives. (a) Deconvolved image obtained from the solution of* (9)*, with relative error of* 4.9×10^{-2}*. (b) Estimated blur kernel; relative error* $= 5.6 \times 10^{-1}$.

the proofs of the main lemmas) until Sections 5 and 6.

Key to our argument is the subspace (of $\mathbb{R}^{K\times N}$) T associated with $\boldsymbol{X}_0 = \boldsymbol{h}\boldsymbol{m}^*$:

$$T = \left\{\boldsymbol{X} : \boldsymbol{X} = \alpha\boldsymbol{h}\boldsymbol{v}^* + \beta\boldsymbol{u}\boldsymbol{m}^*,\ \boldsymbol{v}\in\mathbb{R}^N,\ \boldsymbol{u}\in\mathbb{R}^K,\ \alpha,\beta\in\mathbb{R}\right\}$$

with the (matrix) projection operators

$$\begin{aligned}\mathcal{P}_T(\boldsymbol{X}) &= \boldsymbol{P_H}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{P_M} - \boldsymbol{P_H}\boldsymbol{X}\boldsymbol{P_M}\\ \mathcal{P}_{T^\perp}(\boldsymbol{X}) &= (\boldsymbol{I} - \boldsymbol{P_H})\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P_M}),\end{aligned}$$

where $\boldsymbol{P_H}$ and $\boldsymbol{P_M}$ are the (vector) projection matrices $\boldsymbol{P_H} = \boldsymbol{h}\boldsymbol{h}^*$ and $\boldsymbol{P_M} = \boldsymbol{m}\boldsymbol{m}^*$.

3.1 Theorem 1: Sufficient condition for a nuclear norm minimizer

The following proposition is a specialization of the more general sufficient conditions for verifying the solutions to the nuclear norm minimization problem (9) that have appeared multiple times in the literature in one form or another (see [18], for example).

Proposition 1. *The matrix* $\boldsymbol{X}_0 = \boldsymbol{h}\boldsymbol{m}^*$ *is the unique minimizer to* (9) *if there exists a* $\boldsymbol{Y} \in \operatorname{Range}(\mathcal{A}^*)$ *such that*

$$\langle \boldsymbol{h}\boldsymbol{m}^* - \mathcal{P}_T(\boldsymbol{Y}), \mathcal{P}_T(\boldsymbol{Z})\rangle_F - \langle \mathcal{P}_{T^\perp}(\boldsymbol{Y}), \mathcal{P}_{T^\perp}(\boldsymbol{Z})\rangle_F + \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_* > 0$$

for all $\boldsymbol{Z} \in \operatorname{Null}(\mathcal{A})$.

Since

$$\begin{aligned}\langle \boldsymbol{h}\boldsymbol{m}^* - \mathcal{P}_T(\boldsymbol{Y}), \mathcal{P}_T(\boldsymbol{Z})\rangle_F &- \langle \mathcal{P}_{T^\perp}(\boldsymbol{Y}), \mathcal{P}_{T^\perp}(\boldsymbol{Z})\rangle_F + \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_* \\ &\geq -\|\boldsymbol{h}\boldsymbol{m}^* - \mathcal{P}_T(\boldsymbol{Y})\|_F \|\mathcal{P}_T(\boldsymbol{Z})\|_F - \|\mathcal{P}_{T^\perp}(\boldsymbol{Y})\| \, \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_* + \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_*,\end{aligned}$$

it is enough to find a $\boldsymbol{Y} \in \operatorname{Range}(\mathcal{A}^*)$ such that

$$-\|\boldsymbol{h}\boldsymbol{m}^* - \mathcal{P}_T(\boldsymbol{Y})\|_F \|\mathcal{P}_T(\boldsymbol{Z})\|_F + (1 - \|\mathcal{P}_{T^\perp}(\boldsymbol{Y})\|) \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_* \ > \ 0, \tag{21}$$

for all $\boldsymbol{Z} \in \operatorname{Null}(\mathcal{A})$.

In Lemma 1 in Section 3.4 below we show that $\|\mathcal{A}\| \leq \sqrt{5\mu_1^2 NK/L} =: \gamma \leq \sqrt{N}$ with appropriately high probability. Corollary 2 below also shows that with L obeying (16),

$$\|\mathcal{A}(\mathcal{P}_T(\boldsymbol{Z}))\|_F \geq 2^{-1/2} \|\mathcal{P}_T(\boldsymbol{Z})\|_F \quad \text{for all} \quad \boldsymbol{Z} \in \operatorname{Null}(\mathcal{A}),$$

with the appropriate probability. Then, since

$$\begin{aligned}0 &= \|\mathcal{A}(\boldsymbol{Z})\|_F \\ &\geq \|\mathcal{A}(\mathcal{P}_T(\boldsymbol{Z}))\|_F - \|\mathcal{A}(\mathcal{P}_{T^\perp}(\boldsymbol{Z}))\|_F \\ &\geq \frac{1}{\sqrt{2}} \|\mathcal{P}_T(\boldsymbol{Z})\|_F - \gamma \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_F,\end{aligned}$$

we will have that

$$\|\mathcal{P}_T(\boldsymbol{Z})\|_F \ \leq \ \sqrt{2}\gamma \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_F \ \leq \ \sqrt{2}\gamma \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_*. \tag{22}$$

Applying this fact to (21), we see that it is sufficient to find a $\boldsymbol{Y} \in \operatorname{Range}(\mathcal{A}^*)$ such that

$$\left(1 - \sqrt{2}\gamma \|\boldsymbol{h}\boldsymbol{m}^* - \mathcal{P}_T(\boldsymbol{Y})\|_F - \|\mathcal{P}_{T^\perp}(\boldsymbol{Y})\|\right) \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_* \ > \ 0.$$

Since Lemma 2 also implies that $\mathcal{P}_{T^\perp}(\boldsymbol{Z}) \neq \boldsymbol{0}$ for $\boldsymbol{Z} \in \operatorname{Null}(\mathcal{A})$, our approach will be to construct a $\boldsymbol{Y} \in \operatorname{Range}(\mathcal{A}^*)$ such that

$$\|\boldsymbol{h}\boldsymbol{m}^* - \mathcal{P}_T(\boldsymbol{Y})\|_F \leq \frac{1}{4\sqrt{2}\gamma} \quad \text{and} \quad \|\mathcal{P}_{T^\perp}(\boldsymbol{Y})\| < \frac{3}{4}. \tag{23}$$

In the next section, we will show how such a $\boldsymbol{Y}$ can be found using Gross's *golfing scheme* [20, 34].

3.2 Construction of the dual certificate via golfing

The golfing scheme works by dividing the L linear observations of $\boldsymbol{X}_0$ into P disjoint subsets of size Q, and then using these subsets of observations to iteratively construct the dual certificate $\boldsymbol{Y}$. We index these subsets by $\Gamma_1, \Gamma_2, \ldots, \Gamma_P$; by construction $|\Gamma_p| = Q$, $\bigcup_p \Gamma_p = \{1, \ldots, L\}$, and $\Gamma_p \cap \Gamma_{p'} = \emptyset$. We define $\mathcal{A}_p$ be the operator that returns the measurements indexed by the set Γ_p:

$$\mathcal{A}_p(\boldsymbol{W}) = \{\text{trace}(\boldsymbol{c}_k \boldsymbol{b}_k^* \boldsymbol{W})\}_{k\in\Gamma_p}, \qquad \mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W} = \sum_{k\in\Gamma_p} \boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}\boldsymbol{c}_k\boldsymbol{c}_k^*.$$

The $\mathcal{A}_p^*\mathcal{A}_p$ are random linear operators; the expectation of their action on a fixed matrix $\boldsymbol{W}$ is

$$\mathrm{E}[\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}] = \sum_{k\in\Gamma_p} \boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}.$$

For reasons that will become clear as we proceed through the argument below, we would like this expectation to be as close to a scalar multiple of $\boldsymbol{W}$ as possible for all p. In other words, we would like to partition the L rows of the matrix $\hat{\boldsymbol{B}}$ into P different $Q \times K$ submatrices, each of which is well-conditioned (i.e. the columns are almost orthogonal to one another).

Results from the literature on compressive sensing have shown that such a partition exists for $L\times K$ matrices with orthonormal columns whose rows all have about the same energy. In particular, the proof of Theorem 1.2 in [35] shows that if $\hat{\boldsymbol{B}}$ is a $L \times K$ matrix with $\hat{\boldsymbol{B}}^*\hat{\boldsymbol{B}} = \boldsymbol{I}$, Γ is a randomly selected subset of $\{1, \ldots, L\}$ of size Q, and the rows $\boldsymbol{b}_k^*$ of $\hat{\boldsymbol{B}}$ have coherence μ_1^2 as in (11), then there exists a constant C such that for any $0 < \epsilon < 1$ and $0 < \delta < 1$,

$$Q \;\geq\; C\frac{\mu_1^2 K}{\epsilon^2}\max\{\log K, \log(1/\delta)\},$$

implies

$$\left\|\sum_{k\in\Gamma} \boldsymbol{b}_k\boldsymbol{b}_k^* - \frac{Q}{L}\boldsymbol{I}\right\| \;\leq\; \frac{\epsilon Q}{L}.$$

with probability exceeding $1-\delta$. If our partition $\Gamma_1, \Gamma_2, \ldots, \Gamma_P$ is random, then, applying the above result with $\delta = (KN)^{-1}$ and $\epsilon = 1/4$ tells us that if

$$Q \;\geq\; C\,\mu_1^2 K \log(KN), \tag{24}$$

then

$$\max_{1\leq p\leq P}\left\|\sum_{k\in\Gamma_p} \boldsymbol{b}_k\boldsymbol{b}_k^* - \frac{Q}{L}\boldsymbol{I}\right\| \;\leq\; \frac{Q}{4L} \quad\Rightarrow\quad \max_{1\leq p\leq P}\left\|\sum_{k\in\Gamma_p} \boldsymbol{b}_k\boldsymbol{b}_k^*\right\| \;\leq\; \frac{5Q}{4L}, \tag{25}$$

with positive probability. This means that with Q chosen to obey (24), at least one such partition must exist and we move forward assuming that (25) holds.

Along with the expectation of each of the $\mathcal{A}_p^*\mathcal{A}_p$ being close to a multiple of the identity, we will also need tail bounds stating that $\mathcal{A}_p^*\mathcal{A}_p$ is close to its expectation with high probability. These probabilities can be made smaller by making the subset size Q larger. As detailed below, we will need

$$Q \;\geq\; C_\alpha\, M\log(KN)\log M \quad \text{where} \quad M = \max\left(\mu_1^2 K, \mu_h^2 N\right), \tag{26}$$

to make these probability bounds meaningful. Note that this means (24) will hold.

The construction of $\boldsymbol{Y}$ that obeys the conditions (23) relies on three technical lemmas which are stated below in Section 3.4. Their proofs rely heavily on re-writing different quantities of interest (linear operators, vectors, and scalars) as a sum of independent subexponential random variables and then using a specialized version of the "Matrix Bernstein Inequality" to estimate their sizes. Section 4 below contains a brief overview of these types of probabilistic bounds. The proofs of the key lemmas (2, 3, and 4) are in Section 5. These proofs rely on several miscellaneous lemmas which compute simple expectations and tail bounds for various random variables; these are presented separately in Section 6.

With the Γ_p chosen and the key lemmas established, we construct $\boldsymbol{Y}$ as follows. Let $\boldsymbol{Y}_0 = \boldsymbol{0}$, and then iteratively define

$$\boldsymbol{Y}_p = \boldsymbol{Y}_{p-1} + \frac{L}{Q}\,\mathcal{A}_p^*\mathcal{A}_p\left(\boldsymbol{h}\boldsymbol{m}^* - \mathcal{P}_T(\boldsymbol{Y}_{p-1})\right).$$

We will show that under appropriate conditions on L, taking $\boldsymbol{Y} := \boldsymbol{Y}_P$ will satisfy both parts of (23) with high probability.

Let $\boldsymbol{W}_p$ be the residual between $\boldsymbol{Y}_p$ projected onto T and the target $\boldsymbol{h}\boldsymbol{m}^*$:

$$\boldsymbol{W}_p = \mathcal{P}_T(\boldsymbol{Y}_p) - \boldsymbol{h}\boldsymbol{m}^*.$$

Notice that $\boldsymbol{W}_p \in T$ and

$$\boldsymbol{W}_0 = -\boldsymbol{h}\boldsymbol{m}^*, \qquad \boldsymbol{W}_p = \frac{L}{Q}\left(\frac{Q}{L}\mathcal{P}_T - \mathcal{P}_T\mathcal{A}_p^*\mathcal{A}_p\mathcal{P}_T\right)\boldsymbol{W}_{p-1}. \tag{27}$$

Applying Lemma 2 iteratively to the $\boldsymbol{W}_p$ tells us that

$$\|\boldsymbol{W}_p\|_F \;\le\; \frac{1}{2}\|\boldsymbol{W}_{p-1}\|_F \;\le\; 2^{-p}\|\boldsymbol{h}\boldsymbol{m}^*\|_F = 2^{-p}, \qquad p = 1,\ldots,P, \tag{28}$$

with probability exceeding $1 - 3P(KN)^{-\alpha}$. Thus we will have

$$\|\boldsymbol{h}\boldsymbol{m}^* - \mathcal{P}_T(\boldsymbol{Y}_P)\|_F \le \frac{1}{4\sqrt{2}\gamma},$$

for

$$P \;\ge\; \frac{\log(4\sqrt{2}\gamma)}{\log 2}, \quad \text{which can be achieved with} \quad L \;\ge\; C\,Q\log(KN),$$

which will holds with our assumptions on L in the theorem statement (16) and our choice of Q in (26).

To bound $\|\mathcal{P}_{T^\perp}(\boldsymbol{Y}_p)\|$, we use the expansion

$$\begin{aligned}\boldsymbol{Y}_p = \boldsymbol{Y}_{p-1} - \frac{L}{Q}\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} \;&=\; \boldsymbol{Y}_{p-2} - \frac{L}{Q}\mathcal{A}_{p-1}^*\mathcal{A}_{p-1}\boldsymbol{W}_{p-2} - \frac{L}{Q}\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} \;=\; \cdots \\ = -\sum_{p=1}^{P}\frac{L}{Q}\mathcal{A}_{p-1}^*\mathcal{A}_{p-1}\boldsymbol{W}_p,\end{aligned}$$

and so

$$
\begin{aligned}
\|\mathcal{P}_{T^\perp}(\boldsymbol{Y}_P)\| &= \left\|\mathcal{P}_{T^\perp}\left(\sum_{p=1}^{P}\frac{L}{Q}\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}\right)\right\| \\
&= \frac{L}{Q}\left\|\mathcal{P}_{T^\perp}\left(\sum_{p=1}^{P}\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \frac{Q}{L}\boldsymbol{W}_{p-1}\right)\right\|, \qquad \text{(since } \boldsymbol{W}_{p-1}\in T\text{)} \\
&\le \frac{L}{Q}\left\|\sum_{p=1}^{P}\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \frac{Q}{L}\boldsymbol{W}_{p-1}\right\| \\
&\le \sum_{p=1}^{P}\frac{L}{Q}\left\|\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \frac{Q}{L}\boldsymbol{W}_{p-1}\right\|.
\end{aligned}
$$

Lemma 4 shows that with probability exceeding $1 - P(KN)^{-2}$,

$$\left\|\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \frac{Q}{L}\boldsymbol{W}_{p-1}\right\| \;\le\; 2^{-p}\frac{3Q}{4L}, \quad \text{for all} \quad p = 1,\ldots,P.$$

and so

$$\|\mathcal{P}_{T^\perp}(\boldsymbol{Y}_P)\| \;\le\; \sum_{p=1}^{P} 3\cdot 2^{-p-2} \;<\; \frac{3}{4}.$$

Collecting the results above, we see that both conditions in (23) will hold with probability exceeding $1 - O(K(NK)^{-\alpha})$ when Q is chosen as in (26) and L is chosen as in (16).

3.3 Theorem 2: Stability

With the conditions on L, we can establish three important intermediate results, each of which occurs with probability $1 - O(L(NK)^{-1})$. The first is the existence of a dual certificate $\boldsymbol{Y}$, as constructed in the previous section, that obeys the conditions (23). The second, given to us by Lemma 1, is that the operator $\mathcal{A}\mathcal{A}^*$ is well conditioned, meaning that its eigenvalues obey

$$\frac{\mu_L^2 NK}{2L} \;\le\; \lambda_{\min}(\mathcal{A}\mathcal{A}^*) \;\le\; \lambda_{\max}(\mathcal{A}\mathcal{A}^*) \;\le\; \frac{4.5\mu_1^2 NK}{L}. \tag{29}$$

The third, given to us by Lemma 2 is that in addition, $\mathcal{A}^*\mathcal{A}$ is well conditioned on T: $\|\mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T - \mathcal{P}_T\| \le 1/2$.

With these facts in place, the stability proof follows the template set in [34,36]. We start with two observations; first, the feasibility of $\boldsymbol{X}_0$ implies

$$\|\tilde{\boldsymbol{X}}\|_* \le \|\boldsymbol{X}_0\|_*, \tag{30}$$

and

$$\|\mathcal{A}(\tilde{\boldsymbol{X}} - \boldsymbol{X}_0)\|_2 \le \|\hat{\boldsymbol{y}} - \mathcal{A}(\boldsymbol{X}_0)\|_2 + \|\mathcal{A}(\tilde{\boldsymbol{X}}) - \hat{\boldsymbol{y}}\|_2 \le 2\delta. \tag{31}$$

Set $\tilde{\boldsymbol{X}} = \boldsymbol{X}_0 + \boldsymbol{\xi}$. The result (29) tells us that the orthogonal projection $\mathcal{P}_\mathcal{A} = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}$ onto the range of $\mathcal{A}^*$ is well defined. We then break apart the recovery error as

$$\|\boldsymbol{\xi}\|_F^2 = \|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_F^2 + \|\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_F^2 \tag{32}$$

$$= \|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_F^2 + \|\mathcal{P}_T\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_F^2 + \|\mathcal{P}_{T^\perp}\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_F^2. \tag{33}$$

A direct result of of Proposition 1 is that there exists a constant $C > 0$ such that for all $\boldsymbol{Z} \in \text{Null}(\mathcal{A})$, $\|\boldsymbol{X}_0 + \boldsymbol{Z}\|_* - \|\boldsymbol{X}_0\|_* \geq C\|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\|_*$ (this is developed cleanly in [18]). Since $\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi}) \in \text{Null}(\mathcal{A})$, we have

$$\|\boldsymbol{X}_0 + \mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_* - \|\boldsymbol{X}_0\|_* \;\geq C\|\mathcal{P}_{T^\perp}\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_*.$$

Combining this with (30) and the triangle inequality yields

$$\|\boldsymbol{X}_0\|_* \geq \|\boldsymbol{X}_0\|_* + C\|\mathcal{P}_{T^\perp}\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_* - \|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_*,$$

which implies

$$\begin{aligned}\|\mathcal{P}_{T^\perp}\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_* &\leq C\|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_* \\ &\leq C\sqrt{\min(K,N)}\|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_F.\end{aligned}$$

In addition, since we established (22), with $\gamma = \sqrt{4.5\,\mu_1^2 KN/L}$ given by Lemma 1, for all $\boldsymbol{Z} \in \text{Null}(\mathcal{A})$, we have that

$$\|\mathcal{P}_T\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_F^2 \;\leq\; 2\gamma^2\|\mathcal{P}_{T^\perp}\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_F^2,$$

and as a result

$$\|\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_F^2 \;\leq\; (2\gamma^2+1)\|\mathcal{P}_{T^\perp}\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_F^2.$$

Revisiting (32), we have

$$\begin{aligned}\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_F^2 &\leq (2\gamma^2+1)\|\mathcal{P}_{T^\perp}\mathcal{P}_{\mathcal{A}^\perp}(\boldsymbol{\xi})\|_F^2 + \|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_F^2 \\ &\leq C(2\gamma^2+1)\min(K,N)\|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_F^2 + \|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_F^2,\end{aligned}$$

and then absorbing all the constants into C,

$$\begin{aligned}\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_F &\leq C\gamma\sqrt{\min(K,N)}\|\mathcal{P}_\mathcal{A}(\boldsymbol{\xi})\|_F \\ &\leq C\sqrt{\min(K,N)}\gamma\|\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\|\,\|\mathcal{A}(\boldsymbol{\xi})\|_2.\end{aligned}$$

Using (29), and (31), we obtain the final result

$$\|\tilde{\boldsymbol{X}} - \boldsymbol{X}_0\|_F \leq C\frac{\mu_1}{\mu_L}\sqrt{\min(K,N)}\delta. \tag{34}$$

3.4 Key lemmas

Lemma 1 ($\mathcal{A}\mathcal{A}^*$ is well conditioned.)**.** *Let $\mathcal{A}$ be as defined in* (4)*, with coherences μ_1^2 and μ_L^2 as defined in* (11) *and* (12)*. Suppose that $\mathcal{A}$ is sufficiently underdetermined in that*

$$L \;\leq\; \frac{\mu_L^2 NK}{8\beta^2\log^2(NK)} \tag{35}$$

for some constant $\beta > 1$. Then with probability exceeding $1 - O(L(NK)^{-\beta})$, the eigenvalues of $\mathcal{A}\mathcal{A}^*$ obey

$$0.48\mu_L^2\frac{NK}{L} \leq \lambda_{\min}(\mathcal{A}\mathcal{A}^*) \leq \lambda_{\max}(\mathcal{A}\mathcal{A}^*) \leq 4.5\mu_1^2\frac{NK}{L}.$$

The proof of Lemma 1 in Section 5 decomposes $\mathcal{A}\mathcal{A}^*$ as a sum of independent random matrices, and then applies a Chernoff-like bound discussed in Section 4.

Lemma 2 (Conditioning on T)**.** *With the coherences μ_1^2 and μ_h^2 defined in Section 1.3, let*

$$M = \max(\mu_1^2 K, \mu_h^2 N). \tag{36}$$

Fix $\alpha \geq 1$. Choose the subsets $\Gamma_1, \ldots, \Gamma_P$ described in Section 3.2 so that they have size

$$|\Gamma_p| = Q \;=\; C'_\alpha \cdot M \log(NK) \log(M), \tag{37}$$

where $C'_\alpha = O(\alpha)$ is a constant chosen below, and such that (25) holds. Then the linear operators $\mathcal{A}_1, \ldots, \mathcal{A}_P$ defined in Section 3.2 will obey

$$\max_{1 \leq p \leq P} \left\| \mathcal{P}_T \mathcal{A}_p^* \mathcal{A}_p \mathcal{P}_T - \frac{Q}{L} \mathcal{P}_T \right\| \;\leq\; \frac{Q}{2L},$$

with probability exceeding $1 - 3P(KN)^{-\alpha}$.

Corollary 2. *Let $\mathcal{A}$ be the operator defined in (4), and M be defined as in (36). Then there exists a constant $C'_\alpha = O(\alpha)$ such that*

$$L \;\geq\; C'_\alpha \cdot M \log(KN) \log(M), \tag{38}$$

implies

$$\|\mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T - \mathcal{P}_T\| \;\leq\; \frac{1}{2},$$

with probability exceeding $1 - 3(KN)^{-\alpha}$.

Lemma 3. *Let M, Q, the Γ_p, and the $\mathcal{A}_p$ be the same as in Lemma 2. Let $\boldsymbol{W}_p$ be as in (27), and define*

$$\mu_p^2 = L \max_{\ell \in \Gamma_{p+1}} \|\boldsymbol{W}_p^* \boldsymbol{b}_\ell\|_2^2. \tag{39}$$

Then there exists a constant $C_\alpha = O(\alpha)$ such that if

$$L \;\geq\; C_\alpha \, M \log(KN) \sqrt{\log M} \tag{40}$$

then

$$\mu_p \;\leq\; \frac{\mu_{p-1}}{2}, \quad \textit{for} \;\; p = 1, \ldots, P, \tag{41}$$

with probability exceeding $1 - 2L(KN)^{-\alpha}$.

Lemma 4. *Let α, M, Q, the Γ_p, and the $\mathcal{A}_p$ be the same as in Lemma 2, and μ_p and $\boldsymbol{W}_p$ be the same as in Lemma 3. Assume that (28) and (41) hold:*

$$\|\boldsymbol{W}_{p-1}\|_F \leq 2^{-p+1} \quad \textit{and} \quad \mu_{p-1} \leq 2^{-p+1} \mu_h.$$

Then with probability exceeding $1 - P(KN)^{-\alpha}$,

$$\max_{1 \leq p \leq P} \left\| \mathcal{A}_p^* \mathcal{A}_p \boldsymbol{W}_{p-1} - \frac{Q}{L} \boldsymbol{W}_{p-1} \right\| \;\leq\; 2^{-p} \frac{3Q}{4L}.$$

4 Concentration inequalities

Proving the key lemmas stated in Section 3.4 revolves around estimating the sizes of sums of different subexponential random variables. These random variables are either the absolute value of a sum of independent random scalars, the euclidean norm of a sum of independent random vectors (or equivalently, the Frobenius norm of a sum of random matrices), or the operator norm (maximum singular value) of a sum of random linear operators. In this section, we very briefly overview the tools from probability theory that we will use to make these estimates. The essential tool is the recently developed matrix Bernstein inequality [37].

We start by recalling the classical scalar Bernstein inequality. A nice proof of the result in this form can be found in [38, Chapter 2].

Proposition 2 (Scalar Bernstein, subexponential version)**.** *Let $z_1, \ldots, z_K$ be independent random variables with* $\mathrm{E}[z_k] = 0$ *and*

$$\mathrm{P}\left\{|z_k| > u\right\} \;\leq\; C\mathrm{e}^{-u/\sigma_k}, \tag{42}$$

for some constants C and $\sigma_k,\; k = 1, \ldots, K$ with

$$\sigma^2 = \sum_{k=1}^{K} \sigma_k^2 \quad \text{and} \quad B = \max_{1\leq k\leq K} \sigma_k.$$

Then

$$\mathrm{P}\left\{|z_1 + \cdots + z_K| > u\right\} \;\leq\; 2\exp\left(\frac{-u^2}{2C\sigma^2 + 2Bu}\right),$$

and so

$$|z_1 + \cdots + z_K| \;\leq\; 2\max\left\{\sqrt{C}\sigma\sqrt{t + \log 2},\; 2B(t + \log 2)\right\}$$

with probability exceeding $1 - \mathrm{e}^{-t}$.

To make the statement (and usage) of the concentration inequalities more compact in the vector and matrix case, we will characterize subexponential vectors and matrices using their Orlicz-1 norm.

Definition 1. *Let $\boldsymbol{Z}$ be a random matrix. We will use $\|\cdot\|_{\psi_1}$ to denote the Orlicz-1 norm:*

$$\|Z\|_{\psi_1} = \inf_{u\geq 0}\left\{\mathrm{E}[\exp(\|Z\|/u)] \leq 2\right\},$$

where $\|\boldsymbol{Z}\|$ is the spectral norm of $\boldsymbol{Z}$. In the case where $\boldsymbol{Z}$ is a vector, we take $\|\boldsymbol{Z}\| = \|\boldsymbol{Z}\|_2$.

As the next basic result shows, the Orlicz-1 norm of a random variable can be systematically related to rate at which its distribution function approaches 1 (i.e. σ_k in (42)).

Lemma 5 (Lemma 2.2.1 in [38])**.** *Let z be a random variable which obeys* $\mathrm{P}\left\{|z| > u\right\} \leq \alpha\mathrm{e}^{-\beta u}$. *Then* $\|z\|_{\psi_1} \leq (1 + \alpha)/\beta$.

Using these definitions, we have the following powerful tool for bounding the size of a sum of independent random vectors or matrices, each one of which is subexponential. This result is mostly due to [37], but appears in the form below in [24].

Proposition 3 (Matrix Bernstein, Orlicz norm version)**.** *Let* $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_Q$ *be independent* $K \times N$ *random matrices with* $\mathrm{E}[\boldsymbol{Z}_{\boldsymbol{q}}] = \boldsymbol{0}$*. Let* B *be an upper bound on the Orlicz-1 norms:*

$$\max_{1 \le q \le Q} \|\boldsymbol{Z}_q\|_{\psi_1} \;\le\; B,$$

and define

$$\sigma^2 = \max\left\{ \left\| \sum_{q=1}^{Q} \mathrm{E}[\boldsymbol{Z}_q \boldsymbol{Z}_q^*] \right\|, \left\| \sum_{q=1}^{Q} \mathrm{E}[\boldsymbol{Z}_q^* \boldsymbol{Z}_q] \right\| \right\}. \tag{43}$$

Then there exists a constant C *such that for all* $t \ge 0$

$$\|\boldsymbol{Z}_1 + \cdots + \boldsymbol{Z}_Q\| \;\le\; C \max\left\{ \sigma\sqrt{t + \log(K+N)},\; B \log\left(\frac{\sqrt{Q}B}{\sigma}\right)(t + \log(K+N)) \right\}, \tag{44}$$

with probability at least $1 - e^{-t}$*.*

Essential to establishing our stability result, Theorem 2, is bounding both the upper and lower eigenvalues of the operator $\mathcal{A}\mathcal{A}^*$. We do this in Lemma 1 with a relatively straightforward application of the following Chernoff-like bound for sums of random positive symmetric matrices.

Proposition 4 (**Matrix Chernoff** [37])**.** *Let* $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{\boldsymbol{N}}$ *be independent* $L \times L$ *random self-adjoint matrices whose eigenvalues obey*

$$0 \le \lambda_{\min}(\boldsymbol{X}_n) \le \lambda_{\max}(\boldsymbol{X}_n) \le R \quad \textit{almost surely.}$$

Define

$$\mu_{\min} := \lambda_{\min}\left(\sum_{n=1}^{N} \mathrm{E}[\boldsymbol{X}_n]\right) \quad \textit{and} \quad \mu_{\max} := \lambda_{\max}\left(\sum_{n=1}^{N} \mathrm{E}[\boldsymbol{X}_n]\right).$$

Then

$$\mathrm{P}\left\{ \lambda_{\min}\left(\sum_{n=1}^{N} \boldsymbol{X}_n\right) \le t\mu_{\min} \right\} \;\le\; L\,\mathrm{e}^{-(1-t)^2\mu_{\min}/2R} \quad \textit{for} \quad t \in [0,1], \tag{45}$$

and

$$\mathrm{P}\left\{ \lambda_{\max}\left(\sum_{k} \boldsymbol{X}_k\right) \ge t\mu_{\max} \right\} \;\le\; L\left[\frac{\mathrm{e}}{t}\right]^{t\mu_{\max}/R} \quad \textit{for} \quad t \ge \mathrm{e}. \tag{46}$$

5 Proof of key lemmas

5.1 Proof of Lemma 1

The proof of Lemma 1 is essentially an application of the matrix Chernoff bound in Proposition 4.

Using the matrix form of $\mathcal{A}$,

$$\mathcal{A} = \begin{bmatrix} \Delta_1 \hat{\boldsymbol{B}} & \cdots & \Delta_N \hat{\boldsymbol{B}} \end{bmatrix},$$

we can write $\mathcal{A}\mathcal{A}^*$ as sum of random matrices

$$\mathcal{A}\mathcal{A}^* = \sum_{n=1}^{N} \Delta_n \hat{\boldsymbol{B}} \hat{\boldsymbol{B}}^* \Delta_n^*,$$

where $\Delta_n = \text{diag}(\{c_\ell[n]\}_\ell)$ as in (3). To apply Proposition 4, we will need to condition on the maximum of the magnitudes of the $c_\ell[n]$ not exceeding a certain size. To this end, given an α (which we choose later), we define the event

$$\Gamma_\alpha = \left\{ \max_{\substack{1\le n\le N \\ 1\le \ell\le L/2}} |c_\ell[n]| \le \alpha \right\}.$$

Then

$$\begin{aligned} \text{P}\left\{\lambda_{\max}(\mathcal{A}\mathcal{A}^*) > v\right\} &\le \text{P}\left\{\lambda_{\max}(\mathcal{A}\mathcal{A}^*) > v \mid \Gamma_\alpha\right\} \text{P}\left\{\Gamma_\alpha\right\} + \text{P}\left\{\Gamma_\alpha^c\right\} && (47) \\ &\le \text{P}\left\{\lambda_{\max}(\mathcal{A}\mathcal{A}^*) > v \mid \Gamma_\alpha\right\} + \text{P}\left\{\Gamma_\alpha^c\right\}, && (48) \end{aligned}$$

and similarly for $\text{P}\left\{\lambda_{\min}(\mathcal{A}\mathcal{A}^*) > v\right\}$. Conditioned on Γ_α, the complex Gaussian random variables $c_\ell[n]$ still zero mean and independent; we denote these conditional random variables as $c'_\ell[n]$, and set $\Delta'_n = \text{diag}(\{c'_\ell[n]\}_\ell)$, noting that

$$\text{E}[|c'_\ell[n]|^2] = \text{E}[|c_\ell[n]|^2 \mid \Gamma_\alpha] = \frac{1-(\alpha^2+1)\text{e}^{-\alpha^2}}{1-\text{e}^{-\alpha^2}} =: \sigma_\alpha^2 \le 1.$$

We now apply Proposition 4 with

$$\begin{aligned} R &= \max_n \left\{\lambda_{\max}(\Delta'_n \hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^* \Delta'^*_n)\right\} \\ &\le \max_n \left\{\lambda_{\max}(\Delta'_n)\lambda_{\max}(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^*)\lambda_{\max}(\Delta'^*_n)\right\} \\ &\le \alpha^2, \end{aligned}$$

and

$$\begin{aligned} \mu_{\max} &= \lambda_{\max}\left(\sum_{n=1}^N \text{E}[\Delta'_n \boldsymbol{B}\boldsymbol{B}^*\Delta'^*_n]\right) \\ &= N\lambda_{\max}\left(\text{E}[\Delta'_n \boldsymbol{B}\boldsymbol{B}^*\Delta'^*_n]\right) \\ &\le N\sigma_\alpha^2 \max_\ell \|\boldsymbol{b}_\ell\|_2^2 \\ &\le \mu_1^2 N\frac{K}{L}, \end{aligned}$$

and

$$\begin{aligned} \mu_{\min} &:= \lambda_{\min}\left(\sum_{n=1}^N \text{E}\left[\Delta'_n \hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^*\Delta'^*_n\right]\right) \\ &= N\sigma_\alpha^2 \min_\ell \|\boldsymbol{b}_\ell\|_2^2 \\ &= \sigma_\alpha^2\mu_L^2 N\frac{K}{L}, \end{aligned}$$

to get

$$\text{P}\left\{\lambda_{\min}(\mathcal{A}\mathcal{A}^*) < \frac{\sigma_\alpha^2\mu_L^2 NK}{2L} \;\middle|\; \Gamma_\alpha\right\} \;\le\; L\exp\left(-\frac{\sigma_\alpha^2\mu_L^2 NK}{8\alpha^2 L}\right),$$

where we have take $t = 1/2$ in (45), and

$$\mathrm{P}\left\{ \lambda_{\max}(\mathcal{A}\mathcal{A}^*) > \frac{\mathrm{e}^{3/2}\mu_1^2 NK}{L} \;\middle|\; \Gamma_\alpha \right\} \;\leq\; L\exp\left(-\frac{2\mu_1^2 NK}{\alpha^2 L}\right),$$

where we have taken $t = \mathrm{e}^{3/2}$ in (46).

Taking $\alpha = \beta \log(NK)$ for some $\beta > 1$ results in $\sigma_\alpha^2 \;>\; 0.97$ (for $N, K \geq 4$), and

$$\mathrm{P}\left\{\Gamma_\alpha^c\right\} \;\leq\; L(NK)^{-\beta}.$$

Taking L as in (35) yields

$$\begin{aligned}
\mathrm{P}\left\{ \lambda_{\min}(\mathcal{A}\mathcal{A}^*) < .48\mu_L^2 \frac{NK}{2L} \;\middle|\; \Gamma_\alpha \right\} &\leq L(NK)^{-\beta} \\
\mathrm{P}\left\{ \lambda_{\max}(\mathcal{A}\mathcal{A}^*) > 4.5\mu_1^2 \frac{NK}{L} \;\middle|\; \Gamma_\alpha \right\} &\leq L(NK)^{-8\beta\mu_1^2/\mu_L^2},
\end{aligned}$$

which establishes the lemma (since $\mu_1^2 \geq \mu_L^2$).

5.2 Proof of Lemma 2

The proof of the Lemma and its corollary follow the exact same line of argumentation. We will start with the conditioning of the partial operators $\mathcal{A}_p$ on T; after this, the argument for the conditioning of the full operator $\mathcal{A}$ will be clear.

We start by fixing p, and set $\Gamma = \Gamma_p$. With

$$\boldsymbol{A}_k = \boldsymbol{b}_k \boldsymbol{c}_k^*,$$

where the $\boldsymbol{b}_k \in \mathbb{C}^K$ obey (10),(11),(14) and the $\boldsymbol{c}_k \in \mathbb{C}^N$ are random vectors distributed as in (15), we are interested in how the random operator

$$\mathcal{P}_T \mathcal{A}_p^* \mathcal{A}_p \mathcal{P}_T = \sum_{k \in \Gamma} \mathcal{P}_T(\boldsymbol{A}_k) \otimes \mathcal{P}_T(\boldsymbol{A}_k)$$

concentrates around its mean in the operator norm. This operator is a sum of independent random rank-1 operators on $N \times K$ matrices, and so we can use the matrix Bernstein inequality in Proposition 3 to estimate its deviation.

Since $\boldsymbol{A}_k = \boldsymbol{b}_k \boldsymbol{c}_k^*$, $\mathcal{P}_T(\boldsymbol{A}_k)$ is the rank-2 matrix given by

$$\begin{aligned}
\mathcal{P}_T(\boldsymbol{A}_k) &= \langle \boldsymbol{b}_k, \boldsymbol{h} \rangle \boldsymbol{h} \boldsymbol{c}_k^* + \langle \boldsymbol{m}, \boldsymbol{c}_k \rangle \boldsymbol{b}_k \boldsymbol{m}^* - \langle \boldsymbol{b}_k, \boldsymbol{h} \rangle \langle \boldsymbol{m}, \boldsymbol{c}_k \rangle \boldsymbol{h} \boldsymbol{m}^* \\
&= \boldsymbol{h} \boldsymbol{v}_k^* + \boldsymbol{u}_k \boldsymbol{m}^*,
\end{aligned}$$

where $\boldsymbol{v}_k = \langle \boldsymbol{h}, \boldsymbol{b}_k \rangle \boldsymbol{c}_k$ and $\boldsymbol{u}_k = \langle \boldsymbol{m}, \boldsymbol{c}_k \rangle (\boldsymbol{b}_k - \langle \boldsymbol{b}_k, \boldsymbol{h} \rangle \boldsymbol{h}) = \langle \boldsymbol{m}, \boldsymbol{c}_k \rangle (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_k$.

The linear operator $\mathcal{P}_T(\cdot)$, since it maps $K \times N$ matrices to $K \times N$ matrix, can itself be represented as a $KN \times KN$ matrix that operates on a matrix that has been rasterized (in column order here) into a vector of length KN. We will find it convenient to denote these matrices in block form:

$\{M(i,j)\}_{i,j}$, where $M(i,j)$ is a $K \times K$ matrix that occupies rows $(i-1)K+1, \ldots, iK$ and columns $(j-1)K+1, \ldots, jK$. Using this notation, we can write $\mathcal{P}_T$ as the matrix

$$\mathcal{P}_T = \{\boldsymbol{h}\boldsymbol{h}^*\delta(i,j)\}_{i,j} + \{m[i]m[j]\boldsymbol{I}\}_{i,j} - \{m[i]m[j]\boldsymbol{h}\boldsymbol{h}^*\}_{i,j}, \tag{49}$$

where $\delta(i,j) = 1$ if $i = j$ and is zero otherwise.

We will make repeated use the following three facts about block matrices below:

1. Let $\mathcal{M}$ be an operator that we can write in matrix form as
$$\mathcal{M} = \{\boldsymbol{M}\delta(i,j)\}_{i,j}$$
for some $K \times K$ matrix $\boldsymbol{M}$. Then the action of $\mathcal{M}$ on a matrix $\boldsymbol{X}$ is
$$\mathcal{M}(\boldsymbol{X}) = \boldsymbol{M}\boldsymbol{X},$$
and so $\|\mathcal{M}\| = \|\boldsymbol{M}\|$. Also, $\mathcal{M}^*(\boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{X}$.
2. Now suppose we can write $\mathcal{M}$ in matrix form as
$$\mathcal{M} = \{p[i]^*q[j]\boldsymbol{I}\}_{i,j},$$
for some $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{C}^N$. Then the action of $\mathcal{M}$ on a matrix $\boldsymbol{X}$ is
$$\mathcal{M}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{q}\boldsymbol{p}^*,$$
and so $\|\mathcal{M}\| = \|\boldsymbol{q}\boldsymbol{p}^*\| = \|\boldsymbol{q}\|_2\|\boldsymbol{p}\|_2$. Also, $\mathcal{M}^*(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{p}\boldsymbol{q}^*$.
3. Now let
$$\mathcal{M} = \{p[i]^*q[j]\boldsymbol{M}\}_{i,j}.$$
Then the action of $\mathcal{M}$ on a matrix $\boldsymbol{X}$ is
$$\mathcal{M}(\boldsymbol{X}) = \boldsymbol{M}\boldsymbol{X}\boldsymbol{q}\boldsymbol{p}^*,$$
and so $\|\mathcal{M}\| = \|\boldsymbol{M}\|\,\|\boldsymbol{q}\boldsymbol{p}^*\| = \|\boldsymbol{M}\|\,\|\boldsymbol{q}\|_2\|\boldsymbol{p}\|_2$. Also $\mathcal{M}^*(\boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{X}\boldsymbol{p}\boldsymbol{q}^*$.

We will break $\mathcal{P}_T(\boldsymbol{A}_k) \otimes \mathcal{P}_T(\boldsymbol{A}_k)$ into four different tensor products of rank-1 matrices, and treat each one in turn:

$$\mathcal{P}_T(\boldsymbol{A}_k) \otimes \mathcal{P}_T(\boldsymbol{A}_k) = \boldsymbol{h}\boldsymbol{v}_k^* \otimes \boldsymbol{h}\boldsymbol{v}_k^* + \boldsymbol{h}\boldsymbol{v}_k^* \otimes \boldsymbol{u}_k\boldsymbol{m}^* + \boldsymbol{u}_k\boldsymbol{m}^* \otimes \boldsymbol{h}\boldsymbol{v}_k^* + \boldsymbol{u}_k\boldsymbol{m}^* \otimes \boldsymbol{u}_k\boldsymbol{m}^*. \tag{50}$$

To handle these terms in matrix form, note that if $\boldsymbol{u}_1\boldsymbol{v}_1^*$ and $\boldsymbol{u}_2\boldsymbol{v}_2^*$ are rank-1 matrices, with $\boldsymbol{u}_i \in \mathbb{C}^K$ and $\boldsymbol{v}_i \in \mathbb{C}^N$, then the operator given by their tensor product can be written as

$$\boldsymbol{u}_1\boldsymbol{v}_1^* \otimes \boldsymbol{u}_2\boldsymbol{v}_2^* = \begin{bmatrix} v_1[1]^*v_2[1]\boldsymbol{u}_1\boldsymbol{u}_2^* & v_1[1]^*v_2[2]\boldsymbol{u}_1\boldsymbol{u}_2^* & \cdots & v_1[1]^*v_2[N]\boldsymbol{u}_1\boldsymbol{u}_2^* \\ v_1[2]^*v_2[1]\boldsymbol{u}_1\boldsymbol{u}_2^* & v_1[2]^*v_2[2]\boldsymbol{u}_1\boldsymbol{u}_2^* & \cdots & v_1[2]^*v_2[N]\boldsymbol{u}_1\boldsymbol{u}_2^* \\ \vdots & & \ddots & \\ v_1[N]^*v_2[1]\boldsymbol{u}_1\boldsymbol{u}_2^* & \cdots & \cdots & v_1[N]^*v_2[N]\boldsymbol{u}_1\boldsymbol{u}_2^* \end{bmatrix} = \{v_1[i]^*v_2[j]\boldsymbol{u}_1\boldsymbol{u}_2^*\}_{i,j}.$$

For the expectation of the sum, we compute the following:

$$\begin{aligned}\mathrm{E}[\boldsymbol{h}\boldsymbol{v}_k^* \otimes \boldsymbol{h}\boldsymbol{v}_k^*] &= |\langle \boldsymbol{h}, \boldsymbol{b}_k\rangle|^2 \ \mathrm{E}[\{c_k[i]^* c_k[j] \boldsymbol{h}\boldsymbol{h}^*\}_{i,j}] \\ &= |\langle \boldsymbol{h}, \boldsymbol{b}_k\rangle|^2 \{\delta(i,j)\boldsymbol{h}\boldsymbol{h}^*\}_{i,j},\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}[\boldsymbol{u}_k\boldsymbol{m}^* \otimes \boldsymbol{u}_k\boldsymbol{m}^*] &= \mathrm{E}[|\langle \boldsymbol{m}, \boldsymbol{c}_k\rangle|^2] \{m[i]m[j](\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\}_{i,j} \\ &= \{m[i]m[j](\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\}_{i,j},\end{aligned}$$

since $\mathrm{E}[|\langle \boldsymbol{m}, \boldsymbol{c}_k\rangle|^2] = \|\boldsymbol{m}\|_2^2 = 1$, and

$$\begin{aligned}\mathrm{E}[\boldsymbol{h}\boldsymbol{v}_k^* \otimes \boldsymbol{u}_k\boldsymbol{m}^*] &= \mathrm{E}\{v_k[i]^* m[j] \boldsymbol{h}\boldsymbol{u}_k^*\}_{i,j} \\ &= \langle \boldsymbol{b}_k, \boldsymbol{h}\rangle \{\mathrm{E}[c_k[i]^* \langle \boldsymbol{c}_k, \boldsymbol{m}\rangle] m[j] \boldsymbol{h}\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\}_{i,j} \\ &= \langle \boldsymbol{b}_k, \boldsymbol{h}\rangle \{m[i]m[j]\boldsymbol{h}\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\}_{i,j},\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}[\boldsymbol{u}_k\boldsymbol{m}^* \otimes \boldsymbol{h}\boldsymbol{v}_k^*] &= \langle \boldsymbol{h}, \boldsymbol{b}_k\rangle \{\mathrm{E}[c_k[j]\langle \boldsymbol{m}, \boldsymbol{c}_k\rangle] m[i](\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{h}^*\}_{i,j} \\ &= \langle \boldsymbol{h}, \boldsymbol{b}_k\rangle \{m[i]m[j](\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{h}^*\}_{i,j}.\end{aligned}$$

A straightforward calculation combines these four results with (49) to verify that

$$\mathrm{E}[\mathcal{P}_T(\boldsymbol{A}_k) \otimes \mathcal{P}_T(\boldsymbol{A}_k)] = \mathcal{P}_T(\{\boldsymbol{b}_k\boldsymbol{b}_k^*\delta(i,j)\}_{i,j}\mathcal{P}_T).$$

In light of (25), this means

$$\mathrm{E}\left[\sum_{k\in\Gamma} \mathcal{P}_T(\boldsymbol{A}_k) \otimes \mathcal{P}_T(\boldsymbol{A}_k)\right] = \frac{Q}{L}\mathcal{P}_T + \mathcal{G}, \tag{51}$$

where $\|\mathcal{G}\| \le Q/4L$.

We now derive tail bounds for how far the sum over Γ for each of the terms in (50) deviates from their respective means. Starting with first term, we use the compact notation

$$\mathcal{Z}_k = \boldsymbol{h}\boldsymbol{v}_k^* \otimes \boldsymbol{h}\boldsymbol{v}_k^* - \mathrm{E}[\boldsymbol{h}\boldsymbol{v}_k^* \otimes \boldsymbol{h}\boldsymbol{v}_k^*],$$

for each addend. To apply Proposition 3, we need to uniformly bound the size (Orlicz ψ_1 norm) of each individual $\mathcal{Z}_k$ as well as the variance σ^2 in (43). For the uniform size bound,

$$\begin{aligned}\|\mathcal{Z}_k\| &= |\langle \boldsymbol{h}, \boldsymbol{b}_k\rangle|^2 \|\{(c_k[i]^* c_k[j] - \delta(i,j))\boldsymbol{h}\boldsymbol{h}^*\}_{i,j}\| \\ &= |\langle \boldsymbol{h}, \boldsymbol{b}_k\rangle|^2 \|\{(c_k[i]^* c_k[j] - \delta(i,j))\boldsymbol{I}\}\{\boldsymbol{h}\boldsymbol{h}^*\delta(i,j)\}_{i,j}\| \\ &\le |\langle \boldsymbol{h}, \boldsymbol{b}_k\rangle|^2 \|\boldsymbol{h}\boldsymbol{h}^*\| \|\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I}\| \\ &\le \frac{\mu_h^2}{L} \max(\|\boldsymbol{c}_k\|_2^2, 1).\end{aligned}$$

Applying Lemma 7,

$$\mathrm{P}\left\{\max(\|\boldsymbol{c}_k\|_2^2, 1) > u\right\} \ \le \ 1.2\,\mathrm{e}^{-u/8N},$$

and combined with Lemma 5 this means

$$\|\mathcal{Z}_k\|_{\psi_1} \;\leq\; \frac{\mu_h^2}{L}\|\max(\|\boldsymbol{c}_k\|_2^2,1)\|_{\psi_1} \;\leq\; C\,\frac{\mu_h^2 N}{L}.$$

For the variance, we need to compute $\mathrm{E}[\mathcal{Z}_k^*\mathcal{Z}_k]$. This will be easiest if we rewrite the action of $\mathcal{Z}_k$ on a matrix $\boldsymbol{X}$ as

$$\mathcal{Z}_k(\boldsymbol{X}) = |\langle \boldsymbol{h},\boldsymbol{b}_k\rangle|^4 \boldsymbol{h}\boldsymbol{h}^*\boldsymbol{X}(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I}),$$

and so

$$\mathcal{Z}_k^*\mathcal{Z}_k(\boldsymbol{X}) = |\langle \boldsymbol{h},\boldsymbol{b}_k\rangle|^4 \|\boldsymbol{h}\|_2^2\boldsymbol{h}\boldsymbol{h}^*\boldsymbol{X}(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})^2,$$

and

$$\begin{aligned}\mathrm{E}[\mathcal{Z}_k^*\mathcal{Z}_k(\boldsymbol{X})] &= |\langle \boldsymbol{h},\boldsymbol{b}_k\rangle|^4 \|\boldsymbol{h}\|_2^2\boldsymbol{h}\boldsymbol{h}^*\boldsymbol{X}\,\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})^2] \\ &= N|\langle \boldsymbol{h},\boldsymbol{b}_k\rangle|^4 \boldsymbol{h}\boldsymbol{h}^*\boldsymbol{X},\end{aligned}$$

and finally

$$\begin{aligned}\left\|\sum_{k\in\Gamma}\mathrm{E}[\mathcal{Z}_k^*\mathcal{Z}_k]\right\| &= N\sum_{k\in\Gamma}|\langle \boldsymbol{h},\boldsymbol{b}_k\rangle|^4 \\ &\leq \frac{\mu_h^2 N}{L}\sum_{k\in\Gamma}|\langle \boldsymbol{h},\boldsymbol{b}_k\rangle|^2 \\ &\leq \frac{5\mu_h^2 NQ}{4L^2},\end{aligned}$$

where we have used (25) in the last step. Collecting these results and applying Proposition 3 with $t = \alpha\log(KN)$ yields

$$\left\|\sum_{k\in\Gamma}\boldsymbol{h}\boldsymbol{v}_k^*\otimes\boldsymbol{h}\boldsymbol{v}_k^* - \mathrm{E}[\boldsymbol{h}\boldsymbol{v}_k^*\otimes\boldsymbol{h}\boldsymbol{v}_k^*]\right\| \leq C_\alpha\,\frac{\mu_h\sqrt{N\log(KN)}}{L}\max\left\{\sqrt{Q},\mu_h\sqrt{N\log(KN)}\log(\mu_h^2 N)\right\}, \tag{52}$$

with probability exceeding $1-(KN)^{-\alpha}$.

For the sum over the second term in (50), set

$$\begin{aligned}\mathcal{Z}_k &= \boldsymbol{u}_k\boldsymbol{m}^*\otimes\boldsymbol{u}_k\boldsymbol{m}^* - \mathrm{E}[\boldsymbol{u}_k\boldsymbol{m}^*\otimes\boldsymbol{u}_k\boldsymbol{m}^*] \\ &= \left(|\langle\boldsymbol{m},\boldsymbol{c}_k\rangle|^2-1\right)\{m[i]m[j](\boldsymbol{I}-\boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{b}_k^*(\boldsymbol{I}-\boldsymbol{h}\boldsymbol{h}^*)\}_{i,j},\end{aligned}$$

then using the fact that $\|\boldsymbol{I}-\boldsymbol{h}\boldsymbol{h}^*\|\leq 1$ (since $\|\boldsymbol{h}\|_2 = 1$), we have

$$\begin{aligned}\|\mathcal{Z}_k\| &= \left|\,|\langle\boldsymbol{m},\boldsymbol{c}_k\rangle|^2-1\right|\,\|(\boldsymbol{I}-\boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\|_2^2\,\|\boldsymbol{m}\|_2^2 \\ &\leq \left|\,|\langle\boldsymbol{m},\boldsymbol{c}_k\rangle|^2-1\right|\,\|\boldsymbol{b}_k\|_2^2 \\ &\leq \left|\,|\langle\boldsymbol{m},\boldsymbol{c}_k\rangle|^2-1\right|\,\frac{\mu_1^2 K}{L}.\end{aligned}$$

This is again a subexponential random variable whose size we can characterize using Lemma 9:

$$\||\langle\boldsymbol{m},\boldsymbol{c}_k\rangle|^2-1\|_{\psi_1}\leq C \quad \text{and so} \quad \|\mathcal{Z}_k\|_{\psi_1} \;\leq\; C\,\frac{\mu_1^2 K}{L}.$$

To bound the variance in (44), we again write out the action of $\mathcal{Z}_k$ on an arbitrary $K \times N$ matrix $\boldsymbol{X}$:

$$\mathcal{Z}_k(\boldsymbol{X}) = (|\langle \boldsymbol{m}, \boldsymbol{c}_k\rangle|^2 - 1)(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{X}\boldsymbol{m}\boldsymbol{m}^*,$$

and so

$$\begin{aligned}
\mathrm{E}[\mathcal{Z}_k^*\mathcal{Z}_k(\boldsymbol{X})] &= \mathrm{E}[(|\langle \boldsymbol{m}, \boldsymbol{c}_k\rangle|^2 - 1)^2]\|(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\|_2^2(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{X}\boldsymbol{m}\boldsymbol{m}^* \\
&= \|(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\|_2^2(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{X}\boldsymbol{m}\boldsymbol{m}^*,
\end{aligned}$$

where in the last step we have used the fact that $|\langle \boldsymbol{m}, \boldsymbol{c}_k\rangle|^2$ is a chi-square random variable with two degrees of freedom with variance $\mathrm{E}[(|\langle \boldsymbol{m}, \boldsymbol{c}_k\rangle|^2 - 1)^2] = 1$. This gives us

$$\begin{aligned}
\left\|\sum_{k\in\Gamma}\mathrm{E}[\mathcal{Z}_k^*\mathcal{Z}_k]\right\| &= \left\|\sum_{k\in\Gamma}\|(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\|_2^2(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\right\| \\
&\le \max_{k\in\Gamma}\left(\|(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\|_2^2\right)\left\|\sum_{k\in\Gamma}(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{b}_k^*(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\right\| \\
&\le \frac{\mu_1^2 K}{L}\left\|\sum_{k\in\Gamma}\boldsymbol{b}_k\boldsymbol{b}_k^*\right\| \\
&\le \frac{5\mu_1^2 KQ}{4L^2}.
\end{aligned}$$

Collecting these results and applying Proposition 3 with $t = \alpha\log(KN)$ yields

$$\left\|\sum_{k\in\Gamma}\boldsymbol{u}_k\boldsymbol{m}^*\otimes\boldsymbol{u}_k\boldsymbol{m}^* - \mathrm{E}[\boldsymbol{u}_k\boldsymbol{m}^*\otimes\boldsymbol{u}_k\boldsymbol{m}^*]\right\| \le C_\alpha\frac{\mu_1\sqrt{K\log(KN)}}{L}\max\left\{\sqrt{Q}, \mu_1\sqrt{K\log(KN)}\log(\mu_1^2 K)\right\}, \tag{53}$$

with probability exceeding $1 - (KN)^{-\alpha}$.

The last two terms in (50) are adjoints of one another, so they will have the same operator norm. We now set

$$\begin{aligned}
\mathcal{Z}_k &= \boldsymbol{h}\boldsymbol{v}_k^*\otimes\boldsymbol{u}_k\boldsymbol{m}^* - \mathrm{E}[\boldsymbol{h}\boldsymbol{v}_k^*\otimes\boldsymbol{u}_k\boldsymbol{m}^*] \\
&= \langle \boldsymbol{h}, \boldsymbol{b}_k\rangle\{m[i](c_k[j]\langle \boldsymbol{m}, \boldsymbol{c}_k\rangle - m[j])(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{h}^*\}_{i,j},
\end{aligned}$$

and so the action of $\mathcal{Z}_k$ on an arbitrary matrix $\boldsymbol{X}$ is given by

$$\mathcal{Z}_k(\boldsymbol{X}) = \langle \boldsymbol{h}, \boldsymbol{b}_k\rangle(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\boldsymbol{b}_k\boldsymbol{h}^*\boldsymbol{X}(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{m}\boldsymbol{m}^*,$$

from which we can see

$$\begin{aligned}
\|\mathcal{Z}_k\| &\le |\langle \boldsymbol{h}, \boldsymbol{b}_k\rangle|\,\|\boldsymbol{b}_k\|_2\|(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{m}\|_2 \\
&\le \frac{\mu_h\mu_1\sqrt{K}}{L}\|(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{m}\|_2.
\end{aligned}$$

From Lemmas 10 and 5, we that the random variable $\|(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{m}\|_2$ is subexponential with $\|(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{m}\|_{\psi_1} \le C\sqrt{N}$, and so

$$\|\mathcal{Z}_k\|_{\psi_1} \;\le\; C\frac{\mu_h\mu_1\sqrt{KN}}{L}.$$

For the variance σ^2 in (43), we need to bound the sizes of both $\mathcal{Z}_k^* \mathcal{Z}_k$ and $\mathcal{Z}_k \mathcal{Z}_k^*$. Starting with the former, we have

$$\mathrm{E}[\mathcal{Z}_k^* \mathcal{Z}_k(\boldsymbol{X})] = |\langle \boldsymbol{h}, \boldsymbol{b}_k \rangle|^2 \|(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*)\|_2^2 \boldsymbol{h}\boldsymbol{h}^* \boldsymbol{X} \, \mathrm{E}[(\boldsymbol{c}_k \boldsymbol{c}_k^* - \boldsymbol{I}) \boldsymbol{m}\boldsymbol{m}^* (\boldsymbol{c}_k \boldsymbol{c}_k^* - \boldsymbol{I})],$$

and then applying Lemma 11 yields

$$\begin{aligned} \left\| \sum_{k \in \Gamma} \mathrm{E}[\mathcal{Z}_k^* \mathcal{Z}_k] \right\| &= \sum_{k \in \Gamma} |\langle \boldsymbol{h}, \boldsymbol{b}_k \rangle|^2 \, \|(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_k\|_2^2 \\ &\le \sum_{k \in \Gamma} |\langle \boldsymbol{h}, \boldsymbol{b}_k \rangle|^2 \, \|\boldsymbol{b}_k\|_2^2 \\ &\le \frac{\mu_1^2 K}{L} \sum_{k \in \Gamma} |\langle \boldsymbol{h}, \boldsymbol{b}_k \rangle|^2 \\ &\le \frac{5 \mu_1^2 K Q}{4 L^2}. \end{aligned}$$

For $\mathcal{Z}_k \mathcal{Z}_k^*$,

$$\mathrm{E}[\mathcal{Z}_k \mathcal{Z}_k^*(\boldsymbol{X})] = |\langle \boldsymbol{h}, \boldsymbol{b}_k \rangle|^2 (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_k \boldsymbol{b}_k^* (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{X} \boldsymbol{m}\boldsymbol{m}^* \, \mathrm{E}[(\boldsymbol{c}_k \boldsymbol{c}_k^* - \boldsymbol{I})^2] \boldsymbol{m}\boldsymbol{m}^*,$$

and then applying Lemma 8 yields

$$\begin{aligned} \left\| \sum_{k \in \Gamma} \mathrm{E}[\mathcal{Z}_k \mathcal{Z}_k^*] \right\| &= N \left\| (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \left(\sum_{k \in \Gamma} |\langle \boldsymbol{h}, \boldsymbol{b}_k \rangle|^2 \boldsymbol{b}_k \boldsymbol{b}_k^* \right) (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \right\| \\ &\le N \left\| \sum_{k \in \Gamma} |\langle \boldsymbol{h}, \boldsymbol{b}_k \rangle|^2 \boldsymbol{b}_k \boldsymbol{b}_k^* \right\| \\ &\le \frac{\mu_h^2 N}{L} \left\| \sum_{k \in \Gamma} \boldsymbol{b}_k \boldsymbol{b}_k^* \right\| \\ &\le \frac{5 \mu_h^2 N Q}{4 L^2}. \end{aligned}$$

Collecting these results and applying Proposition 3 with $t = \alpha \log(KN)$ and $M = \max\left\{\mu_1^2 K, \mu_h^2 N\right\}$ yields

$$\left\| \sum_{k \in \Gamma} \boldsymbol{h}\boldsymbol{v}_k^* \otimes \boldsymbol{u}_k \boldsymbol{m}^* - \mathrm{E}[\boldsymbol{h}\boldsymbol{v}_k^* \otimes \boldsymbol{u}_k \boldsymbol{m}^*] \right\| \le C_\alpha \, \frac{\sqrt{M \log(NK)}}{L} \max\left\{ \sqrt{Q}, \sqrt{M \log(NK)} \log(M) \right\}, \tag{54}$$

with probability exceeding $1 - (KN)^{-\alpha}$.

Using the triangle inequality

$$\left\| \mathcal{P}_T \mathcal{A}_p^* \mathcal{A}_p \mathcal{P}_T - \frac{Q}{L} \mathcal{P}_T \right\| \le \left\| \mathcal{P}_T \mathcal{A}_p^* \mathcal{A}_p \mathcal{P}_T - \mathrm{E}[\mathcal{P}_T \mathcal{A}_p^* \mathcal{A}_p \mathcal{P}_T] \right\| + \left\| \mathrm{E}[\mathcal{P}_T \mathcal{A}_p^* \mathcal{A}_p \mathcal{P}_T] - \frac{Q}{L} \mathcal{P}_T \right\|,$$

we can combine (51) with (52), (53), and (54) to establish that

$$\left\| \mathcal{P}_T \mathcal{A}_p^* \mathcal{A}_p \mathcal{P}_T - \frac{Q}{L} \mathcal{P}_T \right\| \le C_\alpha \, \frac{\sqrt{M \log(NK)}}{L} \max\left\{ \sqrt{Q}, \sqrt{M \log(NK)} \log(M) \right\} + \frac{Q}{4L},$$

with probability exceeding $1-3(KN)^{-\alpha}$. With Q chosen as in (37), this becomes

$$\begin{aligned}\left\|\mathcal{P}_T\mathcal{A}_p^*\mathcal{A}_p\mathcal{P}_T-\frac{Q}{L}\mathcal{P}_T\right\| &\leq C_\alpha\,\frac{Q}{L}\,\max\left\{\frac{1}{\sqrt{C'_\alpha\log M}},\frac{1}{C'_\alpha}\right\}+\frac{Q}{4L}\\ &\leq \frac{Q}{2L},\end{aligned}$$

for C'_α chosen appropriately. Applying the union bound establishes the lemma.

To prove the corollary, we take $\Gamma=\{1,\ldots,L\}$ and $Q=L$ above. In this case, we will have $\sum_{k\in\Gamma}\boldsymbol{b}_k\boldsymbol{b}_k^*=\boldsymbol{I}$, and so $\mathcal{G}=0$ in (51). We have

$$\|\mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T-\mathcal{P}_T\|\leq C_\alpha\,\max\left\{\sqrt{\frac{M\log(NK)}{L}},\frac{M\log(NK)\log(M)}{L}\right\},$$

with probability exceeding $1-3(KN)^{-\alpha}$. Then taking L as in (38) will guarantee the desired conditioning.

5.3 Proof of Lemma 3

We start by fixing $\ell\in\Gamma_{p+1}$ and estimating $\|\boldsymbol{W}_p^*\boldsymbol{b}_\ell\|_2$. We can re-write $\boldsymbol{W}_p$ as a sum of independent random matrices: since $\boldsymbol{W}_{p-1}\in T$, $\mathcal{P}_T(\boldsymbol{W}_{p-1})=\boldsymbol{W}_{p-1}$ and

$$\begin{aligned}\boldsymbol{W}_p &= \mathcal{P}_T\left(\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}-\frac{Q}{L}\boldsymbol{W}_{p-1}\right)\\ &= \mathcal{P}_T\left(\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}\boldsymbol{c}_k\boldsymbol{c}_k^*-\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}\right)+\mathcal{P}_T\left(\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}-\frac{Q}{L}\boldsymbol{W}_{p-1}\right)\\ &= \sum_{k\in\Gamma_p}\mathcal{P}_T(\boldsymbol{Z}_k)+\mathcal{P}_T\left(\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}-\frac{Q}{L}\boldsymbol{W}_{p-1}\right),\end{aligned}$$

where $\boldsymbol{Z}_k=\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}(\boldsymbol{c}_k\boldsymbol{c}_k^*-\boldsymbol{I})$. Then

$$\|\boldsymbol{W}_p^*\boldsymbol{b}_\ell\|_2\leq\left\|\sum_{k\in\Gamma_p}\boldsymbol{b}_\ell^*\mathcal{P}_T(\boldsymbol{Z}_k)\right\|_2+\left\|\boldsymbol{b}_\ell^*\mathcal{P}_T\left(\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}-\frac{Q}{L}\boldsymbol{W}_{p-1}\right)\right\|_2. \tag{55}$$

For the second term above

$$\begin{aligned}\left\|\boldsymbol{b}_\ell^*\mathcal{P}_T\left(\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}-\frac{Q}{L}\boldsymbol{W}_{p-1}\right)\right\|_2 &\leq \left\|\mathcal{P}_T\left(\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}-\frac{Q}{L}\boldsymbol{W}_{p-1}\right)\right\|\,\|\boldsymbol{b}_\ell\|_2\\ &\leq \sqrt{\frac{\mu_1^2K}{L}}\left\|\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}-\frac{Q}{L}\boldsymbol{W}_{p-1}\right\|\\ &\leq \sqrt{\frac{\mu_1^2K}{L}}\left\|\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*-\frac{Q}{L}\boldsymbol{I}\right\|\,\|\boldsymbol{W}_{p-1}\|_F\\ &\leq \frac{2^{-p+1}\mu_1\sqrt{K}Q}{4L^{3/2}},\end{aligned}$$

where we have used (25) and the fact that the Frobenius norms of the $\boldsymbol{W}_p$ decrease geometrically with p; see (28).

The first term in (55) is the norm of a sum of independent zero-mean random vectors, which we will bound using Propositions 2 and 3. We set $\boldsymbol{w}_k = \boldsymbol{W}_{p-1}^* \boldsymbol{b}_k$ and expand $\boldsymbol{b}_\ell^* \mathcal{P}_T(\boldsymbol{Z}_k)$ as

$$\boldsymbol{b}_\ell^* \mathcal{P}_T(\boldsymbol{Z}_k) = \langle \boldsymbol{h}, \boldsymbol{b}_\ell \rangle \langle \boldsymbol{b}_k, \boldsymbol{h} \rangle \boldsymbol{w}_k^* (\boldsymbol{c}_k \boldsymbol{c}_k^* - \boldsymbol{I}) + \langle \boldsymbol{b}_k, \boldsymbol{b}_\ell \rangle \boldsymbol{w}_k^* (\boldsymbol{c}_k \boldsymbol{c}_k^* - \boldsymbol{I}) \boldsymbol{m}\boldsymbol{m}^* - \langle \boldsymbol{h}, \boldsymbol{b}_\ell \rangle \langle \boldsymbol{b}_k, \boldsymbol{h} \rangle \boldsymbol{w}_k^* (\boldsymbol{c}_k \boldsymbol{c}_k^* - \boldsymbol{I}) \boldsymbol{m}\boldsymbol{m}^*,$$

and so

$$\left\| \sum_{k \in \Gamma_p} \boldsymbol{b}_\ell^* \mathcal{P}_T(\boldsymbol{Z}_k) \right\|_2 \leq \left\| \sum_{k \in \Gamma_p} \boldsymbol{z}_k \right\|_2 + \left| \sum_{k \in \Gamma_p} z_k \right|, \tag{56}$$

where the $\boldsymbol{z}_k$ are independent random vectors, and the z_k are independent random scalars:

$$\boldsymbol{z}_k = \langle \boldsymbol{b}_\ell, \boldsymbol{h} \rangle \langle \boldsymbol{h}, \boldsymbol{b}_k \rangle (\boldsymbol{c}_k \boldsymbol{c}_k^* - \boldsymbol{I}) \boldsymbol{w}_k, \qquad z_k = \langle \boldsymbol{b}_k, (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_\ell \rangle \, \langle (\boldsymbol{c}_k \boldsymbol{c}_k^* - \boldsymbol{I}) \boldsymbol{m}, \boldsymbol{w}_k \rangle.$$

Using Lemma 12, we have a tail bound for each term in the scalar sum:

$$\mathrm{P}\left\{ |z_k| > \lambda \right\} \leq 2\mathrm{e} \cdot \exp\left(- \frac{\lambda}{\|\boldsymbol{w}_k\|_2 |\langle \boldsymbol{b}_k, (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_\ell \rangle|} \right).$$

Applying the scalar Bernstein inequality (Proposition 2) with

$$B = \max_k \|\boldsymbol{w}_k\|_2 |\langle \boldsymbol{b}_k, (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_\ell \rangle| \;\leq\; \frac{\mu_{p-1}\mu_1^2 K}{L^{3/2}},$$

and

$$\begin{aligned}
\sigma^2 &= \sum_{k \in \Gamma_p} \|\boldsymbol{w}_k\|_2^2 |\langle \boldsymbol{b}_k, (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_\ell \rangle|^2 \\
&\leq \frac{\mu_{p-1}^2}{L} \sum_{k \in \Gamma_p} |\langle \boldsymbol{b}_k, (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_\ell \rangle|^2 \\
&\leq \frac{5\mu_{p-1}^2 Q}{4L^2} \|(\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^*) \boldsymbol{b}_\ell\|_2^2 \\
&\leq \frac{5\mu_{p-1}^2 \mu_1^2 K Q}{4L^3},
\end{aligned}$$

and $t = \alpha \log(KN)$ tells us that

$$\left| \sum_{k \in \Gamma_p} z_k \right| \leq C_\alpha \, \frac{\mu_{p-1}\mu_1 \sqrt{K \log(KN)}}{L^{3/2}} \max\left\{ \sqrt{Q}, \mu_1 \sqrt{K \log(KN)} \right\}, \tag{57}$$

with probability at least $1 - (KN)^{-\alpha}$.

For the vector term in (56), we apply Lemmas 10 and 5 to see that

$$\begin{aligned}
\|\boldsymbol{z}_k\|_{\psi_1} &\leq C\sqrt{N} \|\boldsymbol{w}_k\|_2 |\langle \boldsymbol{b}_\ell, \boldsymbol{h} \rangle \langle \boldsymbol{h}, \boldsymbol{b}_k \rangle| \\
&\leq C \frac{\mu_{p-1}\mu_h^2 \sqrt{N}}{L^{3/2}}.
\end{aligned}$$

For the variance terms, we calculate

$$\begin{aligned}\sum_{k\in\Gamma_p} \mathrm{E}[\boldsymbol{z}_k^*\boldsymbol{z}_k] &= \sum_{k\in\Gamma_p} |\langle \boldsymbol{h},\boldsymbol{b}_\ell\rangle|^2 |\langle \boldsymbol{b}_k,\boldsymbol{h}\rangle|^2 \boldsymbol{w}_k^* \,\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})^2]\boldsymbol{w}_k \\ &= N \sum_{k\in\Gamma_p} |\langle \boldsymbol{h},\boldsymbol{b}_\ell\rangle|^2 |\langle \boldsymbol{b}_k,\boldsymbol{h}\rangle|^2 \|\boldsymbol{w}_k\|_2^2 \qquad \text{(by Lemma 8)} \\ &\le \frac{\mu_{p-1}^2\mu_h^2 N}{L^2} \sum_{k\in\Gamma_p} |\langle \boldsymbol{b}_k,\boldsymbol{h}\rangle|^2 \\ &\le \frac{5\mu_{p-1}^2\mu_h^2 NQ}{4L^3},\end{aligned}$$

and

$$\begin{aligned}\left\|\sum_{k\in\Gamma_p} \mathrm{E}[\boldsymbol{z}_k\boldsymbol{z}_k^*]\right\| &= \left\|\sum_{k\in\Gamma_p} |\langle \boldsymbol{h},\boldsymbol{b}_\ell\rangle|^2 |\langle \boldsymbol{b}_k,\boldsymbol{h}\rangle|^2 \,\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{w}_k\boldsymbol{w}_k^*(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})]\right\| \\ &= \left\|\sum_{k\in\Gamma_p} |\langle \boldsymbol{h},\boldsymbol{b}_\ell\rangle|^2 |\langle \boldsymbol{b}_k,\boldsymbol{h}\rangle|^2 \|\boldsymbol{w}_k\|_2^2 \boldsymbol{I}\right\| \qquad \text{(by Lemma 11)} \\ &\le \frac{\mu_{p-1}^2\mu_h^2}{L^2} \sum_{k\in\Gamma_p} |\langle \boldsymbol{b}_k,\boldsymbol{h}\rangle|^2 \\ &\le \frac{5\mu_{p-1}^2\mu_h^2 Q}{4L^3}.\end{aligned}$$

Thus

$$\left\|\sum_{k\in\Gamma_p} \boldsymbol{z}_k\right\|_2 \le C_\alpha \,\frac{\mu_{p-1}\mu_h\sqrt{N\log(KN)}}{L^{3/2}} \max\left\{\sqrt{Q}, \mu_h\log(\mu_h)\sqrt{\log(KN)}\right\} \tag{58}$$

with probability at least $1-(KN)^{-\alpha}$.

Combining (57) and (58) and taking the union bound over all $\ell \in \Gamma_{p+1}$ yields

$$\mu_p \le \mu_{p-1}\,\frac{C_\alpha\sqrt{MQ}}{L} = \mu_{p-1}\,\frac{C_\alpha M\log(KN)\sqrt{\beta\log M}}{L}.$$

with probability exceeding $1-2Q(KN)^{-\alpha}$. Then taking the union bound over $1\le p\le P$ establishes the lemma.

5.4 Proof of Lemma 4

We start by fixing p and writing

$$\left\|\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \frac{Q}{L}\boldsymbol{W}_{p-1}\right\| \;\le\; \left\|\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \mathrm{E}[\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}]\right\| + \left\|\mathrm{E}[\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}] - \frac{Q}{L}\boldsymbol{W}_{p-1}\right\|.$$

We will derive a concentration inequality to bound the first term, and use (25) for the second. We can write the first term above as the spectral norm of a sum of random rank-1 matrices:

$$\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \mathrm{E}[\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}] = \sum_{k\in\Gamma_p}\boldsymbol{Z}_k, \qquad \boldsymbol{Z}_k := \boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_{p-1}(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I}). \tag{59}$$

We will use Proposition 3 to estimate the size of this random sum; we proceed by calculating the key quantities involved. With $\boldsymbol{w}_k = \boldsymbol{W}_{p-1}^*\boldsymbol{b}_k$, we can bound the size of each term in the sum as

$$\begin{aligned}\|\boldsymbol{Z}_k\| &= \|\boldsymbol{b}_k\boldsymbol{b}_k^*\boldsymbol{W}_p(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\| \\ &= \|\boldsymbol{b}_k\|_2\,\|(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{w}_k\|_2 \\ &\le \mu_1\sqrt{\frac{K}{L}}\,\|(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{w}_k\|_2\end{aligned}$$

and then applying Lemmas 10 and 5 yields

$$\|\boldsymbol{Z}_k\|_{\psi_1} \le C\,\mu_1\sqrt{\frac{KN}{L}}\,\|\boldsymbol{w}_k\|_2 \;\le\; C\,\mu_1\mu_p\sqrt{\frac{KN}{L}}.$$

For the variance terms, we calculate

$$\begin{aligned}\left\|\sum_{k\in\Gamma_p}\mathrm{E}[\boldsymbol{Z}_k^*\boldsymbol{Z}_k]\right\| &= \left\|\sum_{k\in\Gamma_p}\|\boldsymbol{b}_k\|_2^2\,\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{w}_k\boldsymbol{w}_k^*(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})]\right\| \\ &= \sum_{k\in\Gamma_p}\|\boldsymbol{b}_k\|_2^2\|\boldsymbol{w}_k\|_2^2 \qquad \text{(by Lemma 11)} \\ &\le \frac{\mu_1^2K}{L}\sum_{k\in\Gamma_p}\|\boldsymbol{W}_p^*\boldsymbol{b}_k\|_2^2 \\ &\le \frac{5\mu_1^2KQ}{4L^2}\|\boldsymbol{W}_p\|_F^2 \qquad \text{(using (25))},\end{aligned}$$

and

$$\begin{aligned}\left\|\sum_{k\in\Gamma_p}\mathrm{E}[\boldsymbol{Z}_k\boldsymbol{Z}_k^*]\right\| &= \left\|\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{w}_k^*\,\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})^2]\boldsymbol{w}_k\boldsymbol{b}_k^*\right\| \\ &= N\left\|\sum_{k\in\Gamma_p}\|\boldsymbol{w}_k\|_2^2\boldsymbol{b}_k\boldsymbol{b}_k^*\right\| \qquad \text{(by Lemma 8)} \\ &\le \frac{\mu_p^2N}{L}\left\|\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^*\right\| \\ &\le \frac{5\mu_p^2NQ}{4L^2}.\end{aligned}$$

Then with $M = \max\left\{\mu_1^2K, \mu_h^2N\right\}$, we apply Proposition 3 with $t = \alpha\log(KN)$ to get

$$\|\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \mathrm{E}[\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}]\| \le C_\alpha\,2^{-p}\,\frac{\sqrt{M\log(KN)}}{L}\max\left\{\sqrt{Q}, \sqrt{M\log(KN)}\log(M)\right\},$$

with probability exceeding $1 - (KN)^{-\alpha}$. With Q as in (37), this becomes

$$\begin{aligned}\|\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \mathrm{E}[\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}]\| &\le C_\alpha\,2^{-p}\,\frac{Q}{L}\max\left\{\frac{1}{\sqrt{C'_\alpha\log M}}, \frac{1}{C'_\alpha}\right\} \\ &\le 2^{-p}\frac{Q}{4L},\end{aligned}$$

for an appropriate choice of C'_α. Thus

$$\left\|\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \frac{Q}{L}\boldsymbol{W}_{p-1}\right\| \leq \left\|\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1} - \mathrm{E}[\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}]\right\| + \left\|\mathrm{E}[\mathcal{A}_p^*\mathcal{A}_p\boldsymbol{W}_{p-1}] - \frac{Q}{L}\boldsymbol{W}_{p-1}\right\|$$
$$\leq 2^{-p}\frac{Q}{4L} + \left\|\sum_{k\in\Gamma_p}\boldsymbol{b}_k\boldsymbol{b}_k^* - \frac{Q}{L}\right\| \|\boldsymbol{W}_{p-1}\|_F$$
$$\leq 2^{-p}\frac{Q}{4L} + 2^{-p+1}\frac{Q}{4L}.$$

Applying the union bound over all $p = 1, \dots, P$ establishes the lemma.

6 Supporting Lemmas

Lemma 6. *Let $\boldsymbol{c}_k \in \mathbb{C}^N$ be normally distributed as in* (15)*, and let $\boldsymbol{u} \in \mathbb{C}^N$ be an arbitrary vector. Then $|\langle \boldsymbol{c}_k, \boldsymbol{u}\rangle|^2$ is a chi-square random variable with two degrees of freedom and*

$$\mathrm{P}\left\{|\langle \boldsymbol{c}_k, \boldsymbol{u}\rangle|^2 > \lambda\right\} \;\leq\; \mathrm{e}^{-\lambda/\|\boldsymbol{u}\|_2^2}.$$

Lemma 7. *Let $\boldsymbol{c}_k \in \mathbb{C}^N$ be normally distributed as in* (15)*. Then*

$$\mathrm{P}\left\{\|\boldsymbol{c}_k\|_2^2 > Nu\right\} \;\leq\; 1.2\,\mathrm{e}^{-u/8}, \qquad \text{for all } u \geq 0, \tag{60}$$

and since $1.2\mathrm{e}^{-1/8N} \geq 1$ *for all* $N \geq 1$*,*

$$\mathrm{P}\left\{\max(\|\boldsymbol{c}_k\|_2^2, 1) > Nu\right\} \;\leq\; 1.2\,\mathrm{e}^{-u/8}.$$

Proof. It is well-known (see, for example, [39]) that

$$\mathrm{P}\left\{\|\boldsymbol{c}_k\|_2^2 > N(1+\lambda)\right\} \;\leq\; \begin{cases}\mathrm{e}^{-\lambda^2/8} & 0 \leq \lambda \leq 1\\ \mathrm{e}^{-\lambda/8} & \lambda \geq 1\end{cases} \;\leq\; 1.05\,\mathrm{e}^{-\lambda/8}, \quad \lambda \geq 0. \tag{61}$$

Plugging in $\lambda = u - 1$ above yields

$$\mathrm{P}\left\{\|\boldsymbol{c}_k\|_2^2 > Nu\right\} \;\leq\; 1.2\,\mathrm{e}^{-u/8}, \quad u \geq 1.$$

Since $1.2\,\mathrm{e}^{-1/8} > 1$, the bound above can be extended for all $u \geq 0$. □

Lemma 8. *Let $\boldsymbol{c}_k \in \mathbb{C}^N$ be normally distributed as in* (15)*. Then*

$$\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})^2] = N\boldsymbol{I}.$$

Proof. Using the expansion

$$(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})^2 = \|\boldsymbol{c}_k\|_2^2\boldsymbol{c}_k\boldsymbol{c}_k^* - 2\boldsymbol{c}_k\boldsymbol{c}_k^* + \boldsymbol{I},$$

we see that the only non-trivial term is $\boldsymbol{R} = \|\boldsymbol{c}_k\|_2^2\boldsymbol{c}_k\boldsymbol{c}_k^*$. We compute the expectation of an entry in this matrix as

$$\mathrm{E}[R(i,j)] = \sum_{n=1}^N \mathrm{E}[|c_k[n]|^2 c_k[i]c_k[j]^*] = \begin{cases}\sum_n \mathrm{E}[|c_k[n]|^2|c_k[i]|^2] & i = j\\ 0 & i \neq j\end{cases}.$$

For the addends in the diagonal term

$$\mathrm{E}[|c_k[n]|^2|c_k[i]|^2] = \begin{cases} \mathrm{E}[|c_k[n]|^4] = 2 & n = i \\ 1 & n \neq i \end{cases},$$

where the calculation for $n = i$ relies on the fact that $\mathrm{E}[|c_k[n]|^4]$ is the second moment of a chi-square random variable with two degrees of freedom. Thus $\mathrm{E}[\boldsymbol{R}] = (N+1)\boldsymbol{I}$, and

$$\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})^2] = (N+1)\boldsymbol{I} - 2\boldsymbol{I} + \boldsymbol{I} = N\boldsymbol{I}.$$

□

Lemma 9. *Let $\boldsymbol{c}_k \in \mathbb{C}^N$ be normally distributed as in* (15)*, and let $\boldsymbol{v}$ be an arbitrary vector. Then $\mathrm{E}[|\langle \boldsymbol{c}_k, \boldsymbol{v}\rangle|^2] = \|\boldsymbol{v}\|_2^2$ and*

$$\mathrm{P}\left\{\left||\langle \boldsymbol{c}_k, \boldsymbol{v}\rangle|^2 - \|\boldsymbol{v}\|_2^2\right| > \lambda\right\} \;\leq\; 2.1\,\exp\left(-\frac{\lambda}{8\|\boldsymbol{v}\|_2^2}\right).$$

Proof. A slight variation of (61) gives us that

$$\mathrm{P}\left\{\left||\langle \boldsymbol{c}_k, \boldsymbol{v}\rangle|^2 - \|\boldsymbol{v}\|_2^2\right| > \lambda\right\} \;\leq\; \begin{cases} 2\mathrm{e}^{-\lambda^2/8\|\boldsymbol{v}\|_2^2} & 0 \leq \lambda \leq 1 \\ \mathrm{e}^{-\lambda/8\|\boldsymbol{v}\|_2^2} & \lambda > 1 \end{cases}.$$

The lemma follows from combining these two cases into one subexponential bound. □

Lemma 10. *Let $\boldsymbol{c}_k \in \mathbb{C}^N$ be normally distributed as in* (15)*, and let $\boldsymbol{v} \in \mathbb{C}^N$ be an arbitrary vector. Then*

$$\mathrm{P}\left\{\|(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{v}\|_2 > \lambda\right\} \;\leq\; 3\exp\left(-\frac{\lambda}{\sqrt{8N}\|v\|_2}\right).$$

Proof. We have

$$\|(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{v}\|_2 = \|\langle \boldsymbol{v}, \boldsymbol{c}_k\rangle \boldsymbol{c}_k - \boldsymbol{v}\|_2 \leq |\langle \boldsymbol{v}, \boldsymbol{c}_k\rangle| \|\boldsymbol{c}_k\|_2 + \|\boldsymbol{v}\|_2.$$

For the first term above, we have for any $\tau > 0$,

$$\begin{aligned}\mathrm{P}\left\{|\langle \boldsymbol{v}, \boldsymbol{c}_k\rangle|\,\|\boldsymbol{c}_k\|_2 > \lambda\sqrt{N}\|\boldsymbol{v}\|_2\right\} &\leq \mathrm{P}\left\{|\langle \boldsymbol{v}, \boldsymbol{c}_k\rangle| > \sqrt{\lambda}\|\boldsymbol{v}\|_2/\tau\right\} + \mathrm{P}\left\{\|\boldsymbol{c}_k\|_2 > \tau\sqrt{\lambda N}\right\} \\ &= \mathrm{P}\left\{|\langle \boldsymbol{v}, \boldsymbol{c}_k\rangle|^2 > \lambda\|\boldsymbol{v}\|_2^2/\tau^2\right\} + \mathrm{P}\left\{\|\boldsymbol{c}_k\|_2^2 > \tau^2\lambda N\right\}\end{aligned}$$

We can then use the fact that $|\langle \boldsymbol{v}, \boldsymbol{c}_k\rangle|^2$ is a chi-squared random variable along with (60) above to derive the following tail bound:

$$\begin{aligned}\mathrm{P}\left\{|\langle \boldsymbol{v}, \boldsymbol{c}_k\rangle|\,\|\boldsymbol{c}_k\|_2 > \lambda\sqrt{N}\|\boldsymbol{v}\|_2\right\} &\leq \mathrm{e}^{-\lambda/\tau^2} + 1.05\,\mathrm{e}^{-\tau^2\lambda/8} \\ &= 2.05\,\mathrm{e}^{-\lambda/\sqrt{8}},\end{aligned}$$

where we have chosen $\tau^2 = \sqrt{8}$. Thus

$$\mathrm{P}\left\{|\langle \boldsymbol{v}, \boldsymbol{c}_k\rangle|\,\|\boldsymbol{c}_k\|_2 + \|\boldsymbol{v}\|_2 > \lambda\right\} \leq 2.05\,\mathrm{e}^{1/\sqrt{8}} \cdot \mathrm{e}^{-\lambda/\sqrt{8N}}.$$

□

Lemma 11. *Let $\boldsymbol{c}_k \in \mathbb{C}^N$ be normally distributed as in* (15)*, and let $\boldsymbol{v} \in \mathbb{C}^N$ be an arbitrary vector. Then*

$$\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{v}\boldsymbol{v}^*(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})] = \|\boldsymbol{v}\|_2^2\boldsymbol{I}.$$

Proof. We have

$$\begin{aligned}\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{v}\boldsymbol{v}^*(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})] &= \mathrm{E}[|\langle\boldsymbol{v},\boldsymbol{c}_k\rangle|^2\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{c}_k\boldsymbol{c}_k^*\boldsymbol{v}\boldsymbol{v}^* - \boldsymbol{v}\boldsymbol{v}^*\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{v}\boldsymbol{v}^*] \\ &= \mathrm{E}[|\langle\boldsymbol{v},\boldsymbol{c}_k\rangle|^2\boldsymbol{c}_k\boldsymbol{c}_k^*] - \boldsymbol{v}\boldsymbol{v}^*.\end{aligned}$$

Let $R(i,j)$ be the entries of the first matrix above:

$$\begin{aligned}R(i,j) &= \mathrm{E}[|\langle\boldsymbol{v},\boldsymbol{c}_k\rangle|^2 c_k[i]c_k[j]^*] \\ &= \sum_{n_1,n_2} v[n_1]v[n_2]\,\mathrm{E}[c_k[n_1]c_k[n_2]^*c_k[i]c_k[j]^*].\end{aligned}$$

On the diagonal, where $i = j$, all of the terms in the sum above are zero except when $n_1 = n_2$, and so

$$R(i,i) = \sum_{n=1}^{N} |v[n]|^2\,\mathrm{E}\left[|c_k[n]|^2|c_k[i]|^2\right].$$

Using the fact that

$$\mathrm{E}\left[|c_k[n]|^2|c_k[i]|^2\right] = \begin{cases}2 & n = i \\ 1 & n \neq i\end{cases},$$

we see that $R(i,i) = |v[i]|^2 + \|\boldsymbol{v}\|_2^2$. Off the diagonal, where $i \neq j$, we see immediately that $\mathrm{E}[c_k[n_1]c_k[n_2]^*c_k[i]c_k[j]^*]$ will be zero unless one of two (non-overlapping) conditions hold: $(n_1 = i, n_2 = j)$ or $(n_1 = j, n_2 = i)$. Thus

$$R(i,j) = v[i]v[j]\,\mathrm{E}[c_k[i]^2]\,\mathrm{E}[c_k[j]^2] + v[j]v[i]\,\mathrm{E}[|c_k[j]|^2]\,\mathrm{E}[|c_k[i]|^2].$$

Note the lack of absolute values in the first term on the right above; in fact, since the $c_k[i]$ have uniformly distributed phase, $\mathrm{E}[c_k[i]^2] = \mathrm{E}[c_k[j]^2] = 0$, and so $R(i,j) = v[i]v[j]$. As such

$$\mathrm{E}[(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})\boldsymbol{v}\boldsymbol{v}^*(\boldsymbol{c}_k\boldsymbol{c}_k^* - \boldsymbol{I})] = \mathrm{E}[|\langle\boldsymbol{v},\boldsymbol{c}_k\rangle|^2\boldsymbol{c}_k\boldsymbol{c}_k^*] - \boldsymbol{v}\boldsymbol{v}^* = \boldsymbol{v}\boldsymbol{v}^* + \|\boldsymbol{v}\|_2^2\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^* = \boldsymbol{I}.$$

□

Lemma 12. *Let $\boldsymbol{c}_k \in \mathbb{C}^N$ be normally distributed as in* (15)*, and let $\boldsymbol{u},\boldsymbol{v} \in \mathbb{C}^N$ be arbitrary vectors. Then*

$$\mathrm{P}\left\{|\langle\boldsymbol{c}_k,\boldsymbol{v}\rangle\langle\boldsymbol{u},\boldsymbol{c}_k\rangle - \langle\boldsymbol{u},\boldsymbol{v}\rangle| > \lambda\right\} \leq 2\mathrm{e}\cdot\exp\left(-\frac{\lambda}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2}\right).$$

Proof. For any $t > 0$,

$$\begin{aligned}\mathrm{P}\left\{|\langle\boldsymbol{c}_k,\boldsymbol{v}\rangle\langle\boldsymbol{u},\boldsymbol{c}_k\rangle| > \lambda\right\} &\leq \mathrm{P}\left\{|\langle\boldsymbol{c}_k,\boldsymbol{v}\rangle| > t\right\} + \mathrm{P}\left\{|\langle\boldsymbol{u},\boldsymbol{c}_k\rangle| > \lambda/t\right\} \\ &= \mathrm{P}\left\{|\langle\boldsymbol{c}_k,\boldsymbol{v}\rangle|^2 > t^2\right\} + \mathrm{P}\left\{|\langle\boldsymbol{u},\boldsymbol{c}_k\rangle|^2 > \lambda^2/t^2\right\} \\ &\leq \exp\left(-\frac{t^2}{\|\boldsymbol{v}\|_2^2}\right) + \exp\left(-\frac{\lambda^2}{t^2\|\boldsymbol{u}\|_2^2}\right).\end{aligned}$$

Choosing $t^2 = \lambda \|\boldsymbol{v}\|_2 / \|\boldsymbol{u}\|_2$ yields

$$\mathrm{P}\left\{|\langle \boldsymbol{c}_k, \boldsymbol{v}\rangle\langle \boldsymbol{u}, \boldsymbol{c}_k\rangle| > \lambda\right\} \le 2\exp\left(-\frac{\lambda}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2}\right),$$

and so

$$\begin{aligned}
\mathrm{P}\left\{|\langle \boldsymbol{c}_k, \boldsymbol{v}\rangle\langle \boldsymbol{u}, \boldsymbol{c}_k\rangle - \langle \boldsymbol{u}, \boldsymbol{v}\rangle| > \lambda\right\} &\le \mathrm{P}\left\{|\langle \boldsymbol{c}_k, \boldsymbol{v}\rangle\langle \boldsymbol{u}, \boldsymbol{c}_k\rangle| > \lambda - \|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2\right\} \\
&\le 2\exp\left(-\frac{\lambda}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2} + 1\right) \\
&= 2\mathrm{e}\cdot\exp\left(-\frac{\lambda}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2}\right).
\end{aligned}$$

□

References

[1] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford University, March 2002.

[2] Y. Nesterov, H. Wolkowicz, and Y. Ye, "Nonconvex quadratic optimization," in *Handbook of Semidefinite Programming*, ser. International series in operations research and management science, H. Wolkowicz, R. Saigal, and L. Vandenberghe, Eds. Boston, MA: Kluwer Academic Publishers, 2000, pp. 361–420.

[3] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[4] E. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *To appear in Comm. Pure Appl. Math.*, 2012.

[5] M. S. Asif, W. Mantzel, and J. Romberg, "Random channel coding and blind deconvolution," in *Proc. IEEE Allerton Conf. on Comm., Control, and Comp.*, Monticello, IL, September 2009, pp. 1021–1025.

[6] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, December 2005.

[7] M. Rudelson and R. Vershynin, "Geometric approach to error correcting codes and reconstruction of signals," *Int. Math. Res. Not.*, no. 64, pp. 4019–4041, 2005.

[8] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *IEEE Signal Proc. Mag.*, vol. 13, no. 3, pp. 43–64, May 1996.

[9] T. Chan and C. K. Wong, "Total variation blind deconvolution," *IEEE Trans. Image Proc.*, vol. 7, no. 3, pp. 370–375, 1998.

[10] Y. Sato, "A method of self-recovering equalization for multilevel amplitude-modulation," *IEEE Trans. Commun.*, vol. COM-23, no. 6, pp. 679–682, June 1975.

[11] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Inform. Theory*, vol. 40, no. 2, pp. 340–349, 1994.

[12] H. Liu, G. Xu, L. Tong, and T. Kailath, "Recent developments in blind channel equalization : From cyclostationnarity to subspaces," *Signal Process.*, vol. 50, pp. 83–99, April 1996.

[13] L. Tong and S. Perreau, "Multichannel blind identification: From subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, no. 10, pp. 1951–1968, October 1998.

[14] C. R. Johnson, P. Schniter, T. J. Endres, J. D. Behm, D. R. Brown, and R. A. Casas, "Blind equalization using the constant modulus criterion: A review," *Proc. IEEE*, vol. 86, no. 10, pp. 1927–1950, October 1998.

[15] G. B. Giannakis and C. Tepedelenlioglu, "Basis expansion models and diversity techniques for blind indentification and equalization of time-varying channels," *Proc. IEEE*, vol. 86, no. 10, pp. 1969–1986, October 1998.

[16] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-sqaures approach to blind channel identification," *IEEE Trans. Sig. Proc.*, vol. 43, no. 12, pp. 2982–2993, December 1995.

[17] E. Candès and Y. Plan, "Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements," *IEEE Trans. Inform. Theory*, vol. 57, no. 4, pp. 2342—2359, 2011.

[18] B. Recht, "A simpler approach to matrix completion," *J. Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[19] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[20] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[21] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. of Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[22] A. Ahmed and J. Romberg, "Compressive multiplexers for correlated signals," in *Proc. IEEE Asilomar Conf. on Sig. Sys. and Comp.*, Pacific Grove, CA, November 2012.

[23] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inform. Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[24] V. Koltchinkskii, K. Lounici, and A. Tsybakov, "Nuclear norm penalization and optimal rates for noisy low rank matrix completion," `arXiv:1011.6256`, December 2010.

[25] K. Lee and Y. Bresler, "ADMiRA: Atomic decomposition for minimum rank approximation," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4402–4416, September 2010.

[26] J. Romberg and R. Neelamani, "Sparse channel separation using random probes," *Inverse Problems*, vol. 26, no. 11, November 2010.

[27] L. Applebaum, W. Bajwa, M. F. Duarte, and R. Calderbank, "Asynchronous code-division random access using convex optimization," *Physical Communication*, vol. 5, no. 2, pp. 129–147, June 2011.

[28] J. P. Slavinsky, J. Laska, M. A. Davenport, and R. G. Baraniuk, "The compressive multiplexer for multi-channel compressive sensing," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, Prague, Czech Republic, May 2011, pp. 3980–3983.

[29] S. Burer and R. D. C. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming (Series B)*, vol. 95, pp. 329–357, 2003.

[30] ——, "Local minima and convergence in low-rank semidefinite programming," *Mathematical Programming*, vol. 103, no. 3, pp. 427–444, 2005.

[31] J. Lee, B. Recht, N. Srebro, R. R. Salakhutdinov, and J. A. Tropp, "Practical large-scale optimization for max-norm regularization," in *Advances in Neural Information Processing Systems*, 2010.

[32] B. Recht and C. Ré, "Parallel stochastic gradient algorithms for large-scale matrix completion," 2011, submitted to *Mathematical Programming Computation*. Preprint available at http://www.optimization-online.org/DB_HTML/2011/04/3012.html.

[33] M. Schmidt, "minFunc: unconstrained differentiable multivariate optimization in Matlab," http://www.di.ens.fr/~mschmidt/Software/minFunc.html, 2012.

[34] D. Gross, Y.-K. Liu, S. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Phys. Rev. Letters*, vol. 105, no. 15, 2010.

[35] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–986, June 2007.

[36] E. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[37] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. of Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.

[38] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes.* Springer, 1996.

[39] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of johnson and lindenstrauss," *Random Struct. and Alg.*, vol. 22, no. 1, pp. 60–65, 2003.