Information and Inference: A Journal of the IMA (2021) **10**, 285–332
doi:10.1093/imaiai/iaaa018
Advance Access publication on 24 September 2020

Phase retrieval of low-rank matrices by anchored regression

Kiryung Lee[†]
Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA
[†]Corresponding author. Email: lee.8763@osu.edu

Sohail Bahmani
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Yonina C. Eldar
Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel

and

Justin Romberg
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[Received on 17 December 2018; revised on 29 June 2020; accepted on 3 July 2020]

We study the low-rank phase retrieval problem, where our goal is to recover a $d_1 \times d_2$ low-rank matrix from a series of phaseless linear measurements. This is a fourth-order inverse problem, as we are trying to recover factors of a matrix that have been observed, indirectly, through some quadratic measurements. We propose a solution to this problem using the recently introduced technique of anchored regression. This approach uses two different types of convex relaxations: we replace the quadratic equality constraints for the phaseless measurements by a search over a polytope and enforce the rank constraint through nuclear norm regularization. The result is a convex program in the space of $d_1 \times d_2$ matrices. We analyze two specific scenarios. In the first, the target matrix is rank-1, and the observations are structured to correspond to a phaseless blind deconvolution. In the second, the target matrix has general rank, and we observe the magnitudes of the inner products against a series of independent Gaussian random matrices. In each of these problems, we show that anchored regression returns an accurate estimate from a near-optimal number of measurements given that we have access to an anchor matrix of sufficient quality. We also show how to create such an anchor in the phaseless blind deconvolution problem from an optimal number of measurements and present a partial result in this direction for the general rank problem.

1. Introduction

We consider the problem of recovering a low-rank matrix $\boldsymbol{X}_\sharp$ from phaseless linear measurements of the form

$$y_m = |\langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\sharp \rangle|^2 + \xi_m, \quad m = 1, \dots, M. \tag{1}$$

We refer to this inverse problem as *low-rank phase retrieval* (LRPR). LRPR is a combination of two problems that have received a lot of attention over the past decade. The phase retrieval problem, where

the goal is to recover a vector $\boldsymbol{x} \in \mathbb{R}^d$ from M quadratic measurements of the form $|\langle \boldsymbol{x}, \boldsymbol{\phi}_m \rangle|^2$, is known to be solvable when the $\boldsymbol{\phi}_m$ are generic and $M \gtrsim d$ (e.g. see [33] and references therein). There are tractable algorithms for solving the equations that use convex relaxations based on semi-definite programming [14, 16, 61] and polytope constraints [6, 28]. There also exist fast iterative algorithms for nonconvex programming (e.g. [17, 18, 51, 55, 56, 62]). The problem of recovering a $d_1 \times d_2$ matrix of rank r from M linear measurements of the form $\langle \boldsymbol{\Phi}_m, \boldsymbol{X} \rangle$ has also been thoroughly analyzed in the literature for generic $\boldsymbol{\Phi}_m$ [12, 53], $\boldsymbol{\Phi}_m$ that return samples of the matrix [13, 15, 36, 52] and $\boldsymbol{\Phi}_m$ with structured randomness [4, 30]; a survey of these results can be found in [22].

Our contribution in this paper is to show that for certain choices of $\boldsymbol{\Phi}_m$, we can recover $\boldsymbol{X}_\sharp$ from phaseless measurements (1) from far fewer than $d_1 d_2$ measurements by taking advantage of the low-rank structure of $\boldsymbol{X}_\sharp$. Our recovery algorithm uses the recently developed idea of *anchored regression* [6, 7]. The common approaches to estimate $\boldsymbol{X}_\sharp$ from the nonlinear observations (1) lead to nonconvex programs. The anchored regression, however, enables estimation by convex programming as follows. The first step is effectively relaxing the nonlinear equations (1) to convex feasibility constraints. The second step is to use an *anchor matrix* $\boldsymbol{X}_0$, which serves as an initial guess for the solution, to formulate a simple convex program that finds a matrix that is feasible in the relaxed constraints and is best aligned with $\boldsymbol{X}_0$. When the measurements are noiseless ($\xi_m = 0$), we solve

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{X}} & -\text{Re}\,\langle \boldsymbol{X}_0, \boldsymbol{X} \rangle + \lambda \|\boldsymbol{X}\|_* \\ \text{subject to} & |\langle \boldsymbol{\Phi}_m, \boldsymbol{X} \rangle|^2 \le y_m, \quad m = 1, \ldots, M. \end{array} \tag{2}$$

This is a convex program over the space of $d_1 \times d_2$ matrices. Geometrically, each constraint $|\langle \boldsymbol{\Phi}_m, \boldsymbol{X} \rangle|^2 \le y_m$ is a convex set that has the target $\boldsymbol{X}_\sharp$ on its surface. The program finds an extreme point of the intersection of these convex sets by minimizing the linear functional $-\text{Re}\,\langle \boldsymbol{X}_0, \boldsymbol{X} \rangle$ regularized by the nuclear norm $\|\boldsymbol{X}\|_*$ to account for the low-rank structure of the solution. The success of this program in recovering the target (to within a global phase ambiguity) depends on the behavior of the constraints around $\boldsymbol{X}_\sharp$ and having an anchor $\boldsymbol{X}_0$ sufficiently correlated with $\boldsymbol{X}_\sharp$.

When there is noise, we relax the constraints in (2) and solve

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{X}} & -\text{Re}\,\langle \boldsymbol{X}_0, \boldsymbol{X} \rangle + \lambda \|\boldsymbol{X}\|_* \\ \text{subject to} & \frac{1}{M} \sum_{m=1}^{M} (|\langle \boldsymbol{\Phi}_m, \boldsymbol{X} \rangle|^2 - y_m)_+ \le \eta\,, \end{array} \tag{3}$$

where $(\cdot)_+$ denotes the positive part function. This yields a stable solution in the sense that if the conditions for noise-free recovery are met, and we choose η larger than the positive part of the perturbations, that is,

$$\eta = \frac{1}{M} \sum_{m=1}^{M} (-\xi_m)_+ + \epsilon, \quad \text{for some } \epsilon \ge 0,$$

then the solution $\widehat{\boldsymbol{X}}$ to (3) obeys $\|\widehat{\boldsymbol{X}} - e^{\mathrm{j}\theta} \boldsymbol{X}_\sharp\|_{\mathrm{F}} \lesssim \eta$ for some $\theta \in [0, 2\pi)$. Here ϵ denotes an error in estimating the average of the positive part of perturbations by η.

We analyze two scenarios in detail. In the first scenario, the target matrix $\boldsymbol{X}_\sharp$ is of rank r, and the measurement matrices $\boldsymbol{\Phi}_m$ have independent real-valued Gaussian entries. Theorem 4.2 below shows that if we start with an anchor matrix that is sufficiently close to $\boldsymbol{X}_\sharp$, exact recovery occurs when $M \gtrsim r(d_1 + d_2) \log(d_1 + d_2)$. Lemma 4.5 shows that the anchor matrix can be computed from the data by a

variation of spectral initialization [17] when the number of measurements M satisfies $M \gtrsim r^3\kappa^4(d_1 + d_2)\log(d_1+d_2)$, where κ denotes the condition number of $\boldsymbol{X}_\sharp$. We also show that the recovery procedure is stable in the presence of noise.

In our second scenario, the target matrix has rank one, $\boldsymbol{X}_\sharp = \sigma \boldsymbol{u}\boldsymbol{v}^*$ with $\boldsymbol{u} \in \mathbb{C}^{d_1}, \boldsymbol{v} \in \mathbb{C}^{d_2}$, $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1$, as do the measurement matrices, $\boldsymbol{\Phi}_m = \boldsymbol{a}_m\boldsymbol{b}_m^*$. As we discuss below, this scenario is a model for the blind deconvolution of two signals from magnitude measurements in the frequency domain. Our analysis in Theorem 4.1 below takes the $\boldsymbol{a}_m$ and $\boldsymbol{b}_m$ to be complex-valued independent Gaussian random vectors. Under this model, we show that anchored regression produces a stable estimate of $(\boldsymbol{u}, \boldsymbol{v})$ when M is within a logarithmic factor of $d_1 + d_2$. Lemma 4.3 gives a computationally efficient technique for constructing the anchor in a commensurate number of measurements.

2. Application: blind deconvolution from Fourier magnitude observations

LRPR arises in a variation of the blind deconvolution, which estimates two unknown signals from the Fourier magnitudes of the convolution. While blind deconvolution is itself an ill-posed, nonlinear problem, the absence of phase information in the Fourier measurements makes it even more challenging. The type of phaseless blind deconvolution problem we describe below arises in various applications in communications and imaging. In optical communications, high spectral efficiency and robustness against adversarial channel conditions for multiple-input multiple-output channels can be achieved using orthogonal frequency division multiplexing (OFDM). Calibrating these communication channels involves solving a blind deconvolution problem. This problem has to be solved from phaseless observations, as practical direct detection receivers work with intensity-only measurements [5] to provide robustness against synchronization errors, which has been one of the key issues in the OFDM systems [10, 54].

A similar calibration problem arises in Fourier ptychography [25]. In this application, an image is computed from phaseless Fourier domain measurements. If there is uncertainty in the point spread function of the optical system, recovering the image becomes a phaseless blind deconvolution problem.

Blind deconvolution that identifies unknown signals $\boldsymbol{x}, \boldsymbol{h} \in \mathbb{C}^M$ (up to reciprocal scaling) from their circular convolution is in general ill-posed but can be solved with a priori information on $\boldsymbol{x}$ and $\boldsymbol{h}$. The circular convolution of $\boldsymbol{x}$ and $\boldsymbol{h}$ can be equivalently expressed in the Fourier domain as the element-wise product, namely

$$\boldsymbol{F}(\boldsymbol{x} \circledast \boldsymbol{h}) = \sqrt{M}\boldsymbol{F}\boldsymbol{x} \odot \boldsymbol{F}\boldsymbol{h}, \tag{4}$$

where $\boldsymbol{F} \in \mathbb{C}^{M\times M}$ is the unitary discrete Fourier matrix of size M.

We will impose subspace priors on $\boldsymbol{x}$ and $\boldsymbol{h}$, modeling $\boldsymbol{x} \in \mathbb{C}^M$ as being in the low-dimensional column space of $\boldsymbol{D} \in \mathbb{C}^{M\times d_1}$ and $\boldsymbol{h}$ as being in the column space of $\boldsymbol{E} \in \mathbb{C}^{M\times d_2}$. Then $\boldsymbol{x}$ and $\boldsymbol{h}$ are represented as

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{u} \quad \text{and} \quad \boldsymbol{h} = \boldsymbol{E}\bar{\boldsymbol{v}}, \tag{5}$$

for some $\boldsymbol{u} \in \mathbb{C}^{d_1}$ and $\boldsymbol{v} \in \mathbb{C}^{d_2}$. Here $\bar{\boldsymbol{v}}$ denotes the entry-wise complex conjugate of $\boldsymbol{v}$. Let $\boldsymbol{a}_m$ denote the mth column of $\boldsymbol{D}^*\boldsymbol{F}^*$ and $\boldsymbol{b}_m$ denote the mth column of $\boldsymbol{E}^\top\boldsymbol{F}^\top$ for $m = 1, \ldots, M$. Then the Fourier measurement of the convolution at frequency m (after an appropriate normalization) is given as $\boldsymbol{a}_m^*\boldsymbol{u}\boldsymbol{v}^*\boldsymbol{b}_m$. Under this subspace model, it suffices to recover $\boldsymbol{u}$ and $\boldsymbol{v}$.

In particular applications, the subspace model for $\boldsymbol{h}$ might be introduced as a linear approximation of parametric models via principal component analysis. This technique is used for source localization and channel estimation in underwater acoustics [49, 57]. Some analysis in the context of dimensionality reduction of manifolds is provided in [24, 48].

In the scenario where only noisy Fourier magnitudes of the convolution is observed, the corresponding quadratic measurements are given in the form of

$$y_m = |\boldsymbol{a}_m^* \boldsymbol{u}\boldsymbol{v}^* \boldsymbol{b}_m|^2 + \xi_m, \quad m = 1, \dots, M,$$

where $\xi_1, \dots, \xi_M$ denote additive noise. Through the lifting reformulation [3] that substitutes $\boldsymbol{u}\boldsymbol{v}^*$ by a rank-1 matrix $\boldsymbol{X}_\sharp$, the recovery reduces to a LRPR that estimates the unknown rank-1 matrix $\boldsymbol{X}_\sharp$ from its noisy quadratic measurements:

$$y_m = |\langle \boldsymbol{a}_m \boldsymbol{b}_m^*, \boldsymbol{X}_\sharp \rangle|^2 + \xi_m, \quad m = 1, \dots, M. \tag{6}$$

This is a particular instance of LRPR and generates quadratic measurements with rank-1 matrices $\boldsymbol{a}_1 \boldsymbol{b}_1^*, \dots, \boldsymbol{a}_M \boldsymbol{b}_M^*$.

In other words, the recovery combines blind deconvolution and phase retrieval; hence, it suffers from the ambiguities in both problems. Similar to phase retrieval, the absence of the phases in the measurements makes the reconstruction a nonconvex problem, even after it has been lifted. By themselves, both phase retrieval and blind deconvolution amount to solving a system of quadratic equations. However, the phaseless blind deconvolution problem (6) is a system of fourth-order equations. Below, we will show that this system can indeed be tractably solved under certain randomness assumptions on the considered subspaces.

3. Related work

Recovery of a structured signal from nonlinear measurements has received a significant amount of attention in the past decade, particularly in terms of theoretical analysis of various optimization formulations. A prominent example is the phase retrieval problem, which recovers an unknown signal from quadratic measurements. Unique identification of the solution and performance guarantees of optimization algorithms in the case where the unknown signal is sparse have been recently studied in [7, 11, 19, 26, 34, 39, 45].

Another example, discussed in the previous section, is the blind deconvolution problem, which amounts to solving a system of bilinear equations. Although many approaches for blind deconvolution and its variations have been proposed in the communications, signal processing and computational imaging literature, there has been significant progress in recent years in identifying provable performance guarantees. These results offer theoretical guarantees on the number of measurements M in (4) as a function of the subspace dimensions d_1, d_2 (number of columns of $\boldsymbol{D}, \boldsymbol{E}$ in (5)) needed to recover $\boldsymbol{u}, \boldsymbol{v}$. Results that exhibit near-optimal scaling of M versus d_1, d_2 are known both for convex relaxations of the problem and for iterative algorithms that minimize a nonconvex loss [3, 32, 44]. These results have also been extended to sparsity (in place of subspace) models where the recovery is performed through alternating minimization [42]; however, the near optimal result in this work makes some technical, and perhaps too restrictive, assumptions on the success of projection steps.

The blind deconvolution problem can be made easier if we have the freedom to obtain diversified observations. Specifically, the identification of unknown channel impulse responses excited by an

unknown source has been studied extensively in the communications literature since the 1990s (e.g. [50, 65]). These classical results assumed that the channel responses have finite length and provided algebraic performance guarantees. In recent years, its generalization to the blind gain and phase calibration problem has been analyzed and robust optimization algorithms were proposed along with performance guarantees [2, 21, 41, 43, 46, 47, 63]. There also exists further generalization to the off-the-grid sparsity models [20, 66].

The nonlinear recovery problem considered in this paper is motivated to study a version of blind deconvolution where the convolution measurements are observed through certain nonlinearities. Bendory *et al*. [9] studied a similar problem arising in blind ptychgraphy and identified a set of conditions under which a signal can be identified uniquely from the magnitudes of a short-time Fourier transform taken with an unknown window. In this paper, we are more interested in the recovery by a practical convex program from Fourier magnitudes. The lifting reformulation renders the reconstruction problem into phase retrieval of a low-rank matrix.

The problem of recovering a low-rank matrix from phaseless linear measurements can also be interpreted as a generalization of classical subspace learning (i.e principal components analysis). This connection was made explicit in [19], where the problem of estimating a covariance matrix from compressed, streaming data was considered. In a subsequent work, [59] considered the quadratic subspace learning problem in a more general setting. A regularized gradient descent method was proposed to solve the LRPR problem, and they provided an analysis for the accuracy of the initialization step under certain randomness assumptions on the measurement matrices.

Unlike the aforementioned works [19, 59], we take a different approach to solving the LRPR problem that uses the recently introduced *anchored regression* [6, 28] technique for relaxing nonlinear measurements. Unlike lifting techniques, this method recasts phase retrieval as a convex program without increasing the number of optimization variables. Unlike techniques based on nonconvex optimization, its analysis relies only on geometry rather than the trajectory of a certain sequence of iterates, which significantly simplifies the derivations. The anchored regression formulation also makes it straightforward to incorporate structural priors on the data through the introduction of convex regularizers [7]. Importantly, we present performance guarantees for stable recovery of low-rank matrices from its random quadratic measurements, which implies exact reconstruction in the noiseless case. Previously, it was only shown that the initialization by a truncated spectral method provides an accurate approximation [59]. After an early version of this paper [40], another approach to the same problem was independently studied in [1]. In contrast to their work, our approach is not restricted to the case of rank-1 measurement matrices and more importantly, anchored regression provides flexibility that allows nonlinearities in the measurement model beyond the quadratic function.

While a general theory for solving equations with convex nonlinearities has been developed, of which (1) is an example, it still remains to compute the key estimates that depend on the structure of the problem (the low-rankness in our case). Furthermore, it is crucial to design an appropriate initialization scheme that provides a valid anchor matrix. We propose a unified approach to the initialization that takes advantage of the separability of the unknown matrix.

It would be of independent interest to see various estimates on functions of random matrices by the noncommutative Rosenthal inequality [35]. All of the matrix Bernstein inequalities [37, 58] and noncommutative Rosenthal inequality [35] provide tail estimates of a sum of independent random matrices. In applying the matrix Bernstein inequalities, one has to verify that all summands have bounded spectral norms (deterministically or almost surely) or compute their Orlicz norms. On the contrary, the noncommutative Rosenthal inequality [35] first computes moment bounds and then provides a tail estimate by the Markov inequality. Particularly when random matrices are given by a

set of Gaussian random variables, the spectral norm is not bounded almost surely and computing the Orlicz norm of the spectral norm is not trivial. Therefore, it is desirable to derive relevant tail estimates by using the noncommutative Rosenthal inequality. Additionally, the expectations of high-order tensor products of Gaussian random vectors in the appendix might be useful in the study of other applications sharing similar tensor structures.

4. Main results

We have four main results. The first two, presented in Section 4.1, give *sample complexity* bounds that relate the accuracy of the estimate returned by (3) to the number of measurements M observed as in (1). In both cases where the $\boldsymbol{\Phi}_m$s are rank-1 matrices such that

$$\boldsymbol{\Phi}_m = \boldsymbol{a}_m \boldsymbol{b}_m^*, \quad \boldsymbol{a}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{b}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}), \quad m = 1, \dots, M, \tag{7}$$

and when they have i.i.d. standard Gaussian entries, i.e.

$$\operatorname{vec}(\boldsymbol{\Phi}_m) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad m = 1, \dots, M, \tag{8}$$

where $\operatorname{vec}(\boldsymbol{\Phi}_m)$ denotes the column vector obtained by vertically stacking the columns of $\boldsymbol{\Phi}$, we achieve a sample complexity that scales nearly optimally with the size of the target matrix $\boldsymbol{X}_\sharp$ and its rank. These results assume that we have an anchor matrix $\boldsymbol{X}_0$ that is sufficiently correlated with $\boldsymbol{X}_\sharp$.

Our next two main results, presented in Section 4.2, show how such an anchor matrix can be created from the measurements using a spectral initialization method. For the random rank-1 measurements, we are able to construct a sufficiently accurate anchor from a number of observations M that is proportional to the degrees of freedom in the model of $\boldsymbol{X}_\sharp$ up to a logarithmic factor. In the case of Gaussian measurements, we only have a partial result in general and show that a very rough anchor can be bootstrapped into a more accurate one. In a special case where $\boldsymbol{X}_\sharp$ is positive semi-definite or is rank-1, however, the results are near-optimal.

4.1 *Sample complexity*

We begin by presenting theorems that give guarantees on the accuracy of the solution to the convex program (3) in relation to the number of measurements M. In both of the theorems below, we assume that we have an anchor matrix $\boldsymbol{X}_0$ that is roughly aligned with the target $\boldsymbol{X}_\sharp$; we defer the construction of this anchor to Section 4.2.

We start with the case where $\boldsymbol{X}_\sharp$ is rank-1, and the measurements are formed by taking the outer product of two random vectors, $\boldsymbol{\Phi}_m = \boldsymbol{a}_m \boldsymbol{b}_m^*$. As discussed in Section 2 above, this scenario is motivated by problems that involve blind deconvolution from quadratic measurements. Since these applications typically involve the Fourier transform, we formulate our results using complex-valued vectors and matrices.

THEOREM 4.1 Let $\boldsymbol{X}_\sharp = \sigma_\sharp \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^*$ be a complex rank-1 matrix observed as in (6) with $\boldsymbol{\Phi}_m = \boldsymbol{a}_m \boldsymbol{b}_m^*$ for $m = 1 \dots, M$, where $\boldsymbol{a}_1, \dots, \boldsymbol{a}_M$ and $\boldsymbol{b}_1, \dots, \boldsymbol{b}_M$ are independent complex Gaussian random vectors as in (7). Suppose that $\boldsymbol{X}_0 = \boldsymbol{u}_0 \boldsymbol{v}_0^*$ with $\|\boldsymbol{u}_0\|_2 = \|\boldsymbol{v}_0\|_2 = 1$ satisfies

$$\inf_{\theta \in [0, 2\pi)} \left\| \boldsymbol{u}_0 \boldsymbol{v}_0^* - e^{\mathrm{i}\theta} \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \right\|_{\mathrm{F}} \leq \delta \tag{9}$$

for $\delta \leq 0.2$. Then one can set the regularization parameter in (3) such that there exist numerical constants C_1, C_2, C_3 and a constant C_δ that depends only on δ, for which the following holds.[1] If

$$\frac{M}{\log^2 M} \geq C_\delta (d_1 + d_2), \tag{10}$$

then the solution $\widehat{\boldsymbol{X}}$ to (3) satisfies

$$\inf_{\theta \in [0,2\pi)} \|\widehat{\boldsymbol{X}} - e^{\mathrm{i}\theta} \boldsymbol{X}_\sharp\|_{\mathrm{F}} \leq \frac{C_1}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}} \left(\frac{1}{M} \sum_{m=1}^{M} |\xi_m| + \epsilon \right) \tag{11}$$

with probability at least $1 - e^{-C_3 M}$. Furthermore, the left and right singular vectors $\widehat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{v}}$ of $\widehat{\boldsymbol{X}}$ satisfy

$$\sin \angle(\widehat{\boldsymbol{u}}, \boldsymbol{u}_\sharp) \vee \sin \angle(\widehat{\boldsymbol{v}}, \boldsymbol{v}_\sharp) \leq \frac{C_2}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2} \left(\frac{1}{M} \sum_{m=1}^{M} |\xi_m| + \epsilon \right). \tag{12}$$

The sufficient number of measurements for stable recovery of $\boldsymbol{X}_\sharp$ (and hence its factors $\boldsymbol{u}_\sharp$ and $\boldsymbol{v}_\sharp$) required by (10), scales nearly optimally. That is, the sufficient number of samples is proportional to the degrees of freedom of the unknown rank-1 matrix, i.e. $d_1 + d_2$. In Section 4.2 below, we will see that we can also find $\boldsymbol{u}_0, \boldsymbol{v}_0$ that obey (9) from a comparable number of measurements. Combining these results shows that we can recover a $d_1 \times d_2$ rank-1 matrix from phaseless rank-1 measurements when M equals to $d_1 + d_2$ up to a logarithmic factor.

Our second sample complexity result states a performance bound for (3) when the measurements are unstructured Gaussian random matrices and the target is a $d_1 \times d_2$ matrix of rank r. This type of measurement model has served as a standard benchmark in the structured recovery literature, and indeed we do obtain a much tighter bound in this case if the target is well conditioned. To ease the derivation, we state the result for real-valued matrices, but it is straightforward to extend it to the complex-valued case at the cost of making the calculations slightly more involved.

THEOREM 4.2 Let $\boldsymbol{X}_\sharp \in \mathbb{R}^{d_1 \times d_2}$ be of rank r, $\boldsymbol{X}_\sharp = \boldsymbol{U}_\sharp \boldsymbol{\Sigma}_\sharp \boldsymbol{V}_\sharp^\top$ denote the compact singular value decomposition (SVD) of $\boldsymbol{X}_\sharp$, and $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_M \in \mathbb{R}^{d_1 \times d_2}$ be Gaussian random matrices as in (8). Suppose that we have an anchor matrix $\boldsymbol{X}_0 = \boldsymbol{U}_0 \boldsymbol{V}_0^\top$, where $\boldsymbol{U}_0^\top \boldsymbol{U}_0 = \boldsymbol{V}_0^\top \boldsymbol{V}_0 = \mathbf{I}_r$, that satisfies

$$\min \left(\left\| \boldsymbol{U}_0 \boldsymbol{V}_0^\top - \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top \right\|_{\mathrm{F}}, \left\| \boldsymbol{U}_0 \boldsymbol{V}_0^\top + \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top \right\|_{\mathrm{F}} \right) \leq \delta \left\| \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top \right\|_{\mathrm{F}} \tag{13}$$

for δ that obeys

$$\frac{\delta}{1 - \lambda} \leq 0.45\,(2.8 - \kappa), \tag{14}$$

[1] As shown in the proof of Theorem 4.1, given δ, one can choose λ explicitly as $0.9 - \delta$. For specific methods of constructing the anchor matrix, an appropriate value of δ can be determined.

where κ and λ denote the condition number of $\boldsymbol{X}_\sharp$ and the regularization parameter in (3), respectively. Then there exist universal constants C_1, C_2 and a constant C_δ that only depends on δ for which the following holds. If

$$M \geq C_\delta r(d_1 + d_2)\log(d_1 + d_2), \tag{15}$$

then the solution $\widehat{\boldsymbol{X}}$ to (3) satisfies

$$\min\left(\|\widehat{\boldsymbol{X}} - \boldsymbol{X}_\sharp\|_{\mathrm{F}},\ \|\widehat{\boldsymbol{X}} + \boldsymbol{X}_\sharp\|_{\mathrm{F}}\right) \ \leq\ \frac{C_1}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}}\left(\frac{1}{M}\sum_{m=1}^{M}|\xi_m| + \epsilon\right),$$

with probability at least $1 - e^{-C_2 M}$.

Although to the authors' knowledge this is the first result of its kind in the literature, and the bound (15) scales in the rank r and dimensions d_1, d_2 as well as one could hope, we point out a few ways this result could be improved. First, the condition (14) is very restrictive in the sense that it applies only to matrices with a small condition number. Second, constructing $\boldsymbol{U}_0\boldsymbol{V}_0^\top$ that obeys (13) is non-trivial; as we will see in Section 4.2 below, we will only really be able to do this with confidence when $\boldsymbol{X}_\sharp$ is positive semi-definite or is rank-1.

4.2 *Spectral initialization with partial trace*

Our main results, presented as Theorems 4.1 and 4.2 above, give bounds on the number of equations M that are needed to guarantee that the solution to (3) has a certain accuracy. This accuracy depends on the anchor matrix $\boldsymbol{X}_0$ being sufficiently close to the unknown matrix $\boldsymbol{X}_\sharp$. In both cases, we use $\boldsymbol{X}_0 = \boldsymbol{U}_0\boldsymbol{V}_0^*$ as an anchor, where $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^*$ is the compact SVD of an approximation of $\boldsymbol{X}_\sharp$; we are after $\boldsymbol{U}_0, \boldsymbol{V}_0$, each with orthonormal columns, such that for some $\delta > 0$ and a unit modulus z, we have

$$\left\|\boldsymbol{U}_0\boldsymbol{V}_0^* - z\boldsymbol{U}_\sharp\boldsymbol{V}_\sharp^*\right\|_{\mathrm{F}} \leq \delta\left\|\boldsymbol{U}_\sharp\boldsymbol{V}_\sharp^*\right\|_{\mathrm{F}}. \tag{16}$$

In each of the main theorems below, the bounds on M scale like δ^{-2}, and we will achieve the tightest results when we can take δ as a constant independent of the matrix dimensions and rank. In this section, we describe a *data-driven* technique for constructing such an anchor matrix.

To understand the challenges in creating the anchor, let us first recall the now well-known spectral initialization for standard phase retrieval for vectors ($d_2 = 1$ in the formulation above). In this case, we use the observations y_m to form the $d_1 \times d_1$ matrix

$$\widehat{\boldsymbol{R}} = \frac{1}{M}\sum_{m=1}^{M} y_m\boldsymbol{\Phi}_m\boldsymbol{\Phi}_m^*, \tag{17}$$

and then use the leading eigenvector of $\widehat{\boldsymbol{R}}$ as the anchor matrix $\boldsymbol{X}_0$. The idea is that when $y_m = |\langle\boldsymbol{X}_0, \boldsymbol{\Phi}_m\rangle|^2$ and the $\boldsymbol{\Phi}_m$ are random and drawn independent of one another, the expectation of $\mathbb{E}[y_m\boldsymbol{\Phi}_m\boldsymbol{\Phi}_m^*]$ has a leading eigenvector that is exactly $\boldsymbol{X}_0$, and for M large enough, the sum in (17)

provides a good approximation to this expectation. In [17], it is shown that (9) holds for constant δ when $M \gtrsim d_1 \log d_1$.

We might consider using the same initialization when $\boldsymbol{X}_\sharp$ and the $\boldsymbol{\Phi}_m$ are matrices. Using a vectorized version of the above, we can form

$$\widehat{\boldsymbol{R}} = \frac{1}{M}\sum_{m=1}^{M} y_m \mathrm{vec}(\boldsymbol{\Phi}_m)\mathrm{vec}(\boldsymbol{\Phi}_m)^*,$$

compute the leading eigenvector, then reshape into a $d_1 \times d_2$ matrix. We are now guaranteed a good anchor when $M \gtrsim d_1 d_2 \log(d_1 d_2)$. The problem, though, is that this bound is independent of the rank of $\boldsymbol{X}_\sharp$; we are interested in recovery results that scale as closely as possible to the intrinsic number of degrees of freedom $r(d_1 + d_2)$ in our matrix model. Simply finding the largest eigenvector of $\widehat{\boldsymbol{R}}$ and then re-arranging into a $d_1 \times d_2$ matrix will not, by itself, result in a matrix that is rank r, and there is no known algorithm with provable performance guarantees for finding a rank-constrained matrix that is maximally aligned with the column space of $\widehat{\boldsymbol{R}}$ (this is a variation on the 'Sparse principal component analysis (PCA)' problem).

Our approach for estimating the anchor matrix will be to estimate the row and column spaces of $\boldsymbol{X}_\sharp$ individually. We will find a $d_1 \times r$ matrix $\boldsymbol{U}_0$ whose columns are orthonormal and approximately span the column space, a $d_2 \times r$ matrix $\boldsymbol{V}_0$ whose columns are orthonormal and approximately span the row space and then take

$$\boldsymbol{X}_0 = \boldsymbol{U}_0 \boldsymbol{V}_0^*.$$

For the column space estimate $\boldsymbol{U}_0$, we choose $d_2 \times q$ *compression matrices* $\boldsymbol{\Psi}_m$ and form

$$\boldsymbol{\Upsilon} = \frac{1}{M}\sum_{m=1}^{M} y_m \boldsymbol{\Phi}_m \boldsymbol{\Psi}_m \boldsymbol{\Psi}_m^* \boldsymbol{\Phi}_m^*, \tag{18}$$

then take the r leading eigenvectors of $\boldsymbol{\Upsilon}$ as $\boldsymbol{U}_0$. Similarly for the row space, we choose $d_1 \times q$ $\boldsymbol{\Psi}'_m$, form

$$\boldsymbol{\Upsilon}' = \frac{1}{M}\sum_{m=1}^{M} y_m \boldsymbol{\Phi}_m^* \boldsymbol{\Psi}'_m \boldsymbol{\Psi}'^*_m \boldsymbol{\Phi}_m, \tag{19}$$

and take the r leading eigenvectors as $\boldsymbol{V}_0$.

With the measurement matrix $\boldsymbol{\Phi}_m$ random, we want to choose the compression matrices $\boldsymbol{\Psi}_m$ in (18) to meet two criteria:

1. The expectation $\mathbb{E}[\boldsymbol{\Upsilon}]$ has leading eigenvectors that span the same r-dimensional space as the eigenvectors of $\boldsymbol{X}_\sharp \boldsymbol{X}_\sharp^*$.
2. The spectral gap between the rth and $(r+1)$th eigenvalues of $\mathbb{E}[\boldsymbol{\Upsilon}]$ is large enough so that it upper bounds the perturbation error $\|\boldsymbol{\Upsilon} - \mathbb{E}\,\boldsymbol{\Upsilon}\|$ for relatively small M. This allows us to use the classical Davis–Kahan theorem to show that the leading eigenvectors of $\boldsymbol{\Upsilon}$ are approximately aligned with the leading eigenvectors of $\mathbb{E}[\boldsymbol{\Upsilon}]$.

Similar statements hold for the $\boldsymbol{\Psi}'_m$ in (19).

For our blind deconvolution from phaseless measurements application, where $\boldsymbol{X}_\sharp = \sigma \boldsymbol{u}\boldsymbol{v}^*$ and $\boldsymbol{\Phi}_m = \boldsymbol{a}_m \boldsymbol{b}_m^*$, there is a clear way to meet these criteria. If we take

$$\boldsymbol{\Psi}_m = \frac{\boldsymbol{b}_m}{\|\boldsymbol{b}_m\|_2^2} \quad \text{and} \quad \boldsymbol{\Psi}'_m = \frac{\boldsymbol{a}_m}{\|\boldsymbol{a}_m\|_2^2}, \quad m = 1, \dots, M,$$

then

$$\boldsymbol{\Upsilon} = \frac{1}{M}\sum_{m=1}^{M} y_m \boldsymbol{a}_m \boldsymbol{a}_m^* = \frac{\sigma^2}{M}\sum_{m=1}^{M} |\boldsymbol{a}_m^* \boldsymbol{u}|^2 |\boldsymbol{v}^* \boldsymbol{b}_m|^2 \boldsymbol{a}_m \boldsymbol{a}_m^* + \xi_m \boldsymbol{a}_m \boldsymbol{a}_m^*, \tag{20}$$

$$\boldsymbol{\Upsilon}' = \frac{1}{M}\sum_{m=1}^{M} y_m \boldsymbol{b}_m \boldsymbol{b}_m^* = \frac{\sigma^2}{M}\sum_{m=1}^{M} |\boldsymbol{a}_m^* \boldsymbol{u}|^2 |\boldsymbol{v}^* \boldsymbol{b}_m|^2 \boldsymbol{b}_m \boldsymbol{b}_m^* + \xi_m \boldsymbol{b}_m \boldsymbol{b}_m^*. \tag{21}$$

For independent $\boldsymbol{a}_m, \boldsymbol{b}_m$ that follow (7), a simple calculation yields

$$\mathbb{E}\,\boldsymbol{\Upsilon} = \sigma^2 \boldsymbol{u}\boldsymbol{u}^* + \left(\sigma^2 + \frac{1}{M}\sum_{m=1}^{M} \xi_m\right)\mathbf{I}, \quad \mathbb{E}\,\boldsymbol{\Upsilon}' = \sigma^2 \boldsymbol{v}\boldsymbol{v}^* + \left(\sigma^2 + \frac{1}{M}\sum_{m=1}^{M} \xi_m\right)\mathbf{I}.$$

The leading eigenvector for $\boldsymbol{\Upsilon}$ is the left singular vector $\boldsymbol{u}$ for $\boldsymbol{X}_\sharp$, the leading eigenvector of $\boldsymbol{\Upsilon}'$ is the right singular vector $\boldsymbol{v}$, and the spectral gap in both cases is σ^2. That $\boldsymbol{\Upsilon} - \mathbb{E}\,\boldsymbol{\Upsilon}$ and $\boldsymbol{\Upsilon}' - \mathbb{E}\,\boldsymbol{\Upsilon}'$ are small enough so that their leading eigenvectors are close to $\boldsymbol{u}$ and $\boldsymbol{v}$ when M is within a logarithmic factor of $(d_1 + d_2)$ is essentially the content of the following lemma.

LEMMA 4.3 Let $\boldsymbol{\Phi}_m = \boldsymbol{a}_m \boldsymbol{b}_m^*$ be as in (7). Let $\boldsymbol{u}_0 \in \mathbb{C}^{d_1}$ (resp. $\boldsymbol{v}_0 \in \mathbb{C}^{d_2}$) be the leading eigenvector of $\boldsymbol{\Upsilon}$ in (20) (resp. $\boldsymbol{\Upsilon}'$ in (21)) with measurements y_m constructed as in (1). Let $\boldsymbol{u}_\sharp$ and $\boldsymbol{v}_\sharp$ denote the left and right singular vectors of the rank-1 matrix $\boldsymbol{X}_\sharp$. Let $\delta \in (0, 1)$ and $\alpha \in \mathbb{N}$. There exist numerical constants C_1, C_2 that only depend on α, for which the following holds. If

$$\frac{M}{\log^3 M} \geq C_1 \delta^{-2}(d_1 + d_2) \tag{22}$$

and

$$\max_{1 \leq m \leq M} |\xi_m| \leq C_2 \|\boldsymbol{X}_\sharp\|^2 \log M, \tag{23}$$

then (9) holds with probability at least $1 - M^{-\alpha}$.

REMARK 4.4 The inequality (23) requires that the signal-to-noise-ratio is larger than the given threshold. The proof of Lemma 4.3 presents a stronger result that holds by (22) and

$$\frac{M}{\log M} \geq C_2 \alpha \left(\frac{\max_{1 \leq m \leq M} |\xi_m|}{\|\boldsymbol{X}_\sharp\|^2} \vee \delta^{-1} \left(\frac{\max_{1 \leq m \leq M} |\xi_m|}{\|\boldsymbol{X}_\sharp\|^2} \right)^2 \right) \delta^{-1}(d_1 + d_2). \tag{24}$$

Indeed, (23) together with (22) implies (24). Even if (23) is violated, (9) still holds with high probability whenever M is large enough to satisfy (24) that naturally adapts to the signal-to-noise-ratio. To achieve the order of the logarithmic term in (22), it is necessary to satisfy $M \lesssim e^{d_1+d_2}$. Since this upper bound is rather trivial compared to (22), we omit the condition in the statement of Lemma 4.3.

Lemma 4.3 along with Theorem 4.1 give us a clean solution to the phaseless blind deconvolution problem. For generic $\boldsymbol{a}_m, \boldsymbol{b}_m$, the system

$$y_m = |\langle \boldsymbol{u}, \boldsymbol{a}_m \rangle \langle \boldsymbol{b}_m, \boldsymbol{v} \rangle|^2 + \text{noise}, \quad m = 1, \dots, M,$$

can be (stably) solved for $\boldsymbol{u}, \boldsymbol{v}$ when M is within a logarithmic factor of $d_1 + d_2$, the total number of unknowns.

For phaseless measurements of a $d_1 \times d_2$ matrix of rank r, the story is unfortunately not as clean, even when the $\boldsymbol{\Phi}_m$ in (1) are i.i.d. Gaussian. The following lemma gives us a partial result on our ability to create a data-driven anchor. It shows that given an estimate of the row space, this estimate can be leveraged into an accurate estimate of the column space.

LEMMA 4.5 Let $\boldsymbol{X}_\sharp$ and $\boldsymbol{\Phi}_m$s be as in Theorem 4.2. Let $\widehat{\boldsymbol{V}} \in \mathbb{R}^{d_2 \times r}$ satisfy $\widehat{\boldsymbol{V}}^\top \widehat{\boldsymbol{V}} = \mathbf{I}_r$. Suppose that $\widehat{\boldsymbol{V}}$ is given a priori and provides an estimate of the row space of $\boldsymbol{X}_\sharp$ so that

$$\left\| (\mathbf{I}_{d_2} - \widehat{\boldsymbol{V}} \widehat{\boldsymbol{V}}^\top) \boldsymbol{V}_\sharp \boldsymbol{V}_\sharp^\top \right\| \le \delta_{\text{in}} \tag{25}$$

for some $\delta_{\text{in}} < 1$. Take $\boldsymbol{\Upsilon}$ as in (18) with $\boldsymbol{\Psi}_m = \widehat{\boldsymbol{V}}$, and let the columns of $\boldsymbol{U}_0$ be the eigenvectors of $\boldsymbol{\Upsilon}$ corresponding to the r-largest eigenvalues. Fix $\delta_{\text{out}} \in (0, 1)$ and $\alpha \in \mathbb{N}$. Then there exist numerical constants C_1, C_2 that only depend on α, for which the following holds. If

$$\frac{M}{\log^3 M} \ge \frac{C_1 \alpha^3 \kappa^4 r^3 d_1}{\delta_{\text{out}}^2 (1 - \delta_{\text{in}})^2} \tag{26}$$

and

$$\frac{\max_{1 \le m \le M} |\xi_m|}{\|\boldsymbol{X}_\sharp\|} \lesssim \sqrt{r} \log M \wedge \frac{\kappa^2 r^2 \log^2 M}{\delta_{\text{out}}(1 - \delta_{\text{in}})}, \tag{27}$$

then

$$\left\| (\mathbf{I}_{d_1} - \boldsymbol{U}_0 \boldsymbol{U}_0^\top) \boldsymbol{U}_\sharp \boldsymbol{U}_\sharp^\top \right\| \le \delta_{\text{out}}, \tag{28}$$

holds with probability $1 - M^{-\alpha}$, where κ denotes the condition number of $\boldsymbol{X}_\sharp$.

REMARK 4.6 If the noise is weak enough to satisfy (27), then (26) implies

$$\frac{M}{\log M} \ge C_2 \alpha \left(\frac{\kappa^2 \max_{1 \le m \le M} |\xi_m|}{\delta_{\text{out}}(1 - \delta_{\text{in}}) \|\boldsymbol{X}_\sharp\|^2} \vee r \left(\frac{\kappa^2 \max_{1 \le m \le M} |\xi_m|}{\delta_{\text{out}}(1 - \delta_{\text{in}}) \|\boldsymbol{X}_\sharp\|^2} \right)^2 \right) r d_1. \tag{29}$$

Similarly to Remark 4.4, Lemma 4.5 can also be strengthened by substituting (27) by (29). The signal-to-noise-ratio need not be larger than the threshold in (27) whenever M also satisfies (29). Indeed, this version of Lemma 4.5 is proved in Appendix D.

Lemma 4.5 shows that one obtains an estimate of the column space of accuracy δ_{out} from a given estimate of the row space of accuracy δ_{in}. Here the accuracy is measured by the sine of the principal angle between two subspaces. The number of measurements M in (26) that guarantees this result increases as one wishes for a more accurate estimate (smaller δ_{out}) or the input to the initialization method is less accurate (larger δ_{in}).

Furthermore, it is straightforward to exchange the roles of $\boldsymbol{U}_\sharp$ and $\boldsymbol{V}_\sharp$ above. If we have an estimate $\widehat{\boldsymbol{U}}$ of $\boldsymbol{U}_\sharp$, then we can form $\boldsymbol{\Upsilon}'$ as in (19) with $\boldsymbol{\Phi}_m = \widehat{\boldsymbol{U}}$, take its leading eigenvectors and have (under analogous conditions as those in the theorem) an accurate estimate of $\boldsymbol{V}_\sharp$.

The scaling of the number of measurements in (26) has suboptimal dependence on the rank, but its dependence on the side length of the matrix is linear.

Producing an estimate of $\boldsymbol{X}_\sharp$ from matrices $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$ whose ranges approximate its row and column spaces is itself non-trivial. It involves solving another phase retrieval problem, finding a diagonal $\boldsymbol{\Sigma}$ so that

$$y_m \approx |\langle \boldsymbol{U}_0 \boldsymbol{\Sigma} \boldsymbol{V}_0^\top, \boldsymbol{\Phi}_m \rangle|^2, \quad m = 1, \dots, M.$$

Although it might be possible to control the error propagation from the estimates $\boldsymbol{U}_0, \boldsymbol{V}_0$ to the solution of the problem above, this analysis appears to be extremely complicated.[2] However, there are two specific scenarios where we can upper-bound the error in estimating $\boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top$ by the subspace estimation errors.

1. Rank-1 case: let $\sigma_\sharp \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^\top$ be the SVD of $\boldsymbol{X}_\sharp$. Let $\phi := \angle(\boldsymbol{u}_0, \boldsymbol{u}_\sharp)$ and $\psi := \angle(\boldsymbol{v}_0, \boldsymbol{v}_\sharp)$. Then

$$\begin{aligned} &\left\| \boldsymbol{u}_0 \boldsymbol{v}_0^\top - \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^\top \right\|_{\text{F}}^2 \wedge \left\| \boldsymbol{u}_0 \boldsymbol{v}_0^\top + \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^\top \right\|_{\text{F}}^2 = 2 - 2\cos\phi\cos\psi \le 2 - 2\cos^2(\phi \vee \psi) \\ &= 2\sin^2(\phi \vee \psi) = \left\| \left(\mathbf{I}_{d_1} - \boldsymbol{u}_0 \boldsymbol{u}_0^\top \right) \boldsymbol{u}_\sharp \right\|_2^2 \vee \left\| \left(\mathbf{I}_{d_2} - \boldsymbol{v}_0 \boldsymbol{v}_0^\top \right) \boldsymbol{v}_\sharp \right\|_2^2. \end{aligned}$$

2. Positive semi-definite case: let $\boldsymbol{U}_\sharp \boldsymbol{\Lambda}_\sharp \boldsymbol{U}_\sharp^\top$ be the SVD of $\boldsymbol{X}_\sharp$. Then

$$\left\| \boldsymbol{U}_0 \boldsymbol{U}_0^\top - \boldsymbol{U}_\sharp \boldsymbol{U}_\sharp^\top \right\|_{\text{F}}^2 = 2r - 2\left\| \boldsymbol{U}_0^\top \boldsymbol{U}_\sharp \right\|_{\text{F}}^2 = 2\left\| \left(\mathbf{I}_{d_1} - \boldsymbol{U}_0 \boldsymbol{U}_0^\top \right) \boldsymbol{U}_\sharp \right\|_{\text{F}}^2 \le 2r \left\| \left(\mathbf{I}_{d_1} - \boldsymbol{U}_0 \boldsymbol{U}_0^\top \right) \boldsymbol{U}_\sharp \right\|^2.$$

For the above two cases, one can combine Theorem 4.2 and Lemma 4.5 to get a complete analysis of the regularized anchored regression. In the latter case, we still assume that an estimate of $\boldsymbol{U}_\sharp$ is given a priori. Lemma 4.5 provides a refined estimate so that we can invoke Theorem 4.2 with the resulting $\boldsymbol{U}_0$.

[2] An alternative approach is to estimate $\boldsymbol{\Sigma}$ from $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$ through extra independent random measurements. However, this approach doubles the number of observations and may not be interesting in practice. Therefore, we pursue analysis in some special cases without extra observations.

5. Proof of main results

The convex program for phase retrieval of low-rank matrices in (3) is variation to a special case of the anchored regression studied in [7] and the performance guarantees in this paper primarily follow from the main results in [7]. The theorems stated in the previous section are basically obtained by computing the key quantities that determine the sample complexity.

5.1 *Theoretical analysis of regularized anchored regression*

At the core of our analysis is an adaptation of the main result of [7]. The main idea of [7, Theorem 2.1] is to use the small-ball method to find a uniform lower bound for a certain empirical process that is determined by the independent random matrices $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_M$ and indexed by a *deterministic* set $\mathcal{H} \subset \mathbb{C}^{d_1 \times d_2}$ containing $\boldsymbol{\Delta} = \widehat{\boldsymbol{X}} - \boldsymbol{X}_\sharp$. Then this uniform lower bound implies an upper bound for the estimation error $\boldsymbol{\Delta}$.

However, the original statement of [7, Theorem 2.1] cannot be applied directly to the problem of interest in this paper because of two important differences. First, due to technical challenges in our specific problem, as elaborated in Section 4.2, we can only construct a weaker form of anchor compared to that considered originally in [7]. Second, we want to address the case of recovering complex and rank-1 matrices as considered in Theorem 4.1. The results of [7], however, only consider variables and operations in the real space. Therefore, we need to adapt the result of [7] with slight modifications so that it becomes compatible with our setting.

As discussed in Section 4.2, instead of an anchor that approximates the ground truth $\boldsymbol{X}_\sharp$, we require the anchor to approximate $\boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^*$ up to a global phase. To be explicit, we only need to consider a complex phase ambiguity in the case of recovering a complex rank-1 target, where we have $\boldsymbol{U}_\sharp = \boldsymbol{u}_\sharp$ and $\boldsymbol{V}_\sharp = \boldsymbol{v}_\sharp$ and the anchor should basically approximate $\boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^*$. In the case of recovering a real-valued low-rank matrix, the phase ambiguity simply reduces to a sign ambiguity.

With these consideration in mind, here and throughout, we assume that the global phase of the anchor $\boldsymbol{X}_0$ is aligned with $\boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^*$, namely

$$\operatorname{Re} \langle \boldsymbol{X}_0, \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^* \rangle \geq 0, \quad \operatorname{Im} \langle \boldsymbol{X}_0, \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^* \rangle = 0, \tag{30}$$

which, if we operate entirely in the real domain, simply reduces to $\langle \boldsymbol{X}_0, \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top \rangle \geq 0$. The assumption (30) can be made without loss of generality because of the following *equivariance* property. For any $\theta \in [0, 2\pi)$, if we replace the anchor $\boldsymbol{X}_0$ in (3) by $e^{\mathrm{i}\theta} \boldsymbol{X}_0$, then the original solution $\widehat{\boldsymbol{X}}$ accordingly changes to $e^{\mathrm{i}\theta} \widehat{\boldsymbol{X}}$. This property is due to fact that the nuclear norm as well as the constraints in (3) are invariant under the mapping $\boldsymbol{X} \mapsto e^{\mathrm{i}\theta} \boldsymbol{X}$. Since we define the accuracy as the distance to the orbit of $\boldsymbol{X}_\sharp$, i.e. $\{e^{\mathrm{i}\omega} \boldsymbol{X}_\sharp : \omega \in [0, 2\pi)\}$, the mentioned adjustment of the anchor will not affect the accuracy guarantees. Indeed, under (30), the assumption in (16) simplifies to

$$\|\boldsymbol{X}_0 - \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^*\|_{\mathrm{F}} \leq \delta \|\boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^*\|_{\mathrm{F}} = \delta \sqrt{r}. \tag{31}$$

Since $\widehat{\boldsymbol{X}}$ is a minimizer to (3) and $\boldsymbol{X}_\sharp$ is within its feasible set, it naturally follows that $\boldsymbol{\Delta} = \widehat{\boldsymbol{X}} - \boldsymbol{X}_\sharp$ belongs to the set of all ascent directions of the objective function given by

$$\mathcal{A} := \Big\{ \boldsymbol{H} \in \mathbb{C}^{d_1 \times d_2} : \inf_{\boldsymbol{G} \in \lambda \partial \|\boldsymbol{X}_\sharp\|_*} \operatorname{Re} \langle \boldsymbol{X}_0 - \boldsymbol{G}, \boldsymbol{H} \rangle \geq 0 \Big\}.$$

It is desirable to construct the anchor matrix $\boldsymbol{X}_0$ from the available measurements and avoid *sample splitting* schemes. However, for such constructions of the anchor matrix, the set $\mathcal{A}$ will also depend on the measurement matrices $\{\boldsymbol{\Phi}_m\}_{m=1}^M$ that complicates the analysis. To avoid these complications, similar to the approach of [7], we relax $\mathcal{A}$ to some superset that is not dependent on the measurement matrices. Here we consider the superset $\mathcal{A}_\delta$ of $\mathcal{A}$, defined as

$$\mathcal{A}_\delta := \left\{ \boldsymbol{H} \in \mathbb{C}^{d_1 \times d_2} \ : \ \inf_{\boldsymbol{G} \in \lambda \partial \|\boldsymbol{X}_\sharp\|_*} \sqrt{r}\delta \|\boldsymbol{H}\|_{\mathrm{F}} + \mathrm{Re}\,\langle \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^* - \boldsymbol{G}, \boldsymbol{H} \rangle \geq 0 \right\}, \tag{32}$$

which is clearly independent of $\{\boldsymbol{\Phi}_m\}_{m=1}^M$. Inclusion of $\mathcal{A}$ in $\mathcal{A}_\delta$ follows from (31), the triangle inequality and the Cauchy–Schwarz inequality.

To address a technical challenge that only arises when operating in the complex domain, for recovery of complex rank-1 matrices we need to make another modification compared to the original result of [7]. Specifically, similar to [6], with $\boldsymbol{X}_\sharp = \sigma_\sharp \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^*$ as the complex rank-1 ground truth, we introduce the set

$$\mathcal{R}_\delta := \left\{ \boldsymbol{H} \in \mathbb{C}^{d_1 \times d_2} \ : \ \left\| \boldsymbol{H} - \frac{\boldsymbol{X}_\sharp \langle \boldsymbol{X}_\sharp, \boldsymbol{H} \rangle}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2} \right\|_{\mathrm{F}} \geq \frac{\sqrt{1-\delta^2}\, \left| \mathrm{Im} \langle \boldsymbol{X}_\sharp, \boldsymbol{H} \rangle \right|}{\delta \|\boldsymbol{X}_\sharp\|_{\mathrm{F}}} \right\}. \tag{33}$$

Obviously, $\mathcal{R}_\delta$ is only important if we operate in the complex domain; in the real domain, $\mathcal{R}_\delta$ is the entire space and effectively can be ignored. The following lemma, proved in Appendix E, the set $\mathcal{R}_\delta$ also contains $\boldsymbol{\Delta}$ when $\boldsymbol{X}_0$ and $\boldsymbol{X}_\sharp$ are at most δ-apart.

LEMMA 5.1 With $\boldsymbol{X}_\sharp = \sigma_\sharp \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^*$, suppose that (30) and

$$\left\| \boldsymbol{X}_0 - \frac{\boldsymbol{X}_\sharp \langle \boldsymbol{X}_\sharp, \boldsymbol{X}_0 \rangle}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2} \right\|_{\mathrm{F}} \leq \delta \|\boldsymbol{X}_0\|_{\mathrm{F}} \tag{34}$$

hold. Then $\widehat{\boldsymbol{X}} - \boldsymbol{X}_\sharp \in \mathcal{R}_\delta$.

Finally, based on the arguments in [7,Theorem 2.1], our result depends on the following two key quantities defined with respect to the set $\mathcal{H} = \mathcal{A}_\delta \cap \mathcal{R}_\delta$. First, the Rademacher complexity of $\mathcal{H}$ is defined as

$$\mathfrak{C}_M(\mathcal{H}) := \mathbb{E} \sup_{\boldsymbol{H} \in \mathcal{H} \setminus \{\boldsymbol{0}\}} \frac{1}{\sqrt{M}} \sum_{m=1}^M \frac{\epsilon_m \mathrm{Re}(\langle \boldsymbol{X}_\sharp, \boldsymbol{\Phi}_m \rangle \langle \boldsymbol{\Phi}_m, \boldsymbol{H} \rangle)}{\|\boldsymbol{H}\|_{\mathrm{F}}}, \tag{35}$$

where $\epsilon_1, \ldots, \epsilon_M$ are i.i.d. Rademacher random variables independent of everything else. Second, for $\tau > 0$, we also consider a variation of small-ball probability that is defined as

$$P_\tau(\mathcal{H}) := \inf_{\boldsymbol{H} \in \mathcal{H}} \mathbb{P}(\mathrm{Re}(\langle \boldsymbol{X}_\sharp, \boldsymbol{\Phi}_m \rangle \langle \boldsymbol{\Phi}_m, \boldsymbol{H} \rangle) \geq \tau \|\boldsymbol{H}\|_{\mathrm{F}}). \tag{36}$$

Equipped with these notions, the following theorem provides the accuracy guarantees for the regularized anchored regression in the context of LRPR problem.

THEOREM 5.2 (An adaptation of 7, Theorem 2.1 for LRPR). Suppose that $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_M$ in (3) are independent random matrices, and $\boldsymbol{X}_0$ satisfies (16), (30), and (34), where $0 < \delta < 1$. Recalling the definitions (32), (33), (35) and (36), for any $t > 0$, if

$$M \geq 4 \left(\frac{\mathfrak{C}_M(\mathcal{A}_\delta \cap \mathcal{R}_\delta) + t\tau}{\tau P_\tau(\mathcal{A}_\delta \cap \mathcal{R}_\delta)} \right)^2, \tag{37}$$

then the solution $\widehat{\boldsymbol{X}}$ to (3) obeys

$$\inf_{\theta \in [0,2\pi)} \|\widehat{\boldsymbol{X}} - e^{\mathrm{i}\theta} \boldsymbol{X}_\sharp\|_{\mathrm{F}} \leq \frac{2}{\tau P_\tau(\mathcal{A}_\delta \cap \mathcal{R}_\delta)} \left(\frac{1}{M} \sum_{m=1}^{M} |\xi_m| + \epsilon \right)$$

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

REMARK 5.3 A few remarks on Theorem 5.2 are in order.

1. We emphasize again that the required conditions in (30) for the anchor can be made without loss of generality due to the equivariance property discussed above.
2. The original result in [7,Theorem 2.1] considered the problem only in the real domain, where the condition (30) is reduced to the (implicit) assumption $\langle \boldsymbol{X}_0, \boldsymbol{X}_\sharp \rangle \geq 0$. As mentioned above, in this scenario the set $\mathcal{R}_\delta$ becomes trivial (i.e. $\mathcal{R}_\delta = \mathbb{R}^{d_1 \times d_2}$) as well.
3. The additive noise ξ_m to the quadratic measurement $|\langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\sharp \rangle|^2$ is arbitrary fixed. Specifically, we assume that ξ_m does not depend on $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_M$.

Theorems 4.1 and 4.2 are then obtained from Theorem 5.2 by specifying key estimates depending on the corresponding measurement matrices. For the convenience in computing these estimates, we provide a more explicit characterization of $\mathcal{A}_\delta$ as follows. The subdifferential of $\|\cdot\|_*$ at $\boldsymbol{X}_\sharp$, whose SVD is $\boldsymbol{U}_\sharp \boldsymbol{\Sigma}_\sharp \boldsymbol{V}_\sharp^*$, is expressed as

$$\partial \|\boldsymbol{X}_\sharp\|_* = \left\{ \boldsymbol{Z} \,:\, \mathcal{P}_T(\boldsymbol{Z}) = \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^*,\ \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\| \leq 1 \right\}, \tag{38}$$

where $\mathcal{P}_T : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$ denotes the orthogonal projection onto the *tangent space* T of the manifold of rank-r matrices at $\boldsymbol{X}_\sharp$ given by

$$T = \left\{ \boldsymbol{U}_\sharp \widetilde{\boldsymbol{V}}^* + \widetilde{\boldsymbol{U}} \boldsymbol{V}_\sharp^* \,:\, \widetilde{\boldsymbol{V}} \in \mathbb{C}^{d_2 \times r},\ \widetilde{\boldsymbol{U}} \in \mathbb{C}^{d_1 \times r} \right\}$$

and $\mathcal{P}_{T^\perp} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$ denotes the projection onto $T^\perp$, the perpendicular subspace of T. By plugging in the expression of the subdifferential in (38) to (32), we obtain an alternative expression of $\mathcal{A}_\delta$ given by

$$\mathcal{A}_\delta = \left\{ \boldsymbol{H} \in \mathbb{C}^{d_1 \times d_2} \,:\, \sqrt{r}\delta \|\boldsymbol{H}\|_{\mathrm{F}} - \lambda \|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_* + (1-\lambda) \mathrm{Re}\, \langle \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^*, \boldsymbol{H} \rangle \geq 0 \right\}. \tag{39}$$

5.2 *Proof of Theorem 4.2*

All matrices and scalars are real-valued in Theorem 4.2. Thus, $\mathcal{R}_\delta$ becomes trivial and it suffices to compute estimates of $P_\tau(\mathcal{H})$ and $\mathfrak{C}_M(\mathcal{H})$ for $\mathcal{H} = \mathcal{A}_\delta$. The following lemmas respectively provide estimates of $P_\tau(\mathcal{A}_\delta)$ and $\mathfrak{C}_M(\mathcal{A}_\delta)$ whose proofs are deferred to Appendices F and G.

LEMMA 5.4 Suppose the hypotheses in Theorem 4.2 hold. Then for any $\tau' > 0$,

$$\inf_{\boldsymbol{H}\in\mathcal{A}_\delta} \mathbb{P}\Big(\mathrm{Re}(\langle \boldsymbol{X}_\sharp, \boldsymbol{\Phi}_m\rangle\langle \boldsymbol{\Phi}_m, \boldsymbol{H}\rangle) \geq \tau'\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}\|\boldsymbol{H}\|_{\mathrm{F}}\Big) \geq \frac{\exp(-20\tau')}{10}. \tag{40}$$

LEMMA 5.5 Suppose the hypotheses in Theorem 4.2 hold. Then

$$\mathfrak{C}_M(\mathcal{A}_\delta) \leq \frac{C(1-\lambda+\delta)\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}\sqrt{r(d_1+d_2)\log(d_1+d_2)}}{\lambda} \tag{41}$$

for a numerical constant C.

To prove Theorem 4.2, we only need to apply the above estimates in Theorem 5.2. We first show that the assumptions of Theorem 4.2 are sufficient to invoke Theorem 5.2. Following the discussion in Section 5.1, the condition (30) can be satisfied without loss of generality by flipping the sign of $\boldsymbol{X}_0$ if necessary. Fix τ' to a positive constant (e.g. $\tau' = 0.1$). Let $\tau = \tau'\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}$. Then Lemma 5.4 implies that $\tau P_\tau(\mathcal{A}_\delta) \geq c\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}$ for a numerical constant $c > 0$. Choosing $\lambda = 0.9 - \delta$ makes the right-hand side of (41) an increasing function of δ. Then, by Lemma 5.5, the Rademacher complexity $\mathfrak{C}_M(\mathcal{A}_\delta)$ is upper-bounded by $\sqrt{r(d_1+d_2)\log(d_1+d_2)}$ up to a constant solely determined by δ. Therefore, (15) implies that (37) holds whenever $t\tau'$ is dominated by $\sqrt{M}$. We can choose t so that the probability of failure is at most $e^{-2t^2} = e^{-cM}$, for some numerical constant $c > 0$.

5.3 *Proof of Theorem 4.1*

Theorem 4.1 considers recovery of complex-valued rank-1 matrices. We apply Theorem 5.2 for $\mathcal{H} = \mathcal{A}_\delta \cap \mathcal{R}_\delta$ to prove Theorem 4.1. The following lemmas, proved in Appendices H and I, respectively, provide a lower bound on $P_\tau(\mathcal{H})$ and an upper bound on $\mathfrak{C}_M(\mathcal{H})$.

LEMMA 5.6 Suppose the hypotheses in Theorem 4.1 hold. Suppose that $\delta + \lambda < 1$ and $\delta \leq 0.2$. Then there exists a numerical constant $\tau' > 0$ such that

$$\inf_{\boldsymbol{H}\in\mathcal{A}_\delta\cap\mathcal{R}_\delta} \mathbb{P}\Big(\mathrm{Re}(\boldsymbol{b}^*\boldsymbol{X}_\sharp^*\boldsymbol{a}\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}) \geq \tau'\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}\|\boldsymbol{H}\|_{\mathrm{F}}\Big) \geq C_{\tau'},$$

where $C_{\tau'}$ is a positive numerical constant that only depends on τ'.

LEMMA 5.7 Suppose the hypotheses in Theorem 4.1 hold. Then

$$\mathfrak{C}_M(\mathcal{A}_\delta) \leq \frac{C(1-\lambda+\delta)\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}\sqrt{d_1+d_2}\,\log M}{\lambda} \tag{42}$$

for a numerical constant C.

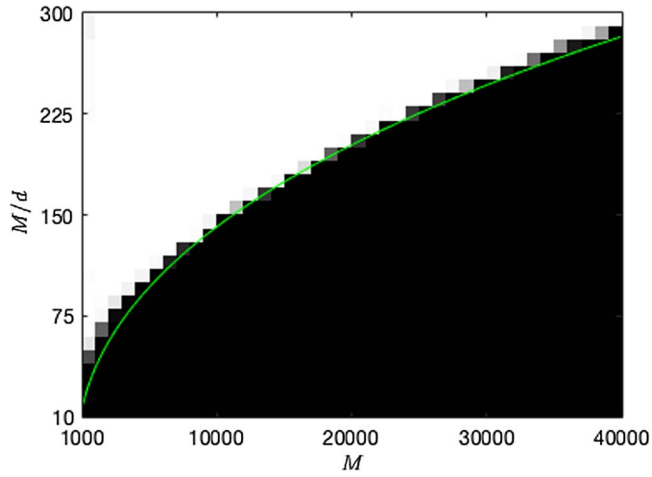

FIG. 1. Empirical phase transition in the noiseless case with rank-1 measurements. The success rate out of 100 trials is plotted in a gray scale (white: all success, black: all failure).

The error bound in (11) then follows from Theorem 5.2 with the above estimates given by Lemmas 5.6 and 5.7. To apply Lemma 5.6, we choose $\lambda = 0.9 - \delta$. Then similar to the proof of Theorem 4.2, the factor $(1 - \lambda + \delta)/\lambda$ becomes an increasing function in δ. The constant C_δ is given by this function of δ together with the result of Lemma 5.6.

Finally, the error bound for the estimation of $\boldsymbol{u}$ and $\boldsymbol{v}$ in (12) follows immediately from the Davis–Kahan theorem (Theorem C.1).

6. Numerical results

We conducted a Monte Carlo simulation to study the empirical performance of the proposed convex programs. Specifically, we considered the optimization problem in (2) in the noiseless case where the measurement matrices are given as the outer product of two Gaussian random vectors and the unknown rank-1 matrix is a square matrix ($d_1 = d_2 = d$). To solve (2), we used the software package `TFOCS` [8] that uses a smoothed conic dual formulation.

Figure 1 illustrates the empirical phase transition. For a fixed number of measurements M, we vary the matrix size d where the ratio M/d belongs to a given interval. In Fig. 1, the convex program provides the exact recovery when d is below a certain threshold determined by M. The sample complexity result by Theorem 4.1 and Lemma 4.3 quantifies this threshold as $CM/\log^\alpha M$ for some constants $C, \alpha > 0$. Alternatively, if the oversampling rate M/d exceeds a polylog factor of M, then the convex program provides the exact recovery. The empirical phase transition occurs at $M/d \approx 0.14 \log^5 M$ or $d \approx 7.3M/\log^5 M$ indicated by the green curve in the figure. Although the requirements for the constants C and α in our proofs seem conservative, our theory is consistent with the empirical performance up to the choice of these constants.

7. Discussions

We proposed a simple initial estimation using partial traces. The regularized anchored regression with the nuclear norm given by this initial estimate provides a stable estimate for LRPR. Performance guarantees were derived for several random measurement models.

One may consider several alternative approaches to LRPR. By applying the lifting trick twice, LRPR can be equivalently cast as a linear inverse problem where the solution is a 4-way tensor of rank-1. Then recovery can be carried out by tensor nuclear norm minimization as a convex relaxation. This fully convexified approach can be attractive as it does not require any initial estimate. However, it is NP-hard to compute the tensor nuclear norm [31], which renders the approach impractical. Recent iterative nonconvex methods such as Wirtinger flow [17] and its variation for the case of sparse signals [11] can be adapted to address low-rank recovery problems. Nevertheless, it will be still required to obtain an accurate initial estimate. Furthermore, the analysis might involve more complications such as the need for resampling in each iteration. In contrast, anchored regression, originally proposed for ordinary phase retrieval and later modified to a regularized version to accommodate various priors on the solution, allows a streamlined analysis by leveraging the geometry of the corresponding convex optimization problem. Its low computational cost also competes with that of the nonconvex approaches.

Acknowledgements

The authors thank the anonymous reviewers for their careful reading of the manuscript and their many insightful comments and suggestions.

Funding

National Science Foundation Computing and Communication Foundations-1718771, by Center for Brain-Inspired Computing, one of six centers in Joint University Microelectronics Program, a Semiconductor Research Corporation program sponsored by Defense Advanced Research Projects Agency, and by the EU Horizon 2020 research and innovation program under 646804-ERC-COG-BNYQ.

References

1. Ahmed, A., Aghasi, A. & Hand, P. (2018) Blind deconvolutional phase retrieval via convex programming. *Advances in Neural Information Processing Systems*, pp. 10030–10040.
2. Ahmed, A. & Demanet, L. (2018) Leveraging diversity and sparsity in blind deconvolution. *IEEE Trans. Inf. Theory*, **64**, 3975–4000.
3. Ahmed, A., Recht, B. & Romberg, J. (2014) Blind deconvolution using convex programming. *IEEE Trans.* Inf. Theory, **60**, 1711–1732.
4. Ahmed, A. & Romberg, J. (2015) Compressive multiplexing of correlated signals. *IEEE Trans. Inf. Theory*, **61**, 479–498.
5. Arik, S. Ö. & Kahn, J. M. (2016) Direct-detection mode-division multiplexing in modal basis using phase retrieval. *Opt. Lett.*, **41**, 4265–4268.
6. Bahmani, S. & Romberg, J. (2017) Phase retrieval meets statistical learning theory: a flexible convex relaxation. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR*, vol. 54. pp. 252–260.
7. Bahmani, S. & Romberg, J. (2019) Solving equations of random convex functions via anchored regression. *Found. Comput. Math.*, **19**, 813–841.
8. Becker, S. R., Candès, E. J. & Grant, M. C. (2011) Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, **3**, 165.

9. Bendory, T., Edidin, D. & Eldar, Y. C. (2018) Blind phaseless short-time Fourier transform recovery. *IEEE Trans. Inf.* Theory, **66**, 3232–3241

10. Bouziane, R. & Killey, R. (2015) Blind symbol synchronization for direct detection optical OFDM using a reduced number of virtual subcarriers. *Opt. Express*, **23**, 6444–6454.

11. Cai, T. T., Li, X. & Ma, Z. (2016) Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *Ann. Stat.*, **44**, 2221–2251.

12. Candès, E. & Li, X. (2014) Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. *Found. Comput. Math.*, **14**, 1017–1026.

13. Candès, E. & Recht, B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772.

14. Candès, E., Strohmer, T. & Voroninski, V. (2013) PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, **66**, 1241–1274.

15. Candès, E. & Tao, T. (2010) The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theory*, **56**, 2053–2080.

16. Candès, E. J., Li, X. & Soltanolkotabi, M. (2015) Phase retrieval from coded diffraction patterns. *Appl. Comp. Harm. Anal.*, **39**, 277–299.

17. Candes, E. J., Li, X. & Soltanolkotabi, M. (2015) Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inf. Theory*, **61**, 1985–2007.

18. Chen, Y. & Candes, E. (2015) Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Advances in Neural Information Processing Systems 28*, (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett eds), pp. 739–747.

19. Chen, Y., Chi, Y. & Goldsmith, A. J. (2015) Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inf. Theory*, **61**, 4034–4059.

20. Chi, Y. (2016) Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. *IEEE J. Sel. Topics Signal Process.*, **10**, 782–794.

21. Cosse, A. (2017) A note on the blind deconvolution of multiple sparse signals from unknown subspaces. *Wavelets and Sparsity XVII*, (Y. M. Lu, D. Van De Ville & M. Papadakis eds) vol. 10394, pp. 330–347. International Society for Optics and Photonics. doi: 10.1117/12.2272836.

22. Davenport, M. A. & Romberg, J. (2016) An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Sel. Topics Signal Process.*, **10**, 608–622.

23. Davis, C. & Kahan, W. M. (1970) The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, **7**, 1–46.

24. Dirksen, S. (2016) Dimensionality reduction with subgaussian matrices: a unified theory. *Found. Comput. Math.*, **16**, 1367–1396.

25. Eckert, R., Tian, L. & Waller, L. (2016) Algorithmic self-calibration of illumination angles in Fourier ptychographic microscopy. *Imaging and Applied Optics 2016*, pp. CT2D.3. Optical Society of America.

26. Eldar, Y. C., Sidorenko, P., Mixon, D. G., Barel, S. & Cohen, O. (2015) Sparse phase retrieval from short-time Fourier measurements. *IEEE Signal Process. Lett.*, **22**, 638–642.

27. Foucart, S. & Rauhut, H. (2013) *A Mathematical Introduction to Compressive Sensing*, vol. **1**. Basel: Birkhäuser.

28. Goldstein, T. & Studer, C. (2017) Convex phase retrieval without lifting via PhaseMax. *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. pp. 1273–1281. Sydney, NSW, Australia: JMLR.org.

29. Golub, G. H. & Van Loan, C. F. (2012) Matrix Computations. *Johns Hopkins Studies in the Mathematical Sciences*. Baltimore, Maryland: Johns Hopkins University Press.

30. Gross, D. (2011) Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theory*, **57**, 1548–1566.

31. Hillar, C. J. & Lim, L.-H. (2013) Most tensor problems are NP-hard. *J. ACM*, **60**, 45.

32. Huang, W. & Hand, P. (2018) Blind deconvolution by a steepest descent algorithm on a quotient manifold. *SIAM J. Imaging Sci.*, **11**, 2757–2785.

33. Jaganathan, K., Eldar, Y. C. & Hassibi, B. (2016) Phase retrieval: an overview of recent developments. *Optical Compressive Imaging*. (B. Raton ed). FL: CRC Press, pp. 263–296.

34. Jaganathan, K., Oymak, S. & Hassibi, B. (2017) Sparse phase retrieval: uniqueness guarantees and recovery algorithms. *IEEE Trans. Signal Process.*, **65**, 2402–2410.

35. Junge, M. & Zeng, Q. (2013) Noncommutative Bennett and Rosenthal inequalities. *Ann. Probab*, **41**, 4287–4316.

36. Keshavan, R. H., Montanari, A. & Oh, S. (2010) Matrix completion from a few entries. *IEEE Trans. Inf. Theory*, **56**, 2980–2998.

37. Koltchinskii, V., Lounici, K. & Tsybakov, A. B. (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat.*, **39**, 2302–2329.

38. Latała, R. (2006) Estimates of moments and tails of Gaussian chaoses. *Ann. Probab.*, **34**, 2315–2331.

39. Lecue, G. & Mendelson, S. (2015) Minimax rate of convergence and the performance of empirical risk minimization in phase recovery. *Electron. J. Probab.*, **20**, 1–29.

40. Lee, K., Bahmani, S., Eldar, Y. & Romberg, J. (2018) Phase retrieval of low-rank matrices. Presented at the 7th International Conference on Computational Harmonic Analysis.

41. Lee, K., Krahmer, F. & Romberg, J. (2018) Spectral methods for passive imaging: nonasymptotic performance and robustness. *SIAM J. Imaging Sci.*, **11**, 2110–2164.

42. Lee, K., Li, Y., Junge, M. & Bresler, Y. (2017) Blind recovery of sparse signals from subsampled convolution. *IEEE Trans. Inf. Theory*, **63**, 802–821.

43. Lee, K., Tian, N. & Romberg, J. (2016) Fast and guaranteed blind multichannel deconvolution under a bilinear channel model. *Information Theory Workshop*. Cambridge, UK.

44. Li, X., Ling, S., Strohmer, T. & Wei, K. (2018) Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Appl. Comput. Harmon. Anal.*.

45. Li, X. & Voroninski, V. (2013) Sparse signal recovery from quadratic measurements via convex programming. *SIAM J. Math. Anal.*, **45**, 3019–3033.

46. Li, Y., Lee, K. & Bresler, Y. (2018) Blind gain and phase calibration via sparse spectral methods. *IEEE Trans. Inf. Theory*, **65**, 3097–3123.

47. Ling, S. & Strohmer, T. (2018) Self-calibration via linear least squares. *SIAM J. Imaging Sci.*, **11**, 252–292.

48. Mantzel, W. & Romberg, J. (2015) Compressed subspace matching on the continuum. *Inf. Inference*, **4**, 79–107.

49. Mantzel, W., Romberg, J. & Sabra, K. (2014) Round-robin multiple source localization. *J. Acoust. Soc. Am.*, **135**, 134–147.

50. Moulines, E., Duhamel, P., Cardoso, J.-F. & Mayrargue, S. (1995) Subspace methods for the blind identification of multichannel FIR filters. *IEEE Trans. Signal Process.*, **43**, 516–525.

51. Netrapalli, P., Jain, P. & Sanghavi, S. (2013) Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems 26*. (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger eds), pp. 2796–2804. Curran Associates, Inc. http://papers.nips.cc/paper/5041-phase-retrieval-using-alternating-minimization.pdf.

52. Recht, B. (2011) A simpler approach to matrix completion. *J. Mach. Learn. Res.*, **12**, 3413–3430.

53. Recht, B., Fazel, M. & Parrilo, P. A. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**, 471–501.

54. Schmidl, T. M. & Cox, D. C. (1997) Robust frequency and timing synchronization for OFDM. *IEEE Trans. Commun.*, **45**, 1613–1621.

55. Sun, J., Qu, Q. & Wright, J. (2018) A geometric analysis of phase retrieval. *Found. Comput. Math.*, **18**, 1131–1198.

56. Tan, Y. S. & Vershynin, R. (2018) Phase retrieval via randomized Kaczmarz: theoretical guarantees. *Inf. Inference*, **8**, 97–123.

57. Tian, N., Byun, S.-H., Sabra, K. & Romberg, J. (2017) Multichannel myopic deconvolution in underwater acoustic channels via low-rank recovery. *J. Acoust. Soc. Am.*, **141**, 3337–3348.

58. Tropp, J. A. (2012) User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, **12**, 389–434.
59. Vaswani, N., Nayer, S. & Eldar, Y. C. (2016) Low-rank phase retrieval. *IEEE Trans. Signal Process.*, **65**, 4059–4074.
60. Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, (Y. Eldar, & G. Kutyniok eds), chap. 5. Cambridge, UK: Cambridge Univ. Press, pp. 210–268.
61. Waldspurger, I., d'Aspremont, A. & Mallat, S. (2015) Phase recovery, MaxCut, and complex semidefinite programming. *Math. Program. Ser. A*, **149**, 47–81.
62. Wang, G., Giannakis, G. B. & Eldar, Y. C. (2017) Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Trans. Inf. Theory*, **64**, 773–794.
63. Wang, L. & Chi, Y. (2016) Blind deconvolution from multiple sparse inputs. *IEEE Signal Process. Lett.*, **23**, 1384–1388.
64. Wedin, P. (1972) Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, **12**, 99–111.
65. Xu, G., Liu, H., Tong, L. & Kailath, T. (1995) A least-squares approach to blind channel identification. *IEEE Trans. Signal Process.*, **43**, 2982–2993.
66. Yang, D., Tang, G. & Wakin, M. B. (2016) Super-resolution of complex exponentials from modulations with unknown waveforms. *IEEE Trans. Inf. Theory*, **62**, 5809–5830.

A. Expectations of symmetric Gaussian tensors

We repeatedly use the expectation of various tensor products of an i.i.d. Gaussian vector, which are summarized below. First, we consider the expectation of the fourth-order tensor product.

Lemma A.1 Let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d)$. Then

$$\mathbb{E}\boldsymbol{g} \otimes \boldsymbol{g} \otimes \boldsymbol{g} \otimes \boldsymbol{g} = \sum_{j,k=1}^{d} (\boldsymbol{e}_j \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_k \otimes \boldsymbol{e}_k + \boldsymbol{e}_j \otimes \boldsymbol{e}_k \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_k + \boldsymbol{e}_j \otimes \boldsymbol{e}_k \otimes \boldsymbol{e}_k \otimes \boldsymbol{e}_j),$$

where $\boldsymbol{e}_j$ denotes the jth column of $\mathbf{I}_d$ for $j = 1, \ldots, d$.

Proof of Lemma A.1 The expectation of $\boldsymbol{g} \otimes \boldsymbol{g} \otimes \boldsymbol{g} \otimes \boldsymbol{g}$ is written as

$$\begin{aligned}
&\mathbb{E} \sum_{i_1,i_2,i_3,i_4=1}^{d} (\boldsymbol{e}_{i_1}\boldsymbol{e}_{i_1}^\top \otimes \boldsymbol{e}_{i_2}\boldsymbol{e}_{i_2}^\top \otimes \boldsymbol{e}_{i_3}\boldsymbol{e}_{i_3}^\top \otimes \boldsymbol{e}_{i_4}\boldsymbol{e}_{i_4}^\top)(\boldsymbol{g} \otimes \boldsymbol{g} \otimes \boldsymbol{g} \otimes \boldsymbol{g}) \\
&= \sum_{i_1,i_2,i_3,i_4=1}^{d} \mathbb{E} g_{i_1} g_{i_2} g_{i_3} g_{i_4} (\boldsymbol{e}_{i_1} \otimes \boldsymbol{e}_{i_2} \otimes \boldsymbol{e}_{i_3} \otimes \boldsymbol{e}_{i_4}),
\end{aligned}$$

where g_i denotes the ith entry of $\boldsymbol{g}$ for $i = 1, \ldots, d$. The proof completes by noting that all odd moments of a standard normal variable vanish.

The following lemma is a direct consequence of Lemma A.1.

Lemma A.2 Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d)$. Then

$$\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{y})\boldsymbol{g}\boldsymbol{g}^\top = (\boldsymbol{x}^\top \boldsymbol{y})\mathbf{I}_d + \boldsymbol{x}\boldsymbol{y}^\top + \boldsymbol{y}\boldsymbol{x}^\top.$$

Next we consider the expectation of an 8-way tensor product applying to a fourth-order tensor product of a unit vector.

LEMMA A.3 Let $\boldsymbol{x} \in \mathbb{S}^{d-1}$ and $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_d)$. Then

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{g})^4(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}) &= 24(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}) \\ &\quad + 12\sum_{l=1}^{d}(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{e}_l + \boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l + \boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{e}_l\otimes\boldsymbol{x} \\ &\qquad + \boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l + \boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{x} + \boldsymbol{e}_l\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{x}) \\ &\quad + 3\sum_{j,k=1}^{d}(\boldsymbol{e}_j\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_k + \boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_k + \boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_j),\end{aligned} \tag{A.1}$$

where $\boldsymbol{e}_l$ denotes the jth column of $\mathbf{I}_d$ for $l = 1, \ldots, d$.

PROOF of Lemma A.3 The expectation $\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{g})^4(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})$ is rewritten as

$$\begin{aligned}&\mathbb{E}(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})^\top(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}) \\ &= \sum_{\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{B}_3,\boldsymbol{B}_4\in\{\boldsymbol{P}_{\boldsymbol{x}},\boldsymbol{P}_{\boldsymbol{x}^\perp}\}} \mathbb{E}(\boldsymbol{B}_1\otimes\boldsymbol{B}_2\otimes\boldsymbol{B}_3\otimes\boldsymbol{B}_4)(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})^\top(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}),\end{aligned} \tag{A.2}$$

where $\boldsymbol{P}_{\boldsymbol{x}}$ and $\boldsymbol{P}_{\boldsymbol{x}^\perp}$ denote the orthogonal projection operators onto the subspace spanned by $\boldsymbol{x}$ and its orthogonal complement, respectively.

If any of $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \boldsymbol{B}_4$ is different from the other three matrices, then the corresponding summand in (A.2) becomes zero since it has a factor that is an odd moment of $\boldsymbol{x}^\top\boldsymbol{g} \sim \mathcal{N}(0,1)$. Therefore, it suffices to consider the following three cases.

Case 1: $\boldsymbol{B}_1 = \boldsymbol{B}_2 = \boldsymbol{B}_3 = \boldsymbol{B}_4 = \boldsymbol{P}_{\boldsymbol{x}}$.

Since

$$\boldsymbol{P}_{\boldsymbol{x}}\otimes\boldsymbol{P}_{\boldsymbol{x}}\otimes\boldsymbol{P}_{\boldsymbol{x}}\otimes\boldsymbol{P}_{\boldsymbol{x}} = (\boldsymbol{x}\boldsymbol{x}^\top\otimes\boldsymbol{x}\boldsymbol{x}^\top\otimes\boldsymbol{x}\boldsymbol{x}^\top\otimes\boldsymbol{x}\boldsymbol{x}^\top) = (\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x})(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x})^\top,$$

it follows that the corresponding summand is written as:

$$\begin{aligned}&(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x})\mathbb{E}(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x})^\top(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})^\top(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}) \\ &= \mathbb{E}(\boldsymbol{x}^\top\boldsymbol{g})^8(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}) = 105(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}).\end{aligned} \tag{A.3}$$

Case 2: Two of $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \boldsymbol{B}_4$ are $\boldsymbol{P}_{\boldsymbol{x}}$ and the other two matrices are $\boldsymbol{P}_{\boldsymbol{x}^\perp}$.

First, we consider the sub-case where $\boldsymbol{B}_1 = \boldsymbol{B}_2 = \boldsymbol{P}_{\boldsymbol{x}}$ and $\boldsymbol{B}_3 = \boldsymbol{B}_4 = \boldsymbol{P}_{\boldsymbol{x}^\perp}$. Since $\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{g}$ and $\boldsymbol{x}^\top\boldsymbol{g}$ are independent, we can replace $\boldsymbol{x}^\top\boldsymbol{g}$ by $\boldsymbol{x}^\top\boldsymbol{g}'$ where $\boldsymbol{g}'$ is an independent copy of $\boldsymbol{g}$. Then the corresponding summand is written as

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{g}'}(\boldsymbol{g}'^\top\boldsymbol{x})^6\,\mathbb{E}_{\boldsymbol{g}}\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{g}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{g}\otimes\boldsymbol{x}\otimes\boldsymbol{x} = 15(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})(\mathbb{E}_{\boldsymbol{g}}\boldsymbol{g}\otimes\boldsymbol{g})\otimes\boldsymbol{x}\otimes\boldsymbol{x} \\ &= 15(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})\mathrm{vec}(\mathbb{E}_{\boldsymbol{g}}\boldsymbol{g}\boldsymbol{g}^\top)\otimes\boldsymbol{x}\otimes\boldsymbol{x} = 15(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})\mathrm{vec}(\mathbf{I}_d)\otimes\boldsymbol{x}\otimes\boldsymbol{x} \\ &= 15\,\mathrm{vec}(\boldsymbol{P}_{\boldsymbol{x}^\perp}\mathbf{I}_d\boldsymbol{P}_{\boldsymbol{x}^\perp})\otimes\boldsymbol{x}\otimes\boldsymbol{x} = 15\,\mathrm{vec}(\mathbf{I}_d - \boldsymbol{P}_{\boldsymbol{x}})\otimes\boldsymbol{x}\otimes\boldsymbol{x} \\ &= 15\Big(-\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x} + \sum_{l=1}^{d}\boldsymbol{e}_l\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{x}\Big).\end{aligned}$$

The summands corresponding to the other sub-cases of Case 2 are calculated similarly, and the partial summation of (A.2) for Case 2 is written as

$$\begin{aligned} &-90\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}+15\sum_{l=1}^{d}(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{e}_l+\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l+\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\\ &\quad+\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l+\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}+\boldsymbol{e}_l\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{x}). \end{aligned} \tag{A.4}$$

Case 3: $\boldsymbol{B}_1=\boldsymbol{B}_2=\boldsymbol{B}_3=\boldsymbol{B}_4=\boldsymbol{P}_{\boldsymbol{x}^\perp}$.

Again by the independence between $\boldsymbol{P}_{\boldsymbol{x}^\perp}\boldsymbol{g}$ and $\boldsymbol{x}^\top\boldsymbol{g}$, the corresponding summand is written as

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{g}'}(\boldsymbol{g}'^{\top}\boldsymbol{x})^4\,\mathbb{E}_{\boldsymbol{g}}(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})\\ &=3(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})\mathbb{E}_{\boldsymbol{g}}(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}). \end{aligned} \tag{A.5}$$

By plugging in the expression of $\mathbb{E}\,\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}$ in Lemma A.1, the right-hand side of (A.5) is written as

$$\begin{aligned} &\underbrace{3\,(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})\sum_{j,k=1}^{d}\boldsymbol{e}_j\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_k}_{(\S)}\\ &\quad+\underbrace{3\,(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})\sum_{j,k=1}^{d}\boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_k}_{(\S\S)}\\ &\quad+\underbrace{3\,(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})\sum_{j,k=1}^{d}\boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_j}_{(\S\S\S)}. \end{aligned} \tag{A.6}$$

The first term (§) in (A.6) is rewritten as

$$\begin{aligned} (\S)&=(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})[\mathrm{vec}(\mathbf{I}_d)\otimes\mathrm{vec}(\mathbf{I}_d)]\\ &=[(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})\mathrm{vec}(\mathbf{I}_d)]\otimes[(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})\mathrm{vec}(\mathbf{I}_d)]\\ &=\mathrm{vec}(\boldsymbol{P}_{\boldsymbol{x}^\perp})\otimes\mathrm{vec}(\boldsymbol{P}_{\boldsymbol{x}^\perp})=\mathrm{vec}(\mathbf{I}_d-\boldsymbol{P}_{\boldsymbol{x}})\otimes\mathrm{vec}(\mathbf{I}_d-\boldsymbol{P}_{\boldsymbol{x}})\\ &=\mathrm{vec}(\boldsymbol{P}_{\boldsymbol{x}})\otimes\mathrm{vec}(\boldsymbol{P}_{\boldsymbol{x}})+\mathrm{vec}(\mathbf{I}_d)\otimes\mathrm{vec}(\mathbf{I}_d)-\mathrm{vec}(\mathbf{I}_d)\otimes\mathrm{vec}(\boldsymbol{P}_{\boldsymbol{x}})-\mathrm{vec}(\boldsymbol{P}_{\boldsymbol{x}})\otimes\mathrm{vec}(\mathbf{I}_d)\\ &=\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}+\sum_{j,k=1}^{d}\boldsymbol{e}_j\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_k-\sum_{l=1}^{d}(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{e}_l+\boldsymbol{e}_l\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{x}). \end{aligned}$$

Similarly, (§§) and (§§§) are written as the sum of rank-1 tensors. Then applying these results to (A.6) provides

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}'}(\boldsymbol{g}'^{\top}\boldsymbol{x})^4\,\mathbb{E}_{\boldsymbol{g}}(\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp}\otimes\boldsymbol{P}_{\boldsymbol{x}^\perp})(\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g}\otimes\boldsymbol{g})\\
&= 9\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x} - 3\sum_{l=1}^{d}(\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{e}_l + \boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l + \boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\\
&\qquad + \boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l + \boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{e}_l\otimes\boldsymbol{x} + \boldsymbol{e}_l\otimes\boldsymbol{e}_l\otimes\boldsymbol{x}\otimes\boldsymbol{x})\\
&+ 3\sum_{j,k=1}^{d}(\boldsymbol{e}_j\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_k + \boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_k + \boldsymbol{e}_j\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_k\otimes\boldsymbol{e}_j).
\end{aligned} \tag{A.7}$$

The identity in (A.1) is then obtained by combining (A.3), (A.4) and (A.7) through (A.2).

B. Moment and tail bounds of random matrices

The following lemma, which provides a central moment bound on a standard normal variable, is a direct consequence of the Khintchine inequality (e.g. 60, Corollary 5.12).

LEMMA B.1 Let $g \sim \mathcal{N}(0,1)$. Then there exists a numerical constant C such that

$$(\mathbb{E}\,|g|^p)^{1/p} \le C\sqrt{p}, \quad \forall p \in \mathbb{N}.$$

We also use moment and tail bounds of random matrices in the spectral norm given by the noncommutative Rosenthal inequality [35, Theorem 0.4].

THEOREM B.2 (Noncommutative Rosenthal inequality [35, Theorem 0.4]). Let $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_M$ be independent random matrices with zero-mean. Then there exists a numerical constant C such that

$$\Big(\mathbb{E}\Big\|\sum_{m=1}^{M}\boldsymbol{Y}_m\Big\|^p\Big)^{1/p} \le C\Big[\sqrt{p}\Big(\Big\|\sum_{m=1}^{M}\mathbb{E}\,\boldsymbol{Y}_m\boldsymbol{Y}_m^*\Big\|^{1/2} \vee \Big\|\sum_{m=1}^{M}\mathbb{E}\,\boldsymbol{Y}_m^*\boldsymbol{Y}_m\Big\|^{1/2}\Big) \vee p\Big(\sum_{m=1}^{M}\mathbb{E}\,\|\boldsymbol{Y}_m\|^p\Big)^{1/p}\Big]$$

for all $1 \le p < \infty$.

Then the following lemma follows immediately from Theorem B.2.

LEMMA B.3 Let $\boldsymbol{g}_1,\dots,\boldsymbol{g}_M \in \mathbb{R}^d$ be independent copies of $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0},\mathbf{I}_d)$, $\boldsymbol{\lambda} = [\lambda_1,\dots,\lambda_M]^\top \in \mathbb{R}^M$, and $\nu \in (0,1)$. Then there exist numerical constants $C_1, C_2 > 0$ such that

$$\Big(\mathbb{E}\Big\|\frac{1}{M}\sum_{m=1}^{M}\lambda_m(\boldsymbol{g}_m\boldsymbol{g}_m^\top - \mathbf{I}_d)\Big\|^p\Big)^{1/p} \le C_1\|\boldsymbol{\lambda}\|_\infty\Big[M^{-1/2}\sqrt{pd} + M^{1/p-1}p(d+p)\Big] \tag{B.1}$$

for all $p \in \mathbb{N}$ and

$$\Big\|\frac{1}{M}\sum_{m=1}^{M}\lambda_m(\boldsymbol{g}_m\boldsymbol{g}_m^\top - \mathbf{I}_d)\Big\| \le \delta$$

holds with probability $1-\nu$ provided

$$M \ge C_2\left(\delta^{-1}\|\boldsymbol{\lambda}\|_\infty \vee \delta^{-2}\|\boldsymbol{\lambda}\|_\infty^2\right)\Big(d\log(M/\nu) \vee \log^2(M/\nu)\Big). \tag{B.2}$$

PROOF OF LEMMA B.3 We apply Theorem B.2 for $\boldsymbol{Y}_m = \lambda_m(\boldsymbol{g}_m\boldsymbol{g}_m^\top - \mathbf{I}_d)$ for $m = 1, \dots, M$. By the triangle inequality, we have

$$(\mathbb{E}\|\boldsymbol{Y}_m\|^p)^{1/p} \le \lambda_m + \lambda_m(\mathbb{E}\|\boldsymbol{g}_m\boldsymbol{g}_m^\top\|^p)^{1/p} = \lambda_m + \lambda_m(\mathbb{E}\|\boldsymbol{g}_m\|_2^{2p})^{1/p} \le C_1\lambda_m(d+p).$$

Here the last step follows since:

$$\|\|\boldsymbol{g}\|_2 - \sqrt{d}\|_{L_{2p}} \le C\sqrt{2p}\left\|\|\boldsymbol{g}\|_2 - \sqrt{d}\right\|_{\psi_2} \le C'\sqrt{p},$$

where $\|\cdot\|_{\psi_2}$ denotes the subgaussian norm. Therefore, we obtain

$$\left(\sum_{m=1}^M \mathbb{E}\,\|\boldsymbol{Y}_m\|^p\right)^{1/p} \le C_3\|\boldsymbol{\lambda}\|_\infty M^{1/p}(d+p). \tag{B.3}$$

Furthermore, the expectation of $\boldsymbol{Y}_m^2 = \lambda_m^2(\boldsymbol{g}_m\boldsymbol{g}_m^\top\boldsymbol{g}_m\boldsymbol{g}_m^\top - 2\boldsymbol{g}_m\boldsymbol{g}_m^\top + \mathbf{I}_d)$ is computed by using Lemma A.2 as $\mathbb{E}\boldsymbol{Y}_m^2 = \lambda_m^2(d+1)\mathbf{I}_d$. Therefore, it follows that:

$$\left\|\sum_{m=1}^M \mathbb{E}\,\boldsymbol{Y}_m^2\right\|^{1/2} = \sqrt{d+1}\|\boldsymbol{\lambda}\|_2 \le C_4\sqrt{Md}\|\boldsymbol{\lambda}\|_\infty. \tag{B.4}$$

Then (B.1) is obtained by plugging in (B.3) and (B.4) to Theorem B.2.

Next, by the Markov inequality, we have

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{M}\sum_{m=1}^M \boldsymbol{Y}_m\right\| > \delta\right) &\le \delta^{-p}\,\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^M \boldsymbol{Y}_m\right\|^p \\ &\le C_1\delta^{-p}\,\|\boldsymbol{\lambda}\|_\infty^p\left[M^{-1/2}\sqrt{pd} + M^{1/p-1}p(d+p)\right]^p. \end{aligned}\tag{B.5}$$

Let $p = \log(M/\nu)$. Then (B.2) implies that the right-hand side of (B.5) is upper-bounded by ν. This completes the proof.

C. Proof of Lemma 4.3

Let $\phi := \angle(\boldsymbol{u}_0, \boldsymbol{u}_\sharp)$ and $\psi := \angle(\boldsymbol{v}_0, \boldsymbol{v}_\sharp)$. Then

$$\inf_{\theta\in[0,2\pi)}\left\|\boldsymbol{u}_0\boldsymbol{v}_0^\top - e^{\mathrm{i}\theta}\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^\top\right\|_{\mathrm{F}}^2 = 2 - 2\cos\phi\cos\psi \le 2 - 2\cos^2(\phi\vee\psi) = 2\sin^2(\phi\vee\psi).$$

Therefore, it suffices to show

$$\sin(\phi\vee\psi) = \sin\phi\vee\sin\psi \le \frac{\delta}{\sqrt{2}}.$$

We will only show $\sin\phi \le \sqrt{\delta/2}$. The derivation of the other part is essentially the same due to symmetry. Without loss of generality, we assume $\|\boldsymbol{X}_\sharp\|_{\mathrm{F}} = 1$ (or equivalently $\sigma_\sharp = 1$).

Since $\boldsymbol{X}_\sharp$ is a scalar multiple of the most dominant eigenvector of $\mathbb{E}\,\boldsymbol{\Upsilon}$, we use the Davis–Kahan theorem [23] to bound the error in estimating $\boldsymbol{u}_\sharp$ as the dominant eigenvector of $\boldsymbol{\Upsilon}$. Among variations of the Davis–Kahan theorem, we use the version given in terms of the principal angle between two

subspaces. The following theorem states this result and is obtained by combining the argument of [29, Corollary 7.2.6] and the $\sin\theta$ theorem for any unitarily invariant norm [64].

THEOREM C.1 (Davis–Kahan $\sin\theta$ theorem). Let $\boldsymbol{A}, \boldsymbol{\Delta} \in \mathbb{C}^{n\times n}$ satisfy that $\boldsymbol{A}$ and $\boldsymbol{A}+\boldsymbol{\Delta}$ are positive semidefinite. Let $\boldsymbol{Q} \in \mathbb{C}^{n\times r}$ (resp. $\widehat{\boldsymbol{Q}} \in \mathbb{C}^{n\times r}$) denote the matrix whose columns are the eigenvectors of $\boldsymbol{A}$ (resp. $\boldsymbol{A}+\boldsymbol{\Delta}$) corresponding to the r-largest eigenvalues. Suppose that $\lambda_r(\boldsymbol{A}) > \lambda_{r+1}(\boldsymbol{A})$. If

$$\|\boldsymbol{\Delta}\| \le \frac{\lambda_r(\boldsymbol{A}) - \lambda_{r+1}(\boldsymbol{A})}{5},$$

then

$$\sin\angle(\operatorname{span}(\boldsymbol{Q}), \operatorname{span}(\widehat{\boldsymbol{Q}})) \le \frac{4\|\boldsymbol{\Delta}\|}{\lambda_r(\boldsymbol{A}) - \lambda_{r+1}(\boldsymbol{A})}.$$

To prove Lemma 4.3, we apply Theorem C.1 to $\boldsymbol{A} = \mathbb{E}\,\boldsymbol{\Upsilon}$ and $\boldsymbol{\Delta} = \boldsymbol{\Upsilon} - \mathbb{E}\,\boldsymbol{\Upsilon}$ with $r = 1$. Since

$$\boldsymbol{A} = \boldsymbol{u}_\sharp \boldsymbol{u}_\sharp^* + \left(1 + \frac{1}{M}\sum_{m=1}^{M} \xi_m\right)\mathbf{I}_{d_1},$$

it follows that:

$$\lambda_k(\boldsymbol{A}) = \lambda_k(\boldsymbol{u}_\sharp \boldsymbol{u}_\sharp^*) + 1 + \frac{1}{M}\sum_{m=1}^{M} \xi_m.$$

Therefore, we obtain

$$\lambda_1(\boldsymbol{A}) - \lambda_2(\boldsymbol{A}) = \lambda_1(\boldsymbol{u}_\sharp \boldsymbol{u}_\sharp^*) - \lambda_2(\boldsymbol{u}_\sharp \boldsymbol{u}_\sharp^*) = 1.$$

It remains to show

$$\|\boldsymbol{\Delta}\| \le \frac{\delta}{4\sqrt{2}}. \tag{C.1}$$

Let us first decompose $\boldsymbol{\Delta}$ into its noise-free portion and the remainder as

$$\boldsymbol{\Delta} = \frac{1}{M}\sum_{m=1}^{M} |\boldsymbol{b}_m^* \boldsymbol{v}_\sharp|^2 |\boldsymbol{a}_m^* \boldsymbol{u}_\sharp|^2 \boldsymbol{a}_m \boldsymbol{a}_m^* - \mathbb{E}\,|\boldsymbol{b}_m^* \boldsymbol{v}_\sharp|^2 |\boldsymbol{a}_m^* \boldsymbol{u}_\sharp|^2 \boldsymbol{a}_m \boldsymbol{a}_m^* + \frac{1}{M}\sum_{m=1}^{M} \xi_m (\boldsymbol{a}_m \boldsymbol{a}_m^* - \mathbf{I}_{d_1}).$$

Then (C.1) is implied by

$$\left\| \frac{1}{M}\sum_{m=1}^{M} |\boldsymbol{b}_m^* \boldsymbol{v}_\sharp|^2 |\boldsymbol{a}_m^* \boldsymbol{u}_\sharp|^2 \boldsymbol{a}_m \boldsymbol{a}_m^* - \mathbb{E}\,|\boldsymbol{b}_m^* \boldsymbol{v}_\sharp|^2 |\boldsymbol{a}_m^* \boldsymbol{u}_\sharp|^2 \boldsymbol{a}_m \boldsymbol{a}_m^* \right\| \le \frac{\delta}{8\sqrt{2}} \tag{C.2}$$

and

$$\left\| \frac{1}{M}\sum_{m=1}^{M} \xi_m (\boldsymbol{a}_m \boldsymbol{a}_m^* - \mathbf{I}_{d_1}) \right\| \le \frac{\delta}{8\sqrt{2}}. \tag{C.3}$$

Indeed, by Lemma B.3, (24) implies that (C.3) holds with probability $1 - \nu/2$ where $\nu = M^{-\alpha}$.

In the remainder of the proof, we show (22) implies (C.2) with probability $1 - \nu/2$. Let

$$\boldsymbol{Y}_m = \boldsymbol{Z}_m - \mathbb{E}\,\boldsymbol{Z}_m, \quad m = 1, \ldots, M,$$

where

$$\boldsymbol{Z}_m = |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^2 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^2 \boldsymbol{a}_m \boldsymbol{a}_m^*. \tag{C.4}$$

Then (C.2) is written as

$$\left\| \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{Y}_m \right\| \le \frac{\delta}{8\sqrt{2}}. \tag{C.5}$$

To show (C.5), we use the noncommutative Rosenthal inequality in Theorem B.2. By direct calculation, we obtain

$$\mathbb{E}\, \boldsymbol{Z}_m = \boldsymbol{u}_\natural \boldsymbol{u}_\natural^* + \mathbf{I}_{d_1}.$$

Next, by plugging in (C.4) into $\mathbb{E}\, \boldsymbol{Z}_m^* \boldsymbol{Z}_m$, we obtain

$$\mathbb{E}\, \boldsymbol{Z}_m^* \boldsymbol{Z}_m = \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{a}_m \boldsymbol{a}_m^*. \tag{C.6}$$

By decomposing the right-hand side of (C.6) with $\boldsymbol{P}_{\boldsymbol{u}} + \boldsymbol{P}_{\boldsymbol{u}^\perp} = \mathbf{I}_{d_1}$, $\mathbb{E}\, \boldsymbol{Z}_m^* \boldsymbol{Z}_m$ is rewritten as

$$\mathbb{E}\, \boldsymbol{Z}_m^* \boldsymbol{Z}_m = \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{P}_{\boldsymbol{u}_\natural} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural} \tag{C.7a}$$

$$+ \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{P}_{\boldsymbol{u}_\natural} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural} \tag{C.7b}$$

$$+ \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \tag{C.7c}$$

$$+ \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \tag{C.7d}$$

$$+ \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{P}_{\boldsymbol{u}_\natural} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \tag{C.7e}$$

$$+ \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural} \tag{C.7f}$$

$$+ \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{P}_{\boldsymbol{u}_\natural} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \tag{C.7g}$$

$$+ \mathbb{E}\, |\boldsymbol{b}_m^* \boldsymbol{v}_\natural|^4 |\boldsymbol{a}_m^* \boldsymbol{u}_\natural|^4 \boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural} \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{P}_{\boldsymbol{u}_\natural}. \tag{C.7h}$$

Since $\boldsymbol{u}_\natural^* \boldsymbol{a}_m$ and $\boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m$ are independent, which follows from $\boldsymbol{a}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{d_1})$, we can substitute $\boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \boldsymbol{a}_m$ by $\boldsymbol{P}_{\boldsymbol{u}_\natural^\perp} \breve{\boldsymbol{a}}_m$, where $\breve{\boldsymbol{a}}_m$ is an independent copy of $\boldsymbol{a}_m$. For a standard complex Gaussian random variable $\breve{g} \sim \mathcal{CN}(0, 1)$, we have

$$\mathbb{E}\, |\breve{g}|^2 = 1,\ \mathbb{E}\, |\breve{g}|^4 = 2,\ \mathbb{E}\, |\breve{g}|^6 = 6,\ \mathbb{E}\, |\breve{g}|^8 = 24.$$

Therefore, by using these even-order moments of $\mathcal{CN}(0,1)$ together with the independence between $\boldsymbol{a}_m$ and $\check{\boldsymbol{a}}_m$, we can compute (C.7a) to (C.7d) as follows:

$$\begin{aligned}
\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^4|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^4\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}_m\boldsymbol{a}_m^*\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}_m\boldsymbol{a}_m^*\boldsymbol{P}_{\boldsymbol{u}_\sharp} &= 48\boldsymbol{P}_{\boldsymbol{u}_\sharp},\\
\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^4|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^4\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}_m\check{\boldsymbol{a}}_m^*\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\check{\boldsymbol{a}}_m\boldsymbol{a}_m^*\boldsymbol{P}_{\boldsymbol{u}_\sharp} &= 12(d_1-1)\boldsymbol{P}_{\boldsymbol{u}_\sharp},\\
\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^4|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^4\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\check{\boldsymbol{a}}_m\boldsymbol{a}_m^*\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}_m\check{\boldsymbol{a}}_m^*\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} &= 12\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\\
\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^4|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^4\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\check{\boldsymbol{a}}_m\check{\boldsymbol{a}}_m^*\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\check{\boldsymbol{a}}_m\check{\boldsymbol{a}}_m^*\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} &= 4(d_1+1)\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}.
\end{aligned}$$

Furthermore, each of the remaining summands (C.7e) to (C.7h) vanishes since it has a factor given as a central Gaussian moments of an odd order.

Applying the above results to (C.7) provides

$$\mathbb{E}\,\boldsymbol{Z}_m^*\boldsymbol{Z}_m = (12d_1+36)\boldsymbol{P}_{\boldsymbol{u}_\sharp} + (4d_1+16)\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}.$$

Then, by the definition of $\boldsymbol{Y}_m$, we have

$$\mathbb{E}\,\boldsymbol{Y}_m^*\boldsymbol{Y}_m = \mathbb{E}\,\boldsymbol{Z}_m^*\boldsymbol{Z}_m - (\mathbb{E}\,\boldsymbol{Z}_m)^*(\mathbb{E}\,\boldsymbol{Z}_m) = (12d_1+32)\boldsymbol{P}_{\boldsymbol{u}_\sharp} + (4d_1+15)\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}.$$

Therefore, for $d_1 \geq 3$, we have

$$\left\|\sum_{m=1}^M \mathbb{E}\,\boldsymbol{Y}_m\boldsymbol{Y}_m^*\right\|^{1/2} \vee \left\|\sum_{m=1}^M \mathbb{E}\,\boldsymbol{Y}_m^*\boldsymbol{Y}_m\right\|^{1/2} \leq C_1\sqrt{Md_1}. \tag{C.8}$$

Next we compute the pth moment of the spectral norm. The pth moment is considered as the norm in L_p. Then by the triangle inequality in L_p, we obtain

$$\left(\mathbb{E}\,\|\boldsymbol{Y}_m\|^p\right)^{1/p} \leq \left(\mathbb{E}\,\|\boldsymbol{Z}_m\|^p\right)^{1/p} + \|\,\mathbb{E}\,\boldsymbol{Z}_m\| \leq \left(\mathbb{E}\,\|\boldsymbol{Z}_m\|^p\right)^{1/p} + 2. \tag{C.9}$$

Again by the triangle inequality, we obtain

$$\begin{aligned}
(\mathbb{E}\,\|\boldsymbol{Z}_m\|^p)^{1/p} &= \left[\mathbb{E}\left(|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^2|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^2\|\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}_m\|_2^2 + |\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^2|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^2\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}_m\|_2^2\right)^p\right]^{1/p}\\
&\leq \left(\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^{2p}|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^{4p}\right)^{1/p} + \left(\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^{2p}|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^{2p}\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\check{\boldsymbol{a}}_m\|_2^{2p}\right)^{1/p}\\
&\leq \left(\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^{2p}\right)^{1/p}\left(\mathbb{E}\,|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^{4p}\right)^{1/p} + \left(\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^{2p}\right)^{1/p}\left(\mathbb{E}\,|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^{2p}\right)^{1/p}\left(\mathbb{E}\,\|\check{\boldsymbol{a}}_m\|_2^{2p}\right)^{1/p}.
\end{aligned} \tag{C.10}$$

Since $\boldsymbol{a}_m^*\boldsymbol{u}_\sharp \sim \mathcal{CN}(0,1)$ and $\boldsymbol{b}_m^*\boldsymbol{v}_\sharp \sim \mathcal{CN}(0,1)$, by Lemma B.1, there exists a numerical constant C_2 such that

$$(\mathbb{E}\,|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^p)^{1/p} = (\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^p)^{1/p} \leq C_2\sqrt{p}.$$

Since $2\|\check{\boldsymbol{a}}_m\|_2^2$ is a chi-square random variable of the degree-of-freedom $2d_1$, we obtain

$$\left(\mathbb{E}\,\|\check{\boldsymbol{a}}_m\|_2^{2p}\right)^{1/p} \leq C_3(d_1+p), \quad \forall p \geq 2.$$

Applying these upper estimates of the moments to (C.10) then to (C.9) provides

$$(\mathbb{E}\,\|\boldsymbol{Y}_m\|^p)^{1/p} \le C_4(p^2 d_1 + p^3),$$

which implies

$$p\left(\sum_{m=1}^{M} \mathbb{E}\,\|\boldsymbol{Y}_m\|^p\right)^{1/p} \le C_4 M^{1/p}(p^3 d_1 + p^4). \tag{C.11}$$

By applying (C.8) and (C.11) to Theorem B.2, we obtain

$$\left(\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{Y}_m\right\|^p\right)^{1/p} \le C_5\left[\sqrt{\frac{pd_1}{M}} + \frac{M^{1/p}(p^3 d_1 + p^4)}{M}\right] \tag{C.12}$$

for all $p \ge 2$ and $d_1 \ge 3$.

Finally, similar to [27, Proposition 7.11], we derive a tail bound from moment bounds. It follows from the Markov inequality that:

$$\mathbb{P}\left(\left\|\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{Y}_m\right\| > \frac{\delta}{8\sqrt{2}}\right) \le \left(\frac{8\sqrt{2}}{\delta}\right)^p \mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{Y}_m\right\|^p. \tag{C.13}$$

By plugging in (C.12) to (C.13), it follows that (C.5) holds with probability ν provided that

$$C_6\left[\sqrt{\frac{pd_1}{M}} + \frac{M^{1/p}(p^3 d_1 + p^4)}{M}\right] \le \delta \nu^{1/p}.$$

Then we set $p = \log(M/\nu)$ so that (22) implies that (22) holds with probability $1 - \nu/2$. Therefore, the probability for violating (C.5) becomes $\nu = M^{-\alpha}$. This completes the proof.

D. Proof of Lemma 4.5

To simplify notation, let

$$\boldsymbol{Z}_m := \langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\sharp\rangle^2 \boldsymbol{\Phi}_m \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{\Phi}_m^\top, \quad m = 1, \dots, M.$$

Then $\boldsymbol{\Upsilon}$ is written as

$$\boldsymbol{\Upsilon} = \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{Z}_m + \underbrace{\xi_m \boldsymbol{\Phi}_m \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{\Phi}_m^\top}_{(\flat)}). \tag{D.1}$$

We derive the expectation of $\boldsymbol{\Upsilon}$ in the following steps: first the expectation of the noise part ($\flat$) in (D.1) is computed as

$$\mathbb{E}\,\xi_m \boldsymbol{\Phi}_m \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{\Phi}_m^\top = \xi_m \mathrm{tr}(\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top)\mathbf{I}_{d_1} = r\xi_m \mathbf{I}_{d_1}. \tag{D.2}$$

Next we compute $\mathbb{E}\,\boldsymbol{Z}_m$ by using Lemma A.1. Let $\boldsymbol{x}_\sharp = \mathrm{vec}(\boldsymbol{X}_\sharp)$ and $\boldsymbol{\phi}_m = \mathrm{vec}(\boldsymbol{\Phi}_m)$ for $m = 1, \dots, M$. Then $\boldsymbol{Z}_m$ is rewritten as

$$\boldsymbol{Z}_m = (\mathrm{tr} \otimes \mathbf{I}_{d_1})\Big[(\widehat{\boldsymbol{V}}^\top \otimes \mathbf{I}_{d_1})\langle \boldsymbol{\phi}_m, \boldsymbol{x}_\sharp\rangle^2 \boldsymbol{\phi}_m \boldsymbol{\phi}_m^\top (\widehat{\boldsymbol{V}} \otimes \mathbf{I}_{d_1})\Big].$$

Since the partial trace operator is linear, the expectation of $\boldsymbol{Z}_m$ is written as

$$\begin{aligned}\mathbb{E}\boldsymbol{Z}_m &= (\mathrm{tr}\otimes \mathbf{I}_{d_1})\Big[(\widehat{\boldsymbol{V}}^\top \otimes \mathbf{I}_{d_1})\mathbb{E}\langle \boldsymbol{\phi}_m, \boldsymbol{x}_\sharp\rangle^2 \boldsymbol{\phi}_m\boldsymbol{\phi}_m^\top (\widehat{\boldsymbol{V}}\otimes \mathbf{I}_{d_1})\Big] \\ &= (\mathrm{tr}\otimes \mathbf{I}_{d_1})\Big[(\widehat{\boldsymbol{V}}^\top \otimes \mathbf{I}_{d_1})(2\boldsymbol{x}_\sharp\boldsymbol{x}_\sharp^\top + \|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2 \mathbf{I}_{d_1d_2})(\widehat{\boldsymbol{V}}\otimes \mathbf{I}_{d_1})\Big] \\ &= 2\boldsymbol{X}_\sharp \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{X}_\sharp^\top + r\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2 \mathbf{I}_{d_1},\end{aligned} \tag{D.3}$$

where the second identity follows from Lemma A.1. Then by combining (D.2) and (D.3), the expectation of $\boldsymbol{\Upsilon}$ is written as

$$\mathbb{E}\,\boldsymbol{\Upsilon} = 2\boldsymbol{X}_\sharp \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{X}_\sharp^\top + \left(r\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2 + \frac{r}{M}\sum_{m=1}^M \xi_m\right)\mathbf{I}_{d_1}. \tag{D.4}$$

It follows from (25) that $\boldsymbol{X}_\sharp \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{X}_\sharp^\top$ in the right-hand side of (D.4) has rank-r and its invariant space coincides with that of $\boldsymbol{X}_\sharp\boldsymbol{X}_\sharp^\top = \boldsymbol{U}_\sharp\boldsymbol{\Sigma}_\sharp^2\boldsymbol{U}_\sharp^\top$. The inclusion of the former subspace to the latter is obvious from the construction. Furthermore, the rank of $\boldsymbol{X}_\sharp \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{X}_\sharp^\top$ is at most r. Indeed, the rth largest singular value of $\boldsymbol{X}_\sharp \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{X}_\sharp^\top$ satisfies

$$\begin{aligned}&\sigma_r(\boldsymbol{X}_\sharp \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{X}_\sharp^\top) \geq \sigma_r(\boldsymbol{X}_\sharp)^2\sigma_r(\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{V}_\sharp\boldsymbol{V}_\sharp^\top) \\ &\geq \sigma_r(\boldsymbol{X}_\sharp)^2\left(\sigma_r(\boldsymbol{V}_\sharp\boldsymbol{V}_\sharp^\top) - \|(\mathbf{I}_{d_2} - \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top)\boldsymbol{V}_\sharp\boldsymbol{V}_\sharp^\top\|\right) \geq (1-\delta_{\mathrm{in}})\sigma_r(\boldsymbol{X}_\sharp)^2,\end{aligned}$$

where the last step follows from (25). Therefore, we deduce that $\boldsymbol{X}_\sharp \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{X}_\sharp^\top$ and $\boldsymbol{X}_\sharp\boldsymbol{X}_\sharp^\top$ have the same invariant subspace.

Recall that the columns of $\boldsymbol{U}_0$ are the eigenvectors of $\boldsymbol{\Upsilon}$ corresponding to the r-largest eigenvalues. Furthermore, the subspace spanned by the top r eigenvectors of $\mathbb{E}\,\boldsymbol{\Upsilon}$ is the same to the column space of $\boldsymbol{U}_\sharp$. Therefore, the Davis–Kahan theorem (Theorem C.1) provides an upper bound for the estimation error measured by the principal angle between subspaces (the left-hand side of (28)). To this end, we apply Theorem C.1 to $\boldsymbol{A} = \mathbb{E}\,\boldsymbol{\Upsilon}$ and $\boldsymbol{\Delta} = \boldsymbol{\Upsilon} - \mathbb{E}\,\boldsymbol{\Upsilon}$ as shown below.

Since the spectral gap in $\boldsymbol{A}$ satisfies

$$\lambda_r(\boldsymbol{A}) - \lambda_{r+1}(\boldsymbol{A}) = \lambda_r(\mathbb{E}\,\boldsymbol{\Upsilon}) - \lambda_{r+1}(\mathbb{E}\,\boldsymbol{\Upsilon}) = \lambda_r(2\boldsymbol{X}_\sharp \widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top \boldsymbol{X}_\sharp^\top) \geq 2(1-\delta_{\mathrm{in}})[\sigma_r(\boldsymbol{X}_\sharp)]^2,$$

the error bound in (28) is obtained by Theorem C.1 provided that

$$\|\boldsymbol{\Delta}\| = \|\boldsymbol{\Upsilon} - \mathbb{E}\,\boldsymbol{\Upsilon}\| \leq \frac{(1-\delta_{\mathrm{in}})\delta_{\mathrm{out}}\sigma_r(\boldsymbol{X}_\sharp)^2}{2}. \tag{D.5}$$

By the triangle inequality, we obtain a sufficient condition for (D.5) given by

$$\left\|\frac{1}{M}\sum_{m=1}^M (\boldsymbol{Z}_m - \mathbb{E}\,\boldsymbol{Z}_m)\right\| \leq \frac{(1-\delta_{\mathrm{in}})\delta_{\mathrm{out}}\sigma_r(\boldsymbol{X}_\sharp)^2}{4} \tag{D.6}$$

and

$$\left\| \frac{1}{M} \sum_{m=1}^{M} \xi_m \left(\boldsymbol{\Phi}_m \widehat{\boldsymbol{V}} \widehat{\boldsymbol{V}}^\top \boldsymbol{\Phi}_m^\top - r \mathbf{I}_{d_1} \right) \right\| \le \frac{(1-\delta_{\text{in}})\delta_{\text{out}} \sigma_r(\boldsymbol{X}_\sharp)^2}{4}. \tag{D.7}$$

In the remainder, we show that (D.6) and (D.6) hold with high probability when the conditions in (26) and (29) are satisfied. First, by Lemma B.3, it follows from (29) that (D.7) holds with probability $1 - M^{-\alpha}/2$. Then it remains to show that (D.6) holds with probability $1 - M^{-\alpha}/2$ when (26) is satisfied. By the Markov inequality,

$$\begin{aligned} &\mathbb{P}\left(\left\| \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{Z}_m - \mathbb{E}\, \boldsymbol{Z}_m) \right\| > \frac{(1-\delta_{\text{in}})\delta_{\text{out}} \sigma_r(\boldsymbol{X}_\sharp)^2}{4} \right) \\ &\le \left(\frac{4}{(1-\delta_{\text{in}})\delta_{\text{out}} \sigma_r(\boldsymbol{X}_\sharp)^2} \right)^p \cdot \mathbb{E} \left\| \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{Z}_m - \mathbb{E}\boldsymbol{Z}_m) \right\|^p \end{aligned}$$

for any $p > 0$. Therefore, (D.6) holds with probability $1 - M^{-\alpha}/2$ if

$$\underbrace{\left(\mathbb{E} \left\| \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{Z}_m - \mathbb{E}\boldsymbol{Z}_m) \right\|^p \right)^{1/p}}_{(\ddagger)} \le \frac{(1-\delta_{\text{in}})\delta_{\text{out}} \sigma_r(\boldsymbol{X}_\sharp)^2 M^{-\alpha/p}}{4}. \tag{D.8}$$

To get an upper estimate of (‡) in (D.8), we apply the noncommutative Rosenthal inequality (Theorem B.2) to $\boldsymbol{Y}_m = \boldsymbol{Z}_m - \mathbb{E}\boldsymbol{Z}_m$ for $m = 1, \dots, M$. The first step is to compute the expectation of $\boldsymbol{Y}_m^2$ as follows: let $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3, \boldsymbol{Q}_4 \in \mathbb{R}^{d_1 \times d_2}$. Note that each entry of $\boldsymbol{Q}_1 \boldsymbol{Q}_2^\top \boldsymbol{Q}_3 \boldsymbol{Q}_4^\top$ is given as a linear combination of the entries of $\operatorname{vec}(\boldsymbol{Q}_1) \otimes \operatorname{vec}(\boldsymbol{Q}_2) \otimes \operatorname{vec}(\boldsymbol{Q}_3) \otimes \operatorname{vec}(\boldsymbol{Q}_4)$. Therefore, there exists a linear map $\mathcal{R} : \mathbb{R}^{(d_1 d_2)^4} \to \mathbb{R}^{d_1 \times d_1}$ that satisfies

$$\mathcal{R}[\operatorname{vec}(\boldsymbol{Q}_1) \otimes \operatorname{vec}(\boldsymbol{Q}_2) \otimes \operatorname{vec}(\boldsymbol{Q}_3) \otimes \operatorname{vec}(\boldsymbol{Q}_4)] = \boldsymbol{Q}_1 \boldsymbol{Q}_2^\top \boldsymbol{Q}_3 \boldsymbol{Q}_4^\top.$$

We also define

$$\boldsymbol{T}_m := \langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\sharp \rangle^4 [\operatorname{vec}(\boldsymbol{\Phi}_m \widehat{\boldsymbol{V}}) \otimes \operatorname{vec}(\boldsymbol{\Phi}_m \widehat{\boldsymbol{V}}) \otimes \operatorname{vec}(\boldsymbol{\Phi}_m \widehat{\boldsymbol{V}}) \otimes \operatorname{vec}(\boldsymbol{\Phi}_m \widehat{\boldsymbol{V}})].$$

Then $\boldsymbol{Z}_m^2$ is written as $\boldsymbol{Z}_m^2 = \mathcal{R}(\boldsymbol{T}_m)$. Since $\mathrm{vec}(\boldsymbol{\Phi}_m\widehat{\boldsymbol{V}}) = (\widehat{\boldsymbol{V}}^\top \otimes \mathbf{I}_{d_1})\mathrm{vec}(\boldsymbol{\Phi}_m)$, it follows that $\mathbb{E}\boldsymbol{T}_m$ is written as:

$$
\begin{aligned}
\mathbb{E}\boldsymbol{T}_m &= [(\widehat{\boldsymbol{V}}^\top \otimes \mathbf{I}_{d_1}) \otimes (\widehat{\boldsymbol{V}}^\top \otimes \mathbf{I}_{d_1}) \otimes (\widehat{\boldsymbol{V}}^\top \otimes \mathbf{I}_{d_1}) \otimes (\widehat{\boldsymbol{V}}^\top \otimes \mathbf{I}_{d_1})]\mathbb{E}(\boldsymbol{\phi}_m^\top \mathrm{vec}(\boldsymbol{X}_\sharp))^4(\boldsymbol{\phi}_m \otimes \boldsymbol{\phi}_m \otimes \boldsymbol{\phi}_m \otimes \boldsymbol{\phi}_m) \\
&= 24\,[\mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}})] \\
&\quad + 12\sum_{k_1=1}^{d}\sum_{k_2=1}^{d}\Big[\mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \\
&\qquad\quad + \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \\
&\qquad\quad + \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \\
&\qquad\quad + \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \\
&\qquad\quad + \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \\
&\qquad\quad + \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}})\Big] \\
&\quad + 3\sum_{j_1,k_1=1}^{d_1}\sum_{j_2,k_2=1}^{d_2}\Big[\mathrm{vec}(\boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \\
&\qquad\quad + \mathrm{vec}(\boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \\
&\qquad\quad + \mathrm{vec}(\boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}) \otimes \mathrm{vec}(\boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}})\Big],
\end{aligned}
$$

where Lemma A.3 is used to compute $\mathbb{E}\boldsymbol{T}_m$ in the second step. Also by the linearity of the map $\mathcal{R}$, it follows that:

$$
\begin{aligned}
\mathbb{E}\boldsymbol{Z}_m^2 &= \mathcal{R}(\mathbb{E}\boldsymbol{T}_m) \\
&= 24\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top \\
&\quad + 12\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2\sum_{l_1=1}^{d}\sum_{l_2=1}^{d}\Big[\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top\boldsymbol{e}_{l_1}\widetilde{\boldsymbol{e}}_{l_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{l_2}\boldsymbol{e}_{l_1}^\top + \boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{l_2}\boldsymbol{e}_{l_1}^\top\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{l_2}\boldsymbol{e}_{l_1}^\top \\
&\qquad\quad + \boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{l_2}\boldsymbol{e}_{l_1}^\top\boldsymbol{e}_{l_1}\widetilde{\boldsymbol{e}}_{l_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top + \boldsymbol{e}_{l_1}\widetilde{\boldsymbol{e}}_{l_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{l_2}\boldsymbol{e}_{l_1}^\top \\
&\qquad\quad + \boldsymbol{e}_{l_1}\widetilde{\boldsymbol{e}}_{l_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top\boldsymbol{e}_{l_1}\widetilde{\boldsymbol{e}}_{l_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top + \boldsymbol{e}_{l_1}\widetilde{\boldsymbol{e}}_{l_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{l_2}\boldsymbol{e}_{l_1}^\top\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top\Big] \\
&\quad + 3\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^4\sum_{j_1,k_1=1}^{d_1}\sum_{j_2,k_2=1}^{d_2}\Big[\boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{j_2}\boldsymbol{e}_{j_1}^\top\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{k_2}\boldsymbol{e}_{k_1}^\top \\
&\qquad\quad + \boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{k_2}\boldsymbol{e}_{k_1}^\top\boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{k_2}\boldsymbol{e}_{k_1}^\top + \boldsymbol{e}_{j_1}\widetilde{\boldsymbol{e}}_{j_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{k_2}\boldsymbol{e}_{k_1}^\top\boldsymbol{e}_{k_1}\widetilde{\boldsymbol{e}}_{k_2}^\top\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\widetilde{\boldsymbol{e}}_{j_2}\boldsymbol{e}_{j_1}^\top\Big].
\end{aligned}
$$

After direct calculation, the above expression for $\mathbb{E}\boldsymbol{Z}_m^2$ simplifies to

$$\begin{aligned}\mathbb{E}\boldsymbol{Z}_m^2 &= 24\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top \\ &\quad + 12(2r+d_1+2)\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top + 12\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2\|\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\|_{\mathrm{F}}^2\mathbf{I}_{d_1} \\ &\quad + 3\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^4 r(r+d_1+1)\mathbf{I}_{d_1}.\end{aligned} \tag{D.9}$$

Then, by combining (D.3) and (D.9), we obtain

$$\begin{aligned}\mathbb{E}\boldsymbol{Y}_m^2 &= \mathbb{E}\boldsymbol{Z}_m^2 - (\mathbb{E}\boldsymbol{Z}_m)^2 \\ &= 20\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top + 4(5r+3d_1+6)\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top \\ &\quad + \left(12\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2\|\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\|_{\mathrm{F}}^2 + \|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^4 r(2r+3d_1+3)\right)\mathbf{I}_{d_1}.\end{aligned}$$

Therefore, the spectral norm of $\mathbb{E}\boldsymbol{Y}_m^2$ is upper-bounded by

$$\|\mathbb{E}\boldsymbol{Y}_m^2\| \le 20\|\boldsymbol{X}_\sharp\|^4 + 4(5r+3d_1+6)\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2\|\boldsymbol{X}_\sharp\|^2 + 12\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2\|\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\|_{\mathrm{F}}^2 + r(2r+3d_1+3)\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^4.$$

Collecting the results for $m = 1,\dots,M$ gives

$$\left\|\sum_{m=1}^M \mathbb{E}\boldsymbol{Y}_m^2\right\|^{1/2} \le Cr^{3/2}\sqrt{Md_1}\|\boldsymbol{X}_\sharp\|^2. \tag{D.10}$$

Moreover, by applying the triangle inequality in L_p twice to (D.3), we obtain

$$\begin{aligned}(\mathbb{E}\,\|\boldsymbol{Y}_m\|^p)^{1/p} &\le \underbrace{\left[\mathbb{E}\left(\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle^2\|\boldsymbol{\Phi}_m\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{\Phi}_m^\top - r\mathbf{I}_{d_1}\|\right)^p\right]^{1/p}}_{(\natural)} \\ &\quad + \underbrace{r\left[\mathbb{E}\left(\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle^2 - \|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2\right)^p\right]^{1/p}}_{(\natural\natural)} + \underbrace{2\|\boldsymbol{X}_\sharp\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{X}_\sharp^\top\|}_{(\natural\natural\natural)}.\end{aligned} \tag{D.11}$$

By the Cauchy–Schwarz inequality in L_2, the first term ($\natural$) on the right-hand side of (D.11) is upper-bounded by

$$(\natural) \le \left(\mathbb{E}\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle^{4p}\right)^{1/2p}\cdot\left(\mathbb{E}\,\|\boldsymbol{\Phi}_m\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{\Phi}_m^\top - r\mathbf{I}_{d_1}\|^{2p}\right)^{1/2p}.$$

Since $\langle\boldsymbol{X}_\sharp,\boldsymbol{\Phi}_m\rangle \sim \mathcal{N}(0,\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2)$, by Lemma B.1, we have

$$\left(\mathbb{E}\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle^{4p}\right)^{1/2p} \le Cp\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2.$$

Then it follows from $\mathrm{vec}(\boldsymbol{\Phi}_m\widehat{\boldsymbol{V}}) = (\widehat{\boldsymbol{V}}^\top\otimes\mathbf{I}_{d_1})\boldsymbol{\phi}_m$ that

$$\mathbb{E}\mathrm{vec}(\boldsymbol{\Phi}_m\widehat{\boldsymbol{V}})\mathrm{vec}(\boldsymbol{\Phi}_m\widehat{\boldsymbol{V}})^\top = \mathbb{E}(\widehat{\boldsymbol{V}}^\top\otimes\mathbf{I}_{d_1})\boldsymbol{\phi}_m\boldsymbol{\phi}_m^\top(\widehat{\boldsymbol{V}}\otimes\mathbf{I}_{d_1})\boldsymbol{\phi}_m = \widehat{\boldsymbol{V}}^\top\widehat{\boldsymbol{V}}\otimes\mathbf{I}_{d_1} = \mathbf{I}_{d_2d_1},$$

which implies that $\boldsymbol{\Phi}_1\widehat{\boldsymbol{V}}, \ldots, \boldsymbol{\Phi}_M\widehat{\boldsymbol{V}} \in \mathbb{R}^{d_1 \times r}$ are independent copies of a standard i.i.d. Gaussian matrix. Thus, Lemma B.3 implies

$$\left(\mathbb{E}\|\boldsymbol{\Phi}_m\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top\boldsymbol{\Phi}_m^\top - r\mathbf{I}_{d_1}\|^{2p}\right)^{1/2p} \leq C\left[\sqrt{rpd_1} + r^{1/2p}p\left(d_1+p\right)\right].$$

Then, by the triangle inequality in L_p and Lemma B.1, (♮♮) is upper-bounded as

$$(\natural\natural) \leq r\left(\mathbb{E}\langle\boldsymbol{\Phi}_m, \boldsymbol{X}_\sharp\rangle^{2p}\right)^{1/p} + r\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2 \leq C'rp\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2.$$

The last term is trivially upper-bounded by $(\natural\natural\natural) \leq \|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2$.

By collecting the above results, we obtain that the L_p-norm of $\|\boldsymbol{Y}_m\|$ is upper-bounded by

$$\left(\mathbb{E}\,\|\boldsymbol{Y}_m\|^p\right)^{1/p} \leq C_1\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2 p\left(r + \sqrt{rpd_1} + r^{1/2p}p\left(d_1+p\right)\right). \tag{D.12}$$

Then, by applying (D.10) and (D.12) to Theorem B.2, we obtain that (‡) in (D.8) is upper-bounded by

$$\begin{aligned}&\left(\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{Z}_m - \mathbb{E}\boldsymbol{Z}_m)\right\|^p\right)^{1/p}\\ &\quad\leq C_3\|\boldsymbol{X}_\sharp\|^2\left(r^{3/2}M^{-1/2}\sqrt{pd_1} + M^{1/p-1}p^2r\left(r + \sqrt{rpd_1} + r^{1/2p}p\left(d_1+p\right)\right)\right).\end{aligned}$$

Finally, we choose $p = \log(M/M^{-\alpha}) = (\alpha+1)\log M$. Then (26) implies (D8). This completes the proof.

E. Proof of Lemma 5.1

Since $\widehat{\boldsymbol{X}}$ is a minimizer to (3), it satisfies

$$\mathrm{Im}\,\langle\boldsymbol{X}_0, \widehat{\boldsymbol{X}}\rangle = 0. \tag{E.1}$$

Then by (E.1) and (30) together with the fact that $\mathrm{rank}(\boldsymbol{X}_\sharp) = 1$, we have

$$\mathrm{Im}\,\langle\boldsymbol{X}_0, \boldsymbol{H}\rangle = 0,$$

where $\boldsymbol{H} = \widehat{\boldsymbol{X}} - \boldsymbol{X}_\sharp$. Let $\mathcal{P}_{\boldsymbol{X}_\sharp}$ denote the orthogonal projection onto $\mathbb{C}\boldsymbol{X}_\sharp$, that is

$$\mathcal{P}_{\boldsymbol{X}_\sharp} : \boldsymbol{M} \mapsto \frac{\boldsymbol{X}_\sharp\langle\boldsymbol{X}_\sharp, \boldsymbol{M}\rangle}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2}.$$

Then it follows that:

$$0 = |\mathrm{Im}\,\langle\boldsymbol{X}_0, \boldsymbol{H}\rangle| \geq |\mathrm{Im}\,\langle\mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{X}_0), \boldsymbol{H}\rangle| - |\mathrm{Im}\,\langle\boldsymbol{X}_0 - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{X}_0), \boldsymbol{H}\rangle|,$$

which is rearranged as

$$|\mathrm{Im}\,\langle\mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{X}_0), \boldsymbol{H}\rangle| \leq |\mathrm{Im}\,\langle\boldsymbol{X}_0 - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{X}_0), \boldsymbol{H}\rangle|. \tag{E..2}$$

By (30), the left-hand side of (E.2) is bounded from below as

$$|\mathrm{Im}\,\langle \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{X}_0), \boldsymbol{H}\rangle| = \frac{\left|\mathrm{Im}(\langle \boldsymbol{X}_0, \boldsymbol{X}_\sharp\rangle\langle \boldsymbol{X}_\sharp, \boldsymbol{H}\rangle)\right|}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}^2}$$
$$= \frac{\langle \boldsymbol{X}_0, \boldsymbol{X}_\sharp\rangle}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}} \cdot \frac{\left|\mathrm{Im}\langle \boldsymbol{X}_\sharp, \boldsymbol{H}\rangle\right|}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}} \geq \sqrt{1-\delta^2} \cdot \|\boldsymbol{X}_0\|_{\mathrm{F}} \cdot \frac{\left|\mathrm{Im}\langle \boldsymbol{X}_\sharp, \boldsymbol{H}\rangle\right|}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}}.$$

Since the linear operator $\iota : \boldsymbol{M} \mapsto \boldsymbol{M} - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{M})$ is self-adjoint and idempotent, the right-hand side of (E..2) is bounded from above as

$$|\mathrm{Im}\,\langle \boldsymbol{X}_0 - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{X}_0), \boldsymbol{H}\rangle| = |\mathrm{Im}\,\langle \boldsymbol{X}_0 - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{X}_0), \boldsymbol{H} - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{H})\rangle|$$
$$\leq \|\boldsymbol{X}_0 - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{X}_0)\|_{\mathrm{F}} \cdot \|\boldsymbol{H} - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{H})\|_{\mathrm{F}} \leq \delta\, \|\boldsymbol{X}_0\|_{\mathrm{F}} \cdot \|\boldsymbol{H} - \mathcal{P}_{\boldsymbol{X}_\sharp}(\boldsymbol{H})\|_{\mathrm{F}}.$$

Applying the above bounds to (E.2) completes the proof.

F. Proof of Lemma 5.4

The following lemma provides a tail probability of the product of two jointly Gaussian variables.

LEMMA F.1 (A variation of [6, Lemma 5]). Let g_1, g_2 be random variables that satisfy

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

Then for all $t > 0$

$$\mathbb{P}(g_1 g_2 > t) \geq \frac{2}{\pi} \cos^{-1}\left(\frac{\sqrt{3-\rho}}{2}\right) \exp\left(-\frac{2t}{1+\rho}\right). \tag{F.1}$$

PROOF OF LEMMA F.1 Let w_1 and w_2 be independent copies of a standard normal random variable following $\mathcal{N}(0,1)$. Then g_1 and g_2 are written as

$$g_1 = \sqrt{\frac{1+\rho}{2}} w_1 + \sqrt{\frac{1-\rho}{2}} w_2 \quad \text{and} \quad g_2 = \sqrt{\frac{1-\rho}{2}} w_1 - \sqrt{\frac{1-\rho}{2}} w_2.$$

With this representation, we have

$$\mathbb{P}(g_1 g_2 > t) = \mathbb{P}\left(\frac{1+\rho}{2} w_1^2 - \frac{1-\rho}{2} w_2^2 > t\right) = \mathbb{P}\left(\frac{\rho-1}{2} + \frac{w_1^2}{w_1^2+w_2^2} > \frac{t}{w_1^2+w_2^2}\right).$$

Since $w_1^2/(w_1^2+w_2^2)$ and $1/(w_1^2+w_2^2)$ respectively depend only on the direction and the ℓ_2 norm of the standard normal random vector $[w_1, w_2]^\top$, they are mutually independent. Furthermore, $R = w_1^2 + w_2^2$ follows the exponential distribution with mean $1/2$ and $w_1/\sqrt{w_1^2+w_2^2}$ is written as $\cos\theta$ where θ is a

uniform random variable on $[0, 2\pi)$. Then it follows that:

$$\begin{aligned}\mathbb{P}\left(\frac{\rho-1}{2}+\frac{w_1^2}{w_1^2+w_2^2}>\frac{t}{w_1^2+w_2^2}\right) &\geq \mathbb{P}\left(\frac{w_1^2}{w_1^2+w_2^2}\geq\frac{3-\rho}{4}\text{ and }\frac{1+\rho}{4}>\frac{t}{w_1^2+w_2^2}\right)\\ &= \mathbb{P}\left(\frac{w_1^2}{w_1^2+w_2^2}\geq\frac{3-\rho}{4}\right)\mathbb{P}\left(w_1^2+w_2^2>\frac{4t}{1+\rho}\right)\\ &= \mathbb{P}\left(\cos^2\theta\geq\frac{3-\rho}{4}\right)\mathbb{P}\left(R>\frac{4t}{1+\rho}\right). \end{aligned} \tag{F.2}$$

The lower bound in (F.1) is obtained by computing the probabilities in (F.2).

We apply Lemma F.1 for $g_1 = \langle \boldsymbol{X}_\sharp, \boldsymbol{\Phi}\rangle/\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}$, $g_2 = \langle \boldsymbol{\Phi}, \boldsymbol{H}\rangle/\|\boldsymbol{H}\|_{\mathrm{F}}$ and $t = \tau'$. Since the probability in (F.1) is a monotone increasing function in ρ, to get a lower bound on the tail probability, it suffices to compute a lower estimate of ρ.

Let $\boldsymbol{X}_\sharp = \boldsymbol{U}_\sharp \boldsymbol{\Sigma}_\sharp \boldsymbol{V}_\sharp^\top$ denote the SVD of $\boldsymbol{X}_\sharp$. Let $\sigma_1, \ldots, \sigma_r$ denote the singular values of $\boldsymbol{X}_\sharp$ in the non-increasing order. Then $\|\boldsymbol{X}_\sharp\|_* = \sum_{k=1}^r \sigma_k$. By the triangle inequality, we have

$$\rho = \frac{\langle \boldsymbol{X}_\sharp, \boldsymbol{H}\rangle}{\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}\|\boldsymbol{H}\|_{\mathrm{F}}} \geq \underbrace{\frac{\langle \|\boldsymbol{X}_\sharp\|_* \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top, \boldsymbol{H}\rangle}{r\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}\|\boldsymbol{H}\|_{\mathrm{F}}}}_{(\flat\flat)} - \underbrace{\frac{|\langle r\boldsymbol{X}_\sharp - \|\boldsymbol{X}_\sharp\|_* \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top, \boldsymbol{H}\rangle|}{r\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}\|\boldsymbol{H}\|_{\mathrm{F}}}}_{(\flat\flat\flat)}. \tag{F.3}$$

Note that, for all $\boldsymbol{H} \in \mathcal{A}_\delta$, the first summand $(\flat\flat)$ is further bounded from below by

$$\frac{\langle \|\boldsymbol{X}_\sharp\|_* \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top, \boldsymbol{H}\rangle}{r\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}\|\boldsymbol{H}\|_{\mathrm{F}}} \geq -\frac{\|\boldsymbol{X}_\sharp\|_*}{r\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}} \cdot \frac{\sqrt{r}\delta}{1-\lambda}.$$

The second term $(\flat\flat\flat)$ can be upper-bounded by the Cauchy–Schwarz inequality with

$$\left\| \boldsymbol{X}_\sharp - \frac{\|\boldsymbol{X}_\sharp\|_* \boldsymbol{U}_\sharp \boldsymbol{V}_\sharp^\top}{r} \right\|_{\mathrm{F}} \leq \frac{\sqrt{r}(\sigma_1 - \sigma_r)}{2}.$$

By plugging in the above estimates to (F.3), we obtain a sufficient condition for $\rho \geq -0.9$ given by

$$\frac{\delta}{1-\lambda} \leq \frac{\sqrt{r}\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}}{\|\boldsymbol{X}_\sharp\|_*} \cdot \left(0.9 - \frac{\sqrt{r}(\sigma_1(\boldsymbol{X}_\sharp) - \sigma_r(\boldsymbol{X}_\sharp))}{2\|\boldsymbol{X}_\sharp\|_{\mathrm{F}}}\right). \tag{F.4}$$

Here the right-hand side of (F.4) is no larger than $(2.8 - \kappa)/2$. Therefore, (14) implies that $1 + \rho \geq 0.1$. Then Lemma F.1 provides the lower bound in (40). This completes the proof.

G. Proof of Lemma 5.5

Without loss of generality, we may assume $\|\boldsymbol{X}_\sharp\|_{\mathrm{F}} = \|\boldsymbol{H}\|_{\mathrm{F}} = 1$. Since $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$ are orthogonal projection operators onto corresponding subspaces, they are self-adjoint and idempotent linear operators. Therefore, it follows that:

$$\langle \boldsymbol{\Phi}_m, \boldsymbol{H}\rangle = \langle \mathcal{P}_T(\boldsymbol{\Phi}_m), \mathcal{P}_T(\boldsymbol{H})\rangle + \langle \mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m), \mathcal{P}_{T^\perp}(\boldsymbol{H})\rangle.$$

Then by Hölder's inequality, we obtain

$$\begin{aligned}\mathfrak{C}_M(\mathcal{A}_\delta) &= \mathbb{E}\sup_{\boldsymbol{H}\in\mathcal{A}_\delta}\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m\langle \boldsymbol{X}_\sharp,\boldsymbol{\Phi}_m\rangle\langle\boldsymbol{\Phi}_m,\boldsymbol{H}\rangle\\ &\le \mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m\mathcal{P}_T(\boldsymbol{\Phi}_m)\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle\right\|_{\mathrm{F}}\cdot\sup_{\boldsymbol{H}\in\mathcal{A}_\delta}\|\mathcal{P}_T(\boldsymbol{H})\|_{\mathrm{F}}\\ &\quad+\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m\mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle\right\|\cdot\sup_{\boldsymbol{H}\in\mathcal{A}_\delta}\|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_*\\ &\le \mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m\mathcal{P}_T(\boldsymbol{\Phi}_m)\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle\right\|_{\mathrm{F}} \qquad \text{(G.1)}\\ &\quad+\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m\mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle\right\|\cdot\frac{\sqrt{r}(1-\lambda+\delta)}{\lambda}, \qquad \text{(G.2)}\end{aligned}$$

where the last step follows from the expression of $\mathcal{A}_\delta$ in (39).

The part in (G.1) is upper-bounded by

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m\mathcal{P}_T(\boldsymbol{\Phi}_m)\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle\right\|_{\mathrm{F}} &\le \sqrt{\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m\mathcal{P}_T(\boldsymbol{\Phi}_m)\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle\right\|_{\mathrm{F}}^2}\\ &= \sqrt{\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\mathcal{P}_T(\boldsymbol{\Phi}_m)\langle\boldsymbol{\Phi}_m,\boldsymbol{X}_\sharp\rangle\|_{\mathrm{F}}^2} = \sqrt{\mathbb{E}\|\mathcal{P}_T(\boldsymbol{\Phi})\langle\boldsymbol{\Phi},\boldsymbol{X}_\sharp\rangle\|_{\mathrm{F}}^2},\end{aligned}$$

where the first step follows from Jensen's inequality; the second step holds since $(\epsilon_m)_{m=1}^M$ is a Rademacher sequence; the last step holds since $\boldsymbol{\Phi}_1,\dots,\boldsymbol{\Phi}_M$ are independent copies of $\boldsymbol{\Phi}$. Indeed, since

$$\mathcal{P}_T(\boldsymbol{\Phi}) = \boldsymbol{U}_\sharp\boldsymbol{U}_\sharp^*\boldsymbol{\Phi} + (\mathbf{I}_{d_1}-\boldsymbol{U}_\sharp\boldsymbol{U}_\sharp^*)\boldsymbol{\Phi}\boldsymbol{V}_\sharp\boldsymbol{V}_\sharp^*,$$

it follows that:

$$\mathbb{E}\|\mathcal{P}_T(\boldsymbol{\Phi})\langle\boldsymbol{\Phi},\boldsymbol{X}_\sharp\rangle\|_{\mathrm{F}}^2 = \mathbb{E}\|\boldsymbol{U}_\sharp\boldsymbol{U}_\sharp^*\boldsymbol{\Phi}\|_{\mathrm{F}}^2\langle\boldsymbol{\Phi},\boldsymbol{X}_\sharp\rangle^2 + \|(\mathbf{I}_{d_1}-\boldsymbol{U}_\sharp\boldsymbol{U}_\sharp^*)\boldsymbol{\Phi}\boldsymbol{V}_\sharp\boldsymbol{V}_\sharp^*\|_{\mathrm{F}}^2\langle\boldsymbol{\Phi},\boldsymbol{X}_\sharp\rangle^2. \qquad \text{(G.3)}$$

The first summand in the right-hand side of (G.3) is computed as

$$\begin{aligned}\mathbb{E}\|\boldsymbol{U}_\sharp^\top\boldsymbol{\Phi}\|_{\mathrm{F}}^2\langle\boldsymbol{U}_\sharp^\top\boldsymbol{\Phi},\boldsymbol{U}_\sharp^\top\boldsymbol{X}_\sharp\rangle^2 &= \mathrm{tr}\Big[\mathbb{E}\langle\mathrm{vec}(\boldsymbol{U}_\sharp^\top\boldsymbol{\Phi}),\mathrm{vec}(\boldsymbol{U}_\sharp^\top\boldsymbol{X}_\sharp)\rangle^2\mathrm{vec}(\boldsymbol{U}_\sharp^\top\boldsymbol{\Phi})\mathrm{vec}(\boldsymbol{U}_\sharp^\top\boldsymbol{\Phi})^\top\Big]\\ &= \mathrm{tr}\left(2\,\mathrm{vec}(\boldsymbol{U}_\sharp^\top\boldsymbol{X}_\sharp)\mathrm{vec}(\boldsymbol{U}_\sharp^\top\boldsymbol{X}_\sharp)^\top+\mathbf{I}_{rd_2}\right) = 2+rd_2,\end{aligned}$$

where the second step follows from Lemma A.2 since $\mathrm{vec}(\boldsymbol{U}_\sharp^\top\boldsymbol{\Phi})\sim\mathcal{N}(\mathbf{0},\mathbf{I}_{rd_2})$.

Let $\boldsymbol{\Phi}'$ be an independent copy of $\boldsymbol{\Phi}$. Since $(\mathbf{I}_{d_1} - \boldsymbol{U}_\sharp \boldsymbol{U}_\sharp^*)\boldsymbol{\Phi}$ is independent of $\boldsymbol{U}_\sharp \boldsymbol{U}_\sharp^* \boldsymbol{\Phi}$, the second summand in the right-hand side of (G.3) is written as

$$\begin{aligned}
&\mathbb{E}\|(\mathbf{I}_{d_1} - \boldsymbol{U}_\sharp \boldsymbol{U}_\sharp^*)\boldsymbol{\Phi}\boldsymbol{V}_\sharp \boldsymbol{V}_\sharp^*\|_{\mathrm{F}}^2 \langle \boldsymbol{U}_\sharp \boldsymbol{U}_\sharp^* \boldsymbol{\Phi}', \boldsymbol{X}_\sharp\rangle^2 \\
&= \mathbb{E}_{\boldsymbol{\Phi}} \left\| \left(\boldsymbol{V}_\sharp \boldsymbol{V}_\sharp^* \otimes (\mathbf{I}_{d_1} - \boldsymbol{U}_\sharp \boldsymbol{U}_\sharp^*) \right) \mathrm{vec}(\boldsymbol{\Phi}) \right\|_2^2 \mathbb{E}_{\boldsymbol{\Phi}'} \langle \boldsymbol{\Phi}', \boldsymbol{X}_\sharp \rangle^2 \\
&= \mathrm{tr}\left(\boldsymbol{V}_\sharp \boldsymbol{V}_\sharp^* \otimes (\mathbf{I}_{d_1} - \boldsymbol{U}_\sharp \boldsymbol{U}_\sharp^*) \right) = r(d_1 - r).
\end{aligned}$$

Therefore, we obtain

$$\mathbb{E}\|\mathcal{P}_T(\boldsymbol{\Phi})\langle \boldsymbol{\Phi}, \boldsymbol{X}_\sharp\rangle\|_{\mathrm{F}}^2 = r(d_1 + d_2 - r) + 2.$$

By Jensen's inequality, the expectation in (G.2) is upper-bounded by

$$\mathbb{E}\left\| \sum_{m=1}^M \epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\sharp\rangle \right\| \le \left(\mathbb{E}\left\| \sum_{m=1}^M \epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\sharp\rangle \right\|^{2p} \right)^{1/2p}$$

for all $p \in \mathbb{N}$. Then we apply the noncommutative Rosenthal inequality (Theorem B.2) for

$$\boldsymbol{Y}_m = \epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\sharp\rangle, \quad m = 1, \dots, M.$$

Since $\mathcal{P}_T(\boldsymbol{X}_\sharp) = \boldsymbol{X}_\sharp$ and $\mathcal{P}_T(\boldsymbol{\Phi}_m)$ is independent from $\mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)$, it follows:

$$\boldsymbol{Y}_m = \epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle \mathcal{P}_T(\boldsymbol{\Phi}'_m), \boldsymbol{X}_\sharp\rangle, \quad m = 1, \dots, M,$$

where $\boldsymbol{\Phi}'_1, \dots, \boldsymbol{\Phi}'_M$ are independent copies of $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_M$. Furthermore, we have $\mathbb{E}\,\boldsymbol{Y}_m = \mathbf{0}$ for $m = 1, \dots, M$. By direct computation with Lemma A.2, we obtain

$$\mathbb{E}\boldsymbol{Y}_m \boldsymbol{Y}_m^\top = \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{V}_\sharp^\perp})\boldsymbol{P}_{\boldsymbol{U}_\sharp^\perp} \quad \text{and} \quad \mathbb{E}\boldsymbol{Y}_m^\top \boldsymbol{Y}_m = \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{U}_\sharp^\perp})\boldsymbol{P}_{\boldsymbol{V}_\sharp^\perp}, \quad m = 1, \dots, M.$$

Therefore, we obtain

$$\left\| \sum_{m=1}^M \mathbb{E}\,\boldsymbol{Y}_m \boldsymbol{Y}_m^\top \right\|^{1/2} \vee \left\| \sum_{m=1}^M \mathbb{E}\,\boldsymbol{Y}_m^\top \boldsymbol{Y}_m \right\|^{1/2} \le \sqrt{M(d_1 + d_2)}.$$

Next we derive an upper bound for $\left(\sum_{m=1}^M \mathbb{E}\|\boldsymbol{Y}_m\|^{2p}\right)^{1/2p}$, which coincides with $M^{1/2p}\left(\mathbb{E}\|\boldsymbol{Y}_m\|^{2p}\right)^{1/2p}$ for any $m \in \{1, \dots, M\}$. Since $\mathcal{P}_T(\boldsymbol{\Phi}_m)$ and $\mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)$ are independent, it follows that the spectral norm of $\boldsymbol{Y}_m$ satisfies:

$$\mathbb{E}\|\boldsymbol{Y}_m\|^{2p} \le \left(\mathbb{E}\|\mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\|^{2p}\right) \cdot \left(\mathbb{E}|\langle \mathcal{P}_T(\boldsymbol{\Phi}_m), \boldsymbol{X}_\sharp\rangle|^{2p}\right) \le (C\sqrt{p})^{2p}\,\mathbb{E}\|\boldsymbol{\Phi}_m\|^{2p},$$

where the last inequality follows from the fact that $\langle \mathcal{P}_T(\boldsymbol{\Phi}_m), \boldsymbol{X}_\sharp\rangle \sim \mathcal{N}(0, 1)$ satisfies:

$$\mathbb{E}|\langle \mathcal{P}_T(\boldsymbol{\Phi}_m), \boldsymbol{X}_\sharp\rangle|^{2p} \le (C\sqrt{p})^{2p}.$$

It remains to get an upper bound on $\mathbb{E}\|\boldsymbol{\Phi}_m\|^{2p}$. Note that $\|\boldsymbol{\Phi}_m\|^2 = \|\boldsymbol{\Phi}_m^\top \boldsymbol{\Phi}_m\|$ where $\boldsymbol{\Phi}_m^\top \boldsymbol{\Phi}_m$ follows the Wishart distribution. Without loss of generality, we may assume $d_1 \le d_2$ (otherwise we consider $\boldsymbol{\Phi}_m \boldsymbol{\Phi}_m^\top$

instead of $\boldsymbol{\Phi}_m^\top \boldsymbol{\Phi}_m$). Then Lemma B.3 implies

$$(\mathbb{E}\|\boldsymbol{\Phi}_m^\top \boldsymbol{\Phi}_m\|^p)^{1/p} \le d_1 + C_1\left(\sqrt{pd_1d_2} + pd_1^{1/p}\left(d_2+p\right)\right). \tag{G.4}$$

Indeed, (G.4) is obtained by Lemma B.3 and the triangle inequality in the Banach space of random variables $L_p(\Omega,\mu)$. Note that $\boldsymbol{\Phi}_m^\top \boldsymbol{\Phi}_m$ is written as

$$\boldsymbol{\Phi}_m^\top \boldsymbol{\Phi}_m = \sum_{k=1}^{d_1} \boldsymbol{g}_k \boldsymbol{g}_k^\top,$$

where $\boldsymbol{g}_1,\dots,\boldsymbol{g}_{d_1}$ are independent copies of $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_2})$. Then it follows by Lemma B.3 that:

$$\left(\mathbb{E}\left\|\frac{1}{d_1}\boldsymbol{\Phi}_m^\top \boldsymbol{\Phi}_m - \mathbf{I}_{d_2}\right\|^p\right)^{1/p} \le C_1\left(d_1^{-1/2}\sqrt{pd_2} + d_1^{1/p-1}p\left(d_2+p\right)\right),$$

which, together with the triangle inequality and the homogeneity of L_p-norm, implies (G.4). Then taking the square root on both sides of (G.4) gives

$$(\mathbb{E}\|\boldsymbol{\Phi}_m\|^{2p})^{1/2p} \le C_2\left(\sqrt{d_1} + (pd_1d_2)^{1/4} + \sqrt{p}d_1^{1/2p}\sqrt{d_2} + pd_1^{1/2p}\right).$$

By collecting the above estimates, Theorem B.2 implies

$$\left(\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\natural\rangle\right\|^{2p}\right)^{1/2p}$$
$$\le C_3\left(\sqrt{p(d_1+d_2)} + M^{1/2p-1/2}p^{3/2}\left(\sqrt{d_1} + (pd_1d_2)^{1/4} + \sqrt{p}d_1^{1/2p}\sqrt{d_2} + pd_1^{1/2p}\right)\right).$$

For the brevity, let $d = d_1 + d_2$. Let us choose $p = 1 \vee \log d$. Then $1/2p - 1/2 \le 0$. Since $M \ge d$, we obtain

$$M^{1/2p-1/2} \le d^{1/2p-1/2} \le \frac{d^{1/2\log d}}{\sqrt{d}} \le C_4 d^{-1/2}.$$

Furthermore, we have

$$\sqrt{d_1} + (pd_1d_2)^{1/4} + \sqrt{p}d_1^{1/2p}\sqrt{d_2} + p^{3/2}d_1^{1/2p}d_2^{1/2p} \le C_5\left(\sqrt{d\log d} + \log^{3/2} d\right).$$

Combining the above estimates provides

$$\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\natural\rangle\right\| \le \left(\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M}\epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{\Phi}_m)\langle \boldsymbol{\Phi}_m, \boldsymbol{X}_\natural\rangle\right\|^{2p}\right)^{1/2p}$$
$$\le C_6\sqrt{(d_1+d_2)\log(d_1+d_2)}.$$

Then the upper bound in (41) is obtained by applying the above estimates to (G.1) and (G.2).

H. Proof of Lemma 5.6

The event is determined by a 1-homogeneous equation in $\boldsymbol{H}$ and $\boldsymbol{X}_\sharp$. Therefore, without loss of generality, we may assume that $\|\boldsymbol{H}\|_{\mathrm{F}} = \|\boldsymbol{X}_\sharp\|_{\mathrm{F}} = 1$. Then $\boldsymbol{X}_\sharp$ is written as $\boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^*$ with $\|\boldsymbol{u}_\sharp\|_2 = \|\boldsymbol{v}_\sharp\|_2 = 1$.

First, we decompose $\mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{H}\boldsymbol{b})$ as

$$\begin{aligned}\mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{H}\boldsymbol{b}) &= \mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp} \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp} \boldsymbol{b}) + \mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp} \boldsymbol{b}) \\ &\quad + \mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp} \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}) + \mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}).\end{aligned}$$

By plugging in $\boldsymbol{P}_{\boldsymbol{u}_\sharp} = \boldsymbol{u}_\sharp \boldsymbol{u}_\sharp^*$ and $\boldsymbol{P}_{\boldsymbol{v}_\sharp} = \boldsymbol{v}_\sharp \boldsymbol{v}_\sharp^*$ to the above identity, we rewrite $\mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{H}\boldsymbol{b})$ as

$$\begin{aligned}\mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{H}\boldsymbol{b}) &= |\boldsymbol{v}_\sharp^* \boldsymbol{b}|^2 |\boldsymbol{u}_\sharp^* \boldsymbol{a}|^2 \,\mathrm{Re}(\boldsymbol{u}_\sharp^* \boldsymbol{H} \boldsymbol{v}_\sharp) + |\boldsymbol{v}_\sharp^* \boldsymbol{b}|^2 |\boldsymbol{u}_\sharp^* \boldsymbol{a}| \,\mathrm{Re}\left(\frac{\boldsymbol{u}_\sharp^* \boldsymbol{a}}{|\boldsymbol{u}_\sharp^* \boldsymbol{a}|} \cdot \boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{H} \boldsymbol{v}_\sharp \right) \\ &\quad + |\boldsymbol{u}_\sharp^* \boldsymbol{a}|^2 |\boldsymbol{v}_\sharp^* \boldsymbol{b}| \,\mathrm{Re}\left(\frac{\overline{\boldsymbol{v}_\sharp^* \boldsymbol{b}}}{|\boldsymbol{v}_\sharp^* \boldsymbol{b}|} \cdot \boldsymbol{u}_\sharp^* \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b} \right) + |\boldsymbol{u}_\sharp^* \boldsymbol{a}| |\boldsymbol{v}_\sharp^* \boldsymbol{b}| \,\mathrm{Re}\left(\frac{\boldsymbol{u}_\sharp^* \boldsymbol{a}}{|\boldsymbol{u}_\sharp^* \boldsymbol{a}|} \cdot \frac{\overline{\boldsymbol{v}_\sharp^* \boldsymbol{b}}}{|\boldsymbol{v}_\sharp^* \boldsymbol{b}|} \cdot \boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b} \right).\end{aligned}$$

The following facts follow from the assumption that $\boldsymbol{a} \sim \mathcal{CN}(\boldsymbol{0}, \mathbf{I}_{d_1})$ and $\boldsymbol{b} \sim \mathcal{CN}(\boldsymbol{0}, \mathbf{I}_{d_2})$ are mutually independent:

1. $|\boldsymbol{u}_\sharp^* \boldsymbol{a}|$, $|\boldsymbol{v}_\sharp^* \boldsymbol{b}|$, $\overline{\boldsymbol{u}_\sharp^* \boldsymbol{a}}/|\boldsymbol{u}_\sharp^* \boldsymbol{a}|$, $\overline{\boldsymbol{v}_\sharp^* \boldsymbol{b}}/|\boldsymbol{v}_\sharp^* \boldsymbol{b}|$, $\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{a}$, and $\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}$ are independent random variables.
2. $|\boldsymbol{u}_\sharp^* \boldsymbol{a}|$ and $|\boldsymbol{v}_\sharp^* \boldsymbol{b}|$ follow the Rayleigh distribution with scale parameter 1.
3. $\overline{\boldsymbol{u}_\sharp^* \boldsymbol{a}}/|\boldsymbol{u}_\sharp^* \boldsymbol{a}|$ and $\overline{\boldsymbol{v}_\sharp^* \boldsymbol{b}}/|\boldsymbol{v}_\sharp^* \boldsymbol{b}|$ follow the uniform distribution on the set of complex number of the unit modulus.

Furthermore, due to the rotation invariance of the Gaussian distribution, $(\overline{\boldsymbol{u}_\sharp^* \boldsymbol{a}}/|\boldsymbol{u}_\sharp^* \boldsymbol{a}|) \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{a}$ has the same distribution with $\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{a}$. Similarly, $(\overline{\boldsymbol{v}_\sharp^* \boldsymbol{b}}/|\boldsymbol{v}_\sharp^* \boldsymbol{b}|) \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}$ and $\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}$ have the same distribution.

Combining the above facts, we obtain that $\mathrm{Re}(\boldsymbol{b}^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}\boldsymbol{a}^* \boldsymbol{H}\boldsymbol{b})$ has the same distribution with

$$x := r_1^2 r_2^2 \,\mathrm{Re}(\boldsymbol{u}_\sharp^* \boldsymbol{H} \boldsymbol{v}_\sharp) + r_1 r_2^2 \,\mathrm{Re}(\boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{H} \boldsymbol{v}_\sharp) + r_1^2 r_2 \,\mathrm{Re}(\boldsymbol{u}_\sharp^* \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}) + r_1 r_2 \,\mathrm{Re}(\boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}),$$

where $r_1, r_2, \boldsymbol{a}, \boldsymbol{b}$ are independent and $r_1, r_2 \sim \mathrm{Rayleigh}(1)$.

Now it suffices to compute the probability of the event $\mathcal{E}$ defined by

$$\mathcal{E} := \{x \geq \tau'\}.$$

For positive constants α, β, we define another event $\mathcal{E}_0$ by

$$\mathcal{E}_0 := \{\alpha \leq r_1 \leq \beta,\ \alpha \leq r_2 \leq \beta\}.$$

For example, we may set $\alpha = 0.9$ and $\beta = 1.1$. Then $\mathbb{P}(\mathcal{E}_0) \geq 0.12$.

Let z_1, z_2, z_3 be random variables defined by

$$z_1 := \mathrm{Re}(\boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{H} \boldsymbol{v}_\sharp), \quad z_2 := \mathrm{Re}(\boldsymbol{u}_\sharp^* \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}) \quad \text{and} \quad z_3 := \mathrm{Re}(\boldsymbol{a}^* \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp} \boldsymbol{H} \boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} \boldsymbol{b}).$$

Since $\boldsymbol{a} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{d_1})$, if $\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{v}_\sharp \neq \mathbf{0}$, then it follows that $\boldsymbol{a}^*\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{v}_\sharp \sim \mathcal{CN}(0, \|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp}\|_{\mathrm{F}}^2)$ and its real part z_1 follows $\mathcal{N}(0, \|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp}\|_{\mathrm{F}}^2/2)$. Otherwise, $\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{v}_\sharp = \mathbf{0}$ implies $z_1 = 0$. Similarly, $z_2 \sim \mathcal{N}(0, \|\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\boldsymbol{H}^*\boldsymbol{P}_{\boldsymbol{u}_\sharp}\|_{\mathrm{F}}^2/2)$ if $\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\boldsymbol{H}^*\boldsymbol{u}_\sharp \neq \mathbf{0}$; $z_2 = 0$ otherwise. By the independence between $\boldsymbol{a}$ and $\boldsymbol{b}$, it follows that $z_1 + z_2 \sim \mathcal{N}(0, \|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2/2)$ if $\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp}\|_{\mathrm{F}}^2 + \|\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2 > 0$; $z_1 + z_2 = 0$ otherwise. In both cases, $z_1 + z_2$ has a symmetric distribution, that is $z_1 + z_2$ is equivalent to $-(z_1 + z_2)$ in distribution.

Furthermore, we can rewrite z_3 as a Gaussian bilinear form, i.e.

$$z_3 = \widetilde{\boldsymbol{a}}^\top \boldsymbol{Q}\widetilde{\boldsymbol{b}}$$

for

$$\boldsymbol{Q} = \frac{1}{2}\begin{bmatrix} \mathrm{Re}(\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}) & -\mathrm{Im}(\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}) \\ \mathrm{Im}(\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}) & \mathrm{Re}(\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}) \end{bmatrix},$$

where

$$\widetilde{\boldsymbol{a}} = \sqrt{2}\begin{bmatrix} \mathrm{Re}(\boldsymbol{a}) \\ \mathrm{Im}(\boldsymbol{a}) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{2d_1}), \qquad \widetilde{\boldsymbol{b}} = \sqrt{2}\begin{bmatrix} \mathrm{Re}(\boldsymbol{b}) \\ \mathrm{Im}(\boldsymbol{b}) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{2d_2}).$$

It follows that z_3 has a symmetric distribution. Furthermore, since z_3 is a Gaussian bilinear form, it has a mixed subexponential-subgaussian tail given by

$$\mathbb{P}(|z_3| \geq t) \leq C\exp\left[-\frac{1}{C}\left(\frac{t^2}{\|\boldsymbol{Q}\|_{\mathrm{F}}^2} \wedge \frac{t}{\|\boldsymbol{Q}\|}\right)\right], \quad \forall t > 0 \tag{H.1}$$

for a numerical constant C. Latała [38] showed that this tail bound is tight with an analogous lower bound given by

$$\mathbb{P}(|z_3| \geq t) \geq \frac{1}{C}\exp\left[-C\left(\frac{t^2}{\|\boldsymbol{Q}\|_{\mathrm{F}}^2} \wedge \frac{t}{\|\boldsymbol{Q}\|}\right)\right], \quad \forall t > 0.$$

By direct calculation, we obtain

$$\|\boldsymbol{Q}\|_{\mathrm{F}} = \frac{1}{\sqrt{2}}\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}$$

and

$$\|\boldsymbol{Q}\| = \frac{1}{2}\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|.$$

Now we are ready to derive a lower bound on the probability of the event $\mathcal{E}$ using the aforementioned properties z_1, z_2, z_3. It follows from the definition of the conditional probability that

$$\frac{\mathbb{P}(\mathcal{E})}{\mathbb{P}(\mathcal{E}_0)} \geq \frac{\mathbb{P}(\mathcal{E} \cap \mathcal{E}_0)}{\mathbb{P}(\mathcal{E}_0)} = \mathbb{P}(\mathcal{E}|\mathcal{E}_0) = \mathbb{P}\Big(r_1 r_2 \,\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) + r_2 z_1 + r_1 z_2 + z_3 \geq \frac{\tau'}{r_1 r_2} \,\Big|\, \mathcal{E}_0\Big). \tag{H.2}$$

As we choose $\alpha < \beta$ as numerical constants, $\mathbb{P}(\mathcal{E}_0)$ is another numerical constant. It remains to show that the lower bound in (H.2) is larger than a numerical constant. We consider the two complementary scenarios below.

Case 1: First, we consider the case when $\boldsymbol{H}$ satisfies

$$\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\| \le \zeta \tag{H.3}$$

for some constant $0 < \zeta < 1$, which we will specify later.

Let $\tau'' > 0$. Then by the inclusion–exclusion principle, the right-hand side of (H.2) is lower-bounded by

$$\begin{aligned}&\mathbb{P}\Big(r_1 r_2 \operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) + r_2 z_1 + r_1 z_2 + z_3 \ge \frac{\tau'}{r_1 r_2} \mid \mathcal{E}_0\Big)\\ &\ge \mathbb{P}\Big(r_1 r_2 \operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) + r_2 z_1 + r_1 z_2 \ge \frac{\tau' + \tau''}{r_1 r_2} \mid \mathcal{E}_0\Big) - \mathbb{P}\Big(z_3 < -\frac{\tau''}{r_1 r_2} \mid \mathcal{E}_0\Big).\end{aligned} \tag{H.4}$$

In the sequel, we will use the fact that for then the tail probability is

$$g \sim \mathcal{N}(0, \varsigma^2) \implies \mathbb{P}(g > t) \text{ is increasing in } \varsigma \text{ and decreasing in } t \ge 0\,. \tag{H.5}$$

The following cases on the sign of $\operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)$ have to be distinguished. First, suppose that $\operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) \le 0$. Conditioned on r_1 and r_2, the random variable $r_2 z_1 + r_1 z_2$ becomes a Gaussian and invoking (H.5) yields

$$\begin{aligned}&\mathbb{P}\left(r_2 z_1 + r_1 z_2 \ge \frac{\tau' + \tau''}{r_1 r_2} - r_1 r_2 \operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) \mid \mathcal{E}_0, r_1, r_2\right)\\ &\ge \mathbb{P}\left(\alpha(z_1 + z_2) \ge \frac{\tau' + \tau''}{\alpha^2} - \beta^2 \operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) \mid \mathcal{E}_0, r_1, r_2\right).\end{aligned}$$

Since the right-hand side of the above inequality is independent of r_1 and r_2, we can conclude that

$$\mathbb{P}\left(r_2 z_1 + r_1 z_2 \ge \frac{\tau' + \tau''}{r_1 r_2} - r_1 r_2 \operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) \mid \mathcal{E}_0\right) \ge \mathbb{P}\left(\alpha(z_1 + z_2) \ge \frac{\tau' + \tau''}{\alpha^2} - \beta^2 \operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)\right).$$

Furthermore, z_3 is symmetric and we obtain an upper estimate of the tail probability of z_3 in (H.4) given by

$$\mathbb{P}\Big(z_3 < -\frac{\tau''}{r_1 r_2} \mid \mathcal{E}_0\Big) \le \mathbb{P}\Big(z_3 < -\frac{\tau''}{\beta^2}\Big) = \frac{1}{2}\,\mathbb{P}\Big(|z_3| \ge \frac{\tau''}{\beta^2}\Big).$$

By combining the above bounds, the lower estimate in (H.4) is further bounded from below by

$$\begin{aligned}&\mathbb{P}\Big(r_1 r_2 \operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) + r_2 z_1 + r_1 z_2 \ge \frac{\tau' + \tau''}{r_1 r_2} \mid \mathcal{E}_0\Big) - \mathbb{P}\Big(z_3 < -\frac{\tau''}{r_1 r_2} \mid \mathcal{E}_0\Big)\\ &\ge \mathbb{P}\Big(z_1 + z_2 \ge \frac{\tau' + \tau''}{\alpha^3} - \frac{\beta^2 \operatorname{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)}{\alpha}\Big) - \frac{1}{2}\,\mathbb{P}\Big(|z_3| \ge \frac{\tau''}{\beta^2}\Big).\end{aligned}$$

Because $z_1 + z_2 \sim \mathcal{N}(0, \|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2/2)$, in order to lower-bound the tail probability of $z_1 + z_2$, we need to compute a lower estimate of its variance. It follows from (39) that every $\boldsymbol{H} \in \mathcal{A}_\delta$ satisfies

$$\|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_* \le \frac{1 - \lambda + \delta}{\lambda}. \tag{H.6}$$

By (H.3) and (H.6) together with Hölder's inequality, we also obtain

$$\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2 \le \|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\| \cdot \|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_* \le \frac{(1-\lambda+\delta)\zeta}{\lambda}. \tag{H.7}$$

Furthermore, by Lemma 5.1, every $\boldsymbol{H} \in \mathcal{R}_\delta$ satisfies

$$\frac{1-\delta^2}{\delta^2} \cdot |\mathrm{Im}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)|^2 \le \|\boldsymbol{H} - \boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp}\|_{\mathrm{F}}^2.$$

Then it follows that:

$$\begin{aligned} 1 = \|\boldsymbol{H}\|_{\mathrm{F}}^2 &= \|\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2 + |\mathrm{Im}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)|^2 + |\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)|^2 \\ &\le \frac{\|\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2}{1-\delta^2} + |\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)|^2. \end{aligned} \tag{H.8}$$

It also follows from (39) that every $\boldsymbol{H} \in \mathcal{A}_\delta$ satisfies:

$$\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) \ge \frac{-\delta}{1-\lambda}. \tag{H.9}$$

The assumption $\lambda + \delta < 1$ implies that the right-hand side of (H.9) is strictly larger than -1.

By (H.9) and $\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) \le 0$, we also have

$$|\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)| \le \frac{\delta}{1-\lambda}. \tag{H.10}$$

Therefore, by applying (H.10) to (H.8), after a rearrangement, we obtain

$$\|\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2 \ge (1-\delta^2)\left(1 - \frac{\delta^2}{(1-\lambda)^2}\right).$$

Then (H.7) implies

$$\|\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp} + \boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp}\|_{\mathrm{F}}^2 \ge (1-\delta^2)\left(1 - \frac{\delta^2}{(1-\lambda)^2}\right) - \frac{(1-\lambda+\delta)\zeta}{\lambda}. \tag{H.11}$$

Now, from (H.10) and (H.11), we obtain

$$\mathbb{P}\left(z_1 + z_2 \ge \frac{\tau'+\tau''}{\alpha^3} - \frac{\beta^2\,\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)}{\alpha}\right) \ge \mathbb{P}\left(z_1 + z_2 \ge \frac{\tau'+\tau''}{\alpha^3} + \frac{\beta^2\delta}{\alpha(1-\lambda)}\right) \ge \mathbb{P}\left(g \ge \frac{t}{\sigma_\zeta}\right) \tag{H.12}$$

for $g \sim \mathcal{N}(0,1)$, where

$$\sigma_\zeta = \sqrt{(1-\delta^2)\left(1 - \frac{\delta^2}{(1-\lambda)^2}\right) - \frac{(1-\lambda+\delta)\zeta}{\lambda}}$$

and

$$t = \frac{\tau'+\tau''}{\alpha^3} + \frac{\beta^2\delta}{\alpha(1-\lambda)}. \tag{H.13}$$

Moreover, the tail bound of z_3 in (H.1) implies

$$\mathbb{P}\Big(|z_3| \geq \frac{\tau''}{\beta^2}\Big) \leq C\exp\Big(-\frac{2}{C}\Big[\frac{\tau''^2/\beta^4}{\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2} \wedge \frac{\tau''/\beta^2}{\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|}\Big]\Big)$$
$$\leq C\exp\Big(-\frac{2}{C}\Big[\frac{\lambda\tau''^2}{\beta^4(1-\lambda+\delta)\zeta} \wedge \frac{\tau''}{\beta^2\zeta}\Big]\Big). \tag{H.14}$$

Note that the tail bound in (H.12) is monotone decreasing in t/σ_ζ. Furthermore, for those ζ that make σ_ζ positive, t/σ_ζ is a monotone increasing in ζ. (The condition $\delta \leq 0.2$ implies the existence of such ζ.) Hence, the tail bound in (H.12) is monotone decreasing in ζ. On the contrary, the upper bound in (H.14) monotonically converges to 0 as $\zeta > 0$ decreases toward 0. Therefore, there exists small enough ζ such that the upper bound in (H.14) becomes less than half of (H.12). Then the lower bound (H.2) is further bounded from below by the half of (H.12). Note that ζ is determined independent from all dimension parameters and hence both ζ and the resulting lower bound for the probability in (H.2) are numerical constants.

Next we consider the complimentary subcase where $\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) > 0$. Similarly to the previous subcase, since z_3 has a symmetric distribution, it follows that:

$$\mathbb{P}\left(r_1r_2\,\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) + r_2z_1 + r_1z_2 \geq \frac{\tau'+\tau''}{r_1r_2} \mid \mathcal{E}_0\right) - \mathbb{P}\left(z_3 < -\frac{\tau''}{r_1r_2} \mid \mathcal{E}_0\right)$$
$$\geq \mathbb{P}\left(\alpha^2\,\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) + r_2z_1 + r_1z_2 \geq \frac{\tau'+\tau''}{\alpha^2} \mid \mathcal{E}_0\right) - \frac{1}{2}\mathbb{P}\left(|z_3| \geq \frac{\tau''}{\beta^2}\right).$$

If $\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) \leq \alpha^{-4}(\tau'+\tau'')$, since z_1+z_2 is a zero-mean Gaussian variable, then it follows that:

$$\mathbb{P}\left(\alpha^2\,\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) + r_2z_1 + r_1z_2 \geq \frac{\tau'+\tau''}{\alpha^2} \mid \mathcal{E}_0\right) \geq \mathbb{P}\left(z_1+z_2 \geq \frac{\tau'+\tau''}{\alpha^3} - \alpha\,\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)\right).$$

Thus by choosing $\tau'+\tau''$ small enough one can satisfy (H.10). Thus, we obtain the desired conclusion as in the previous subcase by repeating the same arguments.

If $\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) > \alpha^{-4}(\tau'+\tau'')$ on the other hand, then

$$\mathbb{P}\Big(\alpha^2\,\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp) + r_2z_1 + r_1z_2 \geq \frac{\tau'+\tau''}{\alpha^2} \mid \mathcal{E}_0\Big) \geq \mathbb{P}\Big(z_1+z_2 \geq \underbrace{\frac{\tau'+\tau''}{\alpha^2\beta} - \frac{\alpha^2\,\mathrm{Re}(\boldsymbol{u}_\sharp^*\boldsymbol{H}\boldsymbol{v}_\sharp)}{\beta}}_{<0}\Big) > \frac{1}{2},$$

which is larger than the other lower bounds on the tail probability.

Case 2: Next we consider the complementary case where

$$\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{H}\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}} > \zeta, \tag{H.15}$$

where ζ is the constant determined in the previous case. In this case, the lower estimate in (H.2) is further bounded from below by

$$\begin{aligned}
&\mathbb{P}\Big(r_1 r_2 \operatorname{Re}(\boldsymbol{u}_\sharp^* \boldsymbol{H} \boldsymbol{v}_\sharp) + r_2 z_1 + r_1 z_2 + z_3 \ge \frac{\tau'}{r_1 r_2} \,|\, \mathcal{E}_0\Big) \\
&\quad \ge \mathbb{P}\Big(- \beta^2 [\operatorname{Re}(\boldsymbol{u}_\sharp^* \boldsymbol{H} \boldsymbol{v}_\sharp)]_- + r_2 z_1 + r_1 z_2 + z_3 \ge \frac{\tau'}{r_1 r_2} \,|\, \mathcal{E}_0\Big) \\
&\quad \ge \mathbb{P}\Big(r_2 z_1 + r_1 z_2 + z_3 \ge \frac{\tau'}{r_1 r_2} + \beta^2(1-\zeta) \,|\, \mathcal{E}_0\Big) \\
&\quad \ge \mathbb{P}\Big(z_3 \ge \frac{\tau' + \tau''}{r_1 r_2} + \beta^2(1-\zeta) \,|\, \mathcal{E}_0\Big) + \mathbb{P}\Big(r_2 z_1 + r_1 z_2 \ge -\frac{\tau''}{r_1 r_2} \,|\, \mathcal{E}_0\Big) - 1 \\
&\quad \ge \mathbb{P}\Big(z_3 \ge \frac{\tau' + \tau''}{\alpha^2} + \beta^2(1-\zeta)\Big) - \mathbb{P}\Big(z_1 + z_2 \ge \frac{\tau''}{\beta^3}\Big) \\
&\quad \ge \frac{1}{2}\mathbb{P}\Big(|z_3| \ge \frac{\tau' + \tau''}{\alpha^2} + \beta^2(1-\zeta)\Big) - \mathbb{P}\Big(z_1 + z_2 \ge \frac{\tau''}{\beta^3}\Big),
\end{aligned}$$

where the second and third steps follow from (H.10) and the inclusion–exclusion principle, respectively. Then, by (H.15), the tail bound on z_3 is lower-bounded by

$$\mathbb{P}(|z_3| \ge t) \ge \frac{1}{C} \exp\Big(- 2C\Big[\frac{t^2}{\zeta^2} \wedge \frac{t}{\zeta}\Big]\Big), \tag{H.16}$$

where t is given in (H.13). Since $\|\boldsymbol{H}\|_{\mathrm{F}} = 1$, the variance of $z_1 + z_2$ is no larger than $1/2$. Thus, the tail bound of $z_1 + z_2$ is upper-bounded by

$$\mathbb{P}\Big(z_1 + z_2 \ge \frac{\tau''}{\beta^3}\Big) \le \exp\Big(- \frac{2\tau''^2}{\beta^6}\Big). \tag{H.17}$$

Note that τ'' still remains a free parameter. For every $t > \zeta$ the lower bound in (H.16) is an exponential tail while the upper bound in (H.17) is a subgaussian tail. Therefore, as τ'' increases while the other parameters are fixed, by (H.13), t also increases as an affine function of τ'' and the lower bound in (H.16) decays slower than the upper bound in (H.17). We may choose τ'' so that the lower bound in (H.16) is larger then four times the upper bound in (H.17). Then the lower bound (H.2) is further bounded below by the resulting value of (H.17). Again, this lower bound is a numerical constant independent of scaling of all dimension parameters.

I. Proof of Lemma 5.7

Without loss of generality, we may assume that $\|\boldsymbol{H}\|_{\mathrm{F}} = \|\boldsymbol{X}_\sharp\|_{\mathrm{F}} = 1$. Then $\boldsymbol{X}_\sharp$ is written as $\boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^*$ where $\boldsymbol{u}_\sharp \in \mathbb{C}^{d_1}$ and $\boldsymbol{v}_\sharp \in \mathbb{C}^{d_2}$ satisfy $\|\boldsymbol{u}_\sharp\|_2 = \|\boldsymbol{v}_\sharp\|_2 = 1$. With this expression of $\boldsymbol{X}_\sharp$, the Rademacher

complexity $\mathfrak{C}_M(\mathcal{A}_\delta)$ is written as

$$\begin{aligned}
\mathfrak{C}_M(\mathcal{A}_\delta) &= \mathbb{E} \sup_{\boldsymbol{H}\in\mathcal{A}_\delta} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathrm{Re}(\boldsymbol{b}_m^* \boldsymbol{v}_\sharp \boldsymbol{u}_\sharp^* \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{H} \boldsymbol{b}_m) \\
&= \mathbb{E} \sup_{\boldsymbol{H}\in\mathcal{A}_\delta} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathrm{Re}\, \langle \boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*, \boldsymbol{H} \rangle \\
&\le \mathbb{E} \sup_{\boldsymbol{H}\in\mathcal{A}_\delta} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathrm{Re}\, \langle \mathcal{P}_T(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*), \mathcal{P}_T(\boldsymbol{H}) \rangle \\
&\quad + \mathbb{E} \sup_{\boldsymbol{H}\in\mathcal{A}_\delta} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathrm{Re}\, \langle \mathcal{P}_{T^\perp}(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*), \mathcal{P}_{T^\perp}(\boldsymbol{H}) \rangle \\
&\le \mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathcal{P}_T(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*) \right\|_{\mathrm{F}} \cdot \sup_{\boldsymbol{H}\in\mathcal{A}_\delta} \|\mathcal{P}_T(\boldsymbol{H})\|_{\mathrm{F}} \\
&\quad + \mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*) \right\| \cdot \sup_{\boldsymbol{H}\in\mathcal{A}_\delta} \|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_*,
\end{aligned}$$

where the first inequality is obtained by taking the supremum of each summand after applying $\boldsymbol{H} = \mathcal{P}_T(\boldsymbol{H}) + \mathcal{P}_{T^\perp}(\boldsymbol{H})$ and the second inequality holds by Hölder's inequality.

Since $\mathcal{P}_T$ is an orthogonal projection onto a subspace, we have $\|\mathcal{P}_T(\boldsymbol{H})\|_{\mathrm{F}} \le \|\boldsymbol{H}\|_{\mathrm{F}} = 1$. Furthermore, for all $\boldsymbol{H} \in \mathcal{A}_\delta$, $\|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_*$ is upper-bounded by (H.6). Therefore, we obtain

$$\begin{aligned}
\mathfrak{C}_M(\mathcal{A}_\delta) &\le \mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathcal{P}_T(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*) \right\|_{\mathrm{F}} \\
&\quad + \mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*) \right\| \cdot \left(\frac{1 - \lambda + \delta}{\lambda} \right).
\end{aligned} \tag{I.1}$$

It remains to compute upper estimates of the expectation terms in (I.1). Since $(\epsilon_m)_{m=1}^M$ is a Rademacher sequence, we have

$$\begin{aligned}
\mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathcal{P}_T(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*) \right\|_{\mathrm{F}} &\le \sqrt{\mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \mathcal{P}_T(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*) \right\|_{\mathrm{F}}^2} \\
= \sqrt{\mathbb{E}\, \frac{1}{M} \sum_{m=1}^{M} \|\mathcal{P}_T(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*)\|_{\mathrm{F}}^2} &= \sqrt{\mathbb{E}\, \|\mathcal{P}_T(\boldsymbol{a}\boldsymbol{a}^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}\boldsymbol{b}^*)\|_{\mathrm{F}}^2},
\end{aligned}$$

where the first step follows from Jensen's inequality and the last step follows since $\boldsymbol{a}_1, \ldots \boldsymbol{a}_M$ (resp. $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$) are independent copies of $\boldsymbol{a}$ (resp. $\boldsymbol{b}$).

Note that $\mathcal{P}_T(\boldsymbol{a}\boldsymbol{a}^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}\boldsymbol{b}^*)$ is written as

$$\begin{aligned}\mathcal{P}_T(\boldsymbol{a}\boldsymbol{a}^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}\boldsymbol{b}^*) &= \boldsymbol{a}^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}\cdot\mathcal{P}_T(\boldsymbol{a}\boldsymbol{b}^*)\\ &= \boldsymbol{a}^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}\cdot(\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp}+\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp}+\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}),\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp}$, $\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp}$ and $\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}$ are mutually orthogonal matrices in the Hilbert space S_2. Thus, the Pythagorean identity implies

$$\begin{aligned}\|\mathcal{P}_T(\boldsymbol{a}\boldsymbol{a}^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}\boldsymbol{b}^*)\|_{\mathrm{F}}^2 &= |\boldsymbol{a}^*\boldsymbol{u}_\sharp|^2|\boldsymbol{b}^*\boldsymbol{v}_\sharp|^2(\|\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp}\|_{\mathrm{F}}^2+\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp}\|_{\mathrm{F}}^2+\|\boldsymbol{P}_{\boldsymbol{u}_\sharp}\boldsymbol{a}\boldsymbol{b}^*\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\|_{\mathrm{F}}^2)\\ &= |\boldsymbol{a}^*\boldsymbol{u}_\sharp|^4|\boldsymbol{b}^*\boldsymbol{v}_\sharp|^4+|\boldsymbol{a}^*\boldsymbol{u}_\sharp|^2|\boldsymbol{b}^*\boldsymbol{v}_\sharp|^4\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}\|_2^2+|\boldsymbol{a}^*\boldsymbol{u}_\sharp|^4|\boldsymbol{b}^*\boldsymbol{v}_\sharp|^2\|\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\boldsymbol{b}\|_2^2.\end{aligned}$$

Since $\boldsymbol{a}\sim\mathcal{CN}(\boldsymbol{0},\mathbf{I}_{d_1})$ and $\boldsymbol{b}\sim\mathcal{CN}(\boldsymbol{0},\mathbf{I}_{d_2})$ are independent, $\boldsymbol{a}^*\boldsymbol{u}_\sharp$, $\boldsymbol{b}^*\boldsymbol{v}_\sharp$, $\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}$ and $\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\boldsymbol{b}$ are all mutually independent. Therefore, exploiting this independence, one can show that the expectation is upper-bounded by

$$\mathbb{E}\,\|\mathcal{P}_T(\boldsymbol{a}\boldsymbol{a}^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}\boldsymbol{b}^*)\|_{\mathrm{F}}^2\le 2\|\boldsymbol{u}_\sharp\|_2^2\|\boldsymbol{v}_\sharp\|_2^2(2+d_1+d_2).$$

By Jensen's inequality, the second expectation in (I.1) is upper-bounded by

$$\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^M\epsilon_m\mathcal{P}_{T^\perp}(\boldsymbol{a}_m\boldsymbol{a}_m^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}_m\boldsymbol{b}_m^*)\right\|\le\left(\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^M\epsilon_m\mathcal{P}_{T^\perp}(\boldsymbol{a}_m\boldsymbol{a}_m^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}_m\boldsymbol{b}_m^*)\right\|^p\right)^{1/p}\tag{I.2}$$

for all $p\in 2\mathbb{N}$. To upper bound the right-hand side of (I.2), we apply Theorem B.2 for

$$\boldsymbol{Y}_m=\epsilon_m\mathcal{P}_{T^\perp}(\boldsymbol{a}_m\boldsymbol{a}_m^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}_m\boldsymbol{b}_m^*)=\epsilon_m\boldsymbol{a}_m^*\boldsymbol{u}_\sharp\boldsymbol{v}_\sharp^*\boldsymbol{b}_m\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}_m\boldsymbol{b}_m^*\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp},\quad m=1,\dots,M,$$

with some $p\in\mathbb{N}$ that satisfies $p\ge 2$. Note that $\mathbb{E}\,\boldsymbol{Y}_m=\boldsymbol{0}$ for all $m=1,\dots,M$. By direct computation, we obtain

$$\mathbb{E}\,\boldsymbol{Y}_m\boldsymbol{Y}_m^*=\|\boldsymbol{u}_\sharp\|_2^2\|\boldsymbol{v}_\sharp\|_2^2\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp})\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\quad\text{and}\quad\mathbb{E}\,\boldsymbol{Y}_m^*\boldsymbol{Y}_m=\|\boldsymbol{u}_\sharp\|_2^2\|\boldsymbol{v}_\sharp\|_2^2\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp})\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp},\quad m=1,\dots,M.$$

Therefore,

$$\left\|\sum_{m=1}^M\mathbb{E}\,\boldsymbol{Y}_m\boldsymbol{Y}_m^*\right\|^{1/2}\vee\left\|\sum_{m=1}^M\mathbb{E}\,\boldsymbol{Y}_m^*\boldsymbol{Y}_m\right\|^{1/2}\le\|\boldsymbol{u}_\sharp\|_2\|\boldsymbol{v}_\sharp\|_2\sqrt{M(d_1+d_2)}.$$

Since the spectral norm of $\boldsymbol{Y}_m$ is upper-bounded by

$$\|\boldsymbol{Y}_m\|=|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp||\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}_m\|_2\|\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\boldsymbol{b}_m\|_2\le 2|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp||\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|(\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}_m\|_2^2+\|\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\boldsymbol{b}_m\|_2^2),$$

it follows that:

$$\begin{aligned}(\mathbb{E}\,\|\boldsymbol{Y}_m\|^p)^{1/p}&\le 2(\mathbb{E}\,|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^p)^{1/p}\cdot(\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^2)^{1/p}\cdot[\mathbb{E}(\|\boldsymbol{P}_{\boldsymbol{u}_\sharp^\perp}\boldsymbol{a}_m\|_2^2+\|\boldsymbol{P}_{\boldsymbol{v}_\sharp^\perp}\boldsymbol{b}_m\|_2^2)^p]^{1/p}\\ &\le 2(\mathbb{E}\,|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^p)^{1/p}\cdot(\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^2)^{1/p}\cdot[\mathbb{E}(\|\boldsymbol{a}_m\|_2^2+\|\boldsymbol{b}_m\|_2^2)^p]^{1/p}.\end{aligned}$$

Since $\boldsymbol{a}_m^*\boldsymbol{u}_\sharp\sim\mathcal{CN}(0,1)$ and $\boldsymbol{b}_m^*\boldsymbol{v}_\sharp\sim\mathcal{CN}(0,1)$, we have

$$(\mathbb{E}\,|\boldsymbol{a}_m^*\boldsymbol{u}_\sharp|^p)^{1/p}=(\mathbb{E}\,|\boldsymbol{b}_m^*\boldsymbol{v}_\sharp|^p)^{1/p}\le C_1\sqrt{p}.$$

for a numerical constant C_1. Since $2(\|\boldsymbol{a}_m\|_2^2 + \|\boldsymbol{b}_m\|_2^2)$ is a chi-square random variable of the degree-of-freedom $2(d_1 + d_2)$, it follows that for $p \geq 2$ we have:

$$(\mathbb{E}(\|\boldsymbol{a}_m\|_2^2 + \|\boldsymbol{b}_m\|_2^2)^p)^{1/p} \leq 2d_1 + 2d_2 + C_2 p$$

for a numerical constant C_2. By collecting these estimates, we obtain

$$p\left(\sum_{m=1}^{M} \mathbb{E}\,\|\boldsymbol{Y}_m\|^p\right)^{1/p} \leq C_3 M^{1/p} p^2 (d_1 + d_2 + p).$$

Applying the above estimates to Theorem B.2 together with (I.2) provides

$$\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M} \epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*)\right\| \leq C_4\left(\sqrt{p(d_1 + d_2)} + M^{1/p - 1/2} p^2 (d_1 + d_2 + p)\right). \tag{I.3}$$

As we set $p = \log M$, (I3) implies

$$\mathbb{E}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^{M} \epsilon_m \mathcal{P}_{T^\perp}(\boldsymbol{a}_m \boldsymbol{a}_m^* \boldsymbol{u}_\sharp \boldsymbol{v}_\sharp^* \boldsymbol{b}_m \boldsymbol{b}_m^*)\right\| \leq C_5\sqrt{(d_1 + d_2)\log M} \cdot \left(1 + \frac{\sqrt{d_1 + d_2}\,\log^{3/2} M}{\sqrt{M}}\right).$$

Then (10) implies that the right-hand side is further upper bounded by $C_6\sqrt{d_1 + d_2}\log M$. Finally, (42) is obtained by plugging in these upper estimates to (I.1).