Low-Rank Matrix Estimation From Rank-One Projections by Unlifted Convex Optimization

Sohail Bahmani[*]
sohail.bahmani@ece.gatech.edu

Kiryung Lee[†]
lee.8763@osu.edu

January 12, 2021

Abstract

We study an estimator with a convex formulation for recovery of low-rank matrices from rank-one projections. Using initial estimates of the factors of the target $d_1 \times d_2$ matrix of rank-r, the estimator admits a practical subgradient method operating in a space of dimension $r(d_1 + d_2)$. This property makes the estimator significantly more scalable than the convex estimators based on lifting and semidefinite programming. Furthermore, we present a streamlined analysis for exact recovery under the real Gaussian measurement model, as well as the partially derandomized measurement model by using the spherical t-design. We show that under both models the estimator succeeds, with high probability, if the number of measurements exceeds $r^2(d_1 + d_2)$ up to some logarithmic factors. This sample complexity improves on the existing results for nonconvex iterative algorithms.

1 Introduction

We consider the problem of estimating a matrix $\boldsymbol{M}_0 \in \mathbb{C}^{d_1 \times d_2}$ of known rank $r \ll \min\{d_1, d_2\}$ from rank-one "sketches" of the form

$$m_i = \boldsymbol{a}_i^* \boldsymbol{M}_0 \boldsymbol{b}_i, \qquad\qquad i = 1, \dots, n, \qquad\qquad (1)$$

for random vectors $\boldsymbol{a}_i \in \mathbb{C}^{d_1}$ and $\boldsymbol{b}_i \in \mathbb{C}^{d_2}$ drawn from certain distributions. More specifically, given the observations $\{(\boldsymbol{a}_i, \boldsymbol{b}_i, m_i)\}_{i=1}^n$, the goal is to estimate factors $\boldsymbol{X}_0 \in \mathbb{C}^{d_1 \times r}$ and $\boldsymbol{Y}_0 \in \mathbb{C}^{d_2 \times r}$ of $\boldsymbol{M}_0$ (i.e., $\boldsymbol{M}_0 = \boldsymbol{X}_0 \boldsymbol{Y}_0^*$).

Depending on the distribution of $(\boldsymbol{a}_i, \boldsymbol{b}_i)_i$, the observation model (1) can describe various low-rank matrix recovery problems including matrix completion [11, 25, 30, 28, 46], phase retrieval [12, 13], blind deconvolution and calibration [1, 10], and sketching [27, 9, 7, 23] to name a few. There are various algorithms proposed in the literature for these problems that can be broadly categorized as follows: the algorithms based on *semidefinite relaxation* [11, 25], the iterative methods based on variants of nonconvex gradient descent [30, 46, 37], *alternating minimization* [28], or approximate message passing [41]. We refer the interested reader to the survey papers [16] and [15] for a broader view of the low-rank matrix recovery literature.

Semidefinite relaxations of the low-rank matrix recovery provide the state of the art sample complexity with linear scaling in the rank and no dependence on the condition number. From the algorithmic perspective, however, these methods suffer from a high computational cost, and

---

[*]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332

[†]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210

more importantly memory usage. This drawback motivated a suite of nonconvex approaches with comparable but weaker sample complexity guarantees [15]. Another line of research, originally studied for phase retrieval, proposed recovery by a convex program that avoids lifting and the semidefinite constraint altogether [3, 24, 4, 2]. This new convex approach admits inherent robustness due to convex geometry of the optimization formulation together with flexibility in adopting a number of off-the-shelf numerical convex optimization algorithms. We apply this framework to the low-rank desketching problem in this paper. We extend and streamline the analysis of the special case of rank-1 recovery with Gaussian factors provided in [2], to a more general low-rank recovery problem.

1.1 Anchored regression

Let $\widetilde{\boldsymbol{X}}_0$ and $\widetilde{\boldsymbol{Y}}_0$ be a pair of matrices for which $\widetilde{\boldsymbol{X}}_0 \widetilde{\boldsymbol{Y}}_0^*$ approximates the ground truth matrix $\boldsymbol{M}_0$. With $\langle \cdot, \cdot \rangle$ being used throughout the paper to denote the real-valued inner product defined as

$$\langle \boldsymbol{U}, \boldsymbol{V} \rangle \stackrel{\text{def}}{=} \operatorname{Re}\left(\operatorname{tr}(\boldsymbol{U}^*\boldsymbol{V})\right),$$

our proposed estimator is formulated as

$$(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}) \in \underset{\boldsymbol{X} \in \mathbb{C}^{d_1 \times r}, \boldsymbol{Y} \in \mathbb{C}^{d_2 \times r}}{\operatorname{argmax}} \langle \widetilde{\boldsymbol{X}}_0, \boldsymbol{X} \rangle + \langle \widetilde{\boldsymbol{Y}}_0, \boldsymbol{Y} \rangle - \frac{1}{n} \sum_{i=1}^{n} \ell_i(\boldsymbol{X}, \boldsymbol{Y}), \tag{2}$$

where

$$\ell_i(\boldsymbol{X}, \boldsymbol{Y}) \stackrel{\text{def}}{=} \frac{1}{2}\|\boldsymbol{X}^*\boldsymbol{a}_i\|^2 + \frac{1}{2}\|\boldsymbol{Y}^*\boldsymbol{b}_i\|^2 + |\boldsymbol{a}_i^*\boldsymbol{X}\boldsymbol{Y}^*\boldsymbol{b}_i - m_i|.$$

The optimization in (2) is effectively a convex program and can be solved efficiently. To clarify this fact, observe that the functions $\ell_i(\boldsymbol{X}, \boldsymbol{Y})$ can be written equivalently as

$$\ell_i(\boldsymbol{X}, \boldsymbol{Y}) = \sup_{\phi \colon |\phi|=1} \frac{1}{2}\|\boldsymbol{X}^*\boldsymbol{a}_i\|^2 + \frac{1}{2}\|\boldsymbol{Y}^*\boldsymbol{b}_i\|^2 + \operatorname{Re}(\phi\,(\boldsymbol{a}_i^*\boldsymbol{X}\boldsymbol{Y}^*\boldsymbol{b}_i - m_i)) \tag{3}$$

$$= \sup_{\phi \colon |\phi|=1} \frac{1}{2}\|\boldsymbol{X}^*\boldsymbol{a}_i + \phi\,\boldsymbol{Y}^*\boldsymbol{b}_i\|^2 - \operatorname{Re}(\phi\,m_i)$$

For any fixed ϕ, the argument of the supremum is clearly convex in $[\boldsymbol{X}; \boldsymbol{Y}] \in \mathbb{C}^{(d_1+d_2)\times r}$. Therefore, $\ell_i(\boldsymbol{X}, \boldsymbol{Y})$ is also a convex function of $[\boldsymbol{X}; \boldsymbol{Y}]$, meaning that (2) is a convex program.

Because of the specific form of the loss functions $\ell_i(\boldsymbol{X}, \boldsymbol{Y})$, the estimator in (2) can be viewed as a "convexification" of the (nonconvex) least absolute deviation (LAD) estimator by quadratic regularization. Previously, [2] has studied similar estimators for observations in the form of difference of convex functions, with the bilinear observations for rank-1 matrices as a special case. In this paper we provide a streamlined analysis of the estimator tailored to desketching of a low-rank matrix from its rank-one measurements in (1). The following provides the high-level description of the sample complexity we have established for the anchored regression estimator. The precise statements are provided in Theorems 1 and 2.

Theorem. *Given a "good" approximation of the unknown rank-r matrix $\boldsymbol{M}_0$ as $\widetilde{\boldsymbol{X}}_0\widetilde{\boldsymbol{Y}}_0^* \approx \boldsymbol{M}_0$, with high probability, the proposed anchored regression recovers $\boldsymbol{M}_0$ exactly, from $O(r(d_1+d_2)\operatorname{polylog}(d_1+d_2))$ random rank-one measurements. The hidden constant in the sample complexity depends also on the "low-rank condition number" of $\boldsymbol{M}_0$.*

The sufficient number of samples for the exact recovery provided by the theorem is near optimal compared to the degrees of freedom of the rank-r matrix model. The random sketching models and the size of the neighborhood will be specified in the following section.

1.2 Sketching Models

We study the sample complexity of our proposed estimator under two different random measurements models. The first model, which we refer to as the *real Gaussian sketching*, simply uses the outer product of two independent Gaussian vectors as the rank-one sketching matrix. The second model, called the *partially derandomized sketching*, mimics the behavior of the firs model but the random rank-one sketching matrix takes realizations from a finite set. More precisely, the first few moments of the sketching matrix of the second model are designed to coincide with those of the real Gaussian model.

1.2.1 Real Gaussian sketching

This model simply considers the random vectors $[\boldsymbol{a}_i;\ \boldsymbol{b}_i] \in \mathbb{R}^{d_1+d_2}$ to be independent copies of a random vector $[\boldsymbol{a};\ \boldsymbol{b}] \in \mathbb{R}^{d_1+d_2}$ satisfying

$$[\boldsymbol{a};\ \boldsymbol{b}] \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_{d_1+d_2}\right). \tag{4}$$

1.2.2 Partially derandomized sketching

Next we consider a measurement distribution supported on a special finite set whose first $2t$ moments coincide with those of the real Gaussian model. A configuration of such a set is called a *complex projective t-design* and designs were first introduced by Delsarte et al. [19]. A variant of this model has been previously employed for the phase retrieval [5, 26, 32]. The concept of t-design has been utilized in coding theory and quantum information theory, particularly in the analysis of randomized algorithms. One drawback is the size of the set generally grows exponentially in the dimension while the exponent is proportional to the parameter t. Concrete constructions are widely available for the special case with degree 2 and numerical algorithms for an approximate design for higher order are also available in the literature. We refer to [26] for more details on the t-design model and the related references. Below we describe the version of the model that is relevant for our purposes.

Let P_{Sym^t} denote the totally symmetric subspace of $(\mathbb{C}^d)^{\otimes t}$ such that all elements are invariant under every possible permutation of t factors (see, e.g., [33]). Then a weighted t-design is defined as follows [26].

Definition 1. *Let $t \in \mathbb{N}$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N \in \mathbb{C}^d$ be unit vectors. The set $\{\boldsymbol{w}_i\}_{i=1}^N$ with corresponding weights $\{p_i\}_{i=1}^N$ such that $p_i \geq 0$ for all $i = 1, \ldots, N$, and $\sum_{i=1}^N p_i = 1$ is a weighted complex projective t-design of dimension n and cardinality N, if*

$$\sum_{i=1}^N p_i \left(\boldsymbol{w}_i \boldsymbol{w}_i^*\right)^{\otimes t} = \binom{d+t-1}{t}^{-1} P_{\mathrm{Sym}^t},$$

where P_{Sym^t} denotes the projector onto the totally symmetric subspace Sym^t of $(\mathbb{C}^d)^{\otimes t}$.

Our second sketching model is given by the concatenation of two independent random vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in the following construction: Given a weighted t-design $\{(\boldsymbol{w}_i, p_i)\}_{i=1}^{N_1}$ in $\mathbb{C}^{d_1}$ with $t \geq 2$, let $\boldsymbol{a}$ be a random vector given by

$$\mathbb{P}\left\{\boldsymbol{a} = \sqrt{d_1}\boldsymbol{w}_i\right\} = p_i, \quad i = 1, \ldots, N_1. \tag{5}$$

Then $\boldsymbol{a}$ satisfies

$$\mathbb{E}(\boldsymbol{a}\boldsymbol{a}^*)^{\otimes t} = d_1^t \binom{d_1+t-1}{t}^{-1} P_{\mathrm{Sym}^t}. \tag{6}$$

Similarly, given a weighted t-design $\{(\boldsymbol{w}_i', p_i')\}_{i=1}^{N_2}$ in $\mathbb{C}^{d_2}$ with $t \geq 2$, let $\boldsymbol{b}$ be a random vector given by

$$\mathbb{P}\left\{\boldsymbol{b} = \sqrt{d_2}\boldsymbol{w}_i'\right\} = p_i', \quad i = 1, \ldots, N_2. \tag{7}$$

Then $\boldsymbol{b}$ satisfies

$$\mathbb{E}(\boldsymbol{b}\boldsymbol{b}^*)^{\otimes t} = d_2^t \binom{d_2 + t - 1}{t}^{-1} P_{\mathrm{Sym}^t}. \tag{8}$$

By construction, a weighted t-design is automatically a weighted t'-design for all $t' \leq t$. Given accurate anchor matrices, a performance guarantee for (2) can be derived only requiring the moment conditions of up to the 4th order. However, obtaining such accurate anchor matrices require a more stringent condition on higher moments of up to order $2t = \Omega(\log(d_1 + d_2))$. This is why we assume that $t \geq 2$ in constructing partially derandomized measurement vectors with t-designs.

1.3 Spectral initialization

Our main results rely on the availability of $\widetilde{\boldsymbol{X}}_0$ and $\widetilde{\boldsymbol{Y}}_0^*$ such that $\widetilde{\boldsymbol{X}}_0\widetilde{\boldsymbol{Y}}_0^*$ is close to the ground truth matrix $\boldsymbol{M}_0$. To provide a stand-alone theory that does not require any oracle information, we also analyze a specific method to obtain such matrices $\widetilde{\boldsymbol{X}}_0$ and $\widetilde{\boldsymbol{Y}}_0^*$ described below.

Let $\mathcal{A} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^n$ denote the linear operator representing the rank-one measurements in (1), i.e., it is defined by

$$\boldsymbol{M} \mapsto \mathcal{A}(\boldsymbol{M}) = \left(\frac{1}{\sqrt{n}}\boldsymbol{a}_i^*\boldsymbol{M}\boldsymbol{b}_i\right)_{i=1}^n.$$

Then its adjoint operator, denoted by $\mathcal{A}^*$, is given by

$$\boldsymbol{y} \mapsto \mathcal{A}^*(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{a}_i\boldsymbol{b}_i^*.$$

The *spectral method* computes an estimate $\widetilde{\boldsymbol{M}}_0$ of the unknown matrix $\boldsymbol{M}_0$ as the best rank-r approximation of $\mathcal{A}^*\mathcal{A}(\boldsymbol{M}_0)$ with respect to the Frobenius norm. Under the two random sketching models, we obtain suitable upper bounds on the approximation error through matrix concentration inequalities.

Next we factorize the estimated matrix into $\widetilde{\boldsymbol{M}}_0 = \widetilde{\boldsymbol{X}}_0\widetilde{\boldsymbol{Y}}_0^*$ through the singular value decomposition. Let $\widetilde{\boldsymbol{M}}_0 = \widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{\Sigma}}_0\widetilde{\boldsymbol{V}}_0^*$ be the compact singular value decomposition of $\widetilde{\boldsymbol{M}}_0$. Then we choose $\widetilde{\boldsymbol{X}}_0$ and $\widetilde{\boldsymbol{Y}}_0$ by

$$\widetilde{\boldsymbol{X}}_0 = \widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{\Sigma}}_0^{1/2} \quad \text{and} \quad \widetilde{\boldsymbol{Y}}_0 = \widetilde{\boldsymbol{V}}_0\widetilde{\boldsymbol{\Sigma}}_0^{1/2},$$

so that they have the same singular values. This particular decomposition provides a set of useful properties, such as the identity

$$\widetilde{\boldsymbol{\Sigma}}_0 = \widetilde{\boldsymbol{X}}_0^*\widetilde{\boldsymbol{X}}_0 = \widetilde{\boldsymbol{Y}}_0^*\widetilde{\boldsymbol{Y}}_0,$$

that are utilized in the proof of our main results.

The precise statement of the requirements for the spectral initialization, and the corresponding sample complexity under the two considered measurement models are provided in Section 2. The pertaining derivations are provided in Section 6.

1.4 Discussion and Related work

Under the real Gaussian sketching model and given an initial estimate satisfying (9), we demonstrate that, with high probability, the estimator in (2) recovers $\boldsymbol{M}_0$ exactly, provided the number of measurements scales as $n \geq Cdr$, where $d = \max(d_1, d_2)$. This sample complexity coincides with the sample complexity achieved by the estimators based on lifting and semidefinite relaxation [14, 9]. On the other hand, our estimator is formulated through an explicit factorization only with $r(d_1+d_2)$ variables while the lifted convex estimator over $d_1 d_2$ variables [14, 9]. Furthermore, because the methods based on semidefinite relaxation do not operate in the factorized domain, they often need singular value calculations which further complicates their scalability.

However, computationally inexpensive methods used to find the initial estimates obeying (9), often lead to a suboptimal overall sample complexity. In fact, we show that, with high probability, the spectral initialization succeeds if $n \geq Cdr^2$ which dominates the sample complexity $n \geq Cdr$ for the "oracle-assisted" estimator mentioned above.

The *iterative hard thresholding* algorithm is studied under variants of the restricted isometry property for low-rank matrices in [28] and [23]. This algorithm is computationally less expensive than the generic convex optimization algorithms that solve the semidefinite relaxation, because it only requires to perform low-rank SVDs in its iterations rather than the full SVDs. However, the fact that the iterative hard thresholding method operates in the lifted domain, is an obstruction to its scalability. Several other iterative algorithms have been proposed and analyzed under the real Gaussian sketching model. Earlier methods used *resampling* to draw fresh measurements per iteration. Therefore, these methods need to terminate after finitely many iterations, which only allows for approximate recovery up to a prescribed accuracy ϵ. Prior work on this approach achieve the sample complexities $O(dr^4 \log^2 d \log(1/\epsilon))$ [51], $O(dr^3 \log(1/\epsilon))$ [36], and $O(dr^2 \log^4 d \log(1/\epsilon))$ [44]. In more recent work, [42] and [35] studied performance of the nonconvex gradient descent and established the sample complexities $O(dr^6 \log^2 d)$ and $O(dr^4 \log d)$, respectively. Our estimator outperforms these results for nonconvex approaches. In fact, our estimator would have achieved the ideal sample complexity should there be an initialization with the sample complexity $O(dr)$. The hidden constant in this sample complexity, similar to the sample complexity of the nonconvex methods, depends on the "low-rank condition number" of the ground truth matrix, defined precisely below in Section 2. It is also worthwhile to mention that the existing results in the literature often focus on the case where the rank-one measurement matrices or the ground truth matrix are symmetric. The model (1) considered in this paper allows for a general choice of the measurement factors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$. For simplicity, we only consider independent factors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, but the provided framework can be adapted to the case of dependent factors by modifying some of the relevant calculations.

2 Main results

Our main results demonstrate how many observations suffice for the estimator (2) to reconstruct the unknown matrix $\boldsymbol{X}_0 \boldsymbol{Y}_0^*$. Our first theorem provides a sample complexity that guarantees accuracy of the estimator (2) under the real Gaussian sketching model. Throughout we use $\kappa \geq 1$ to denote the (low-rank) condition number of $\boldsymbol{M}_0$, which refers to the ratio of the largest and smallest *non-zero* singular values of $\boldsymbol{M}_0$, i.e.,

$$\kappa = \frac{\sigma_1(\boldsymbol{M}_0)}{\sigma_r(\boldsymbol{M}_0)}.$$

Theorem 1 (Real Gaussian desketching)**.** *Let $([\boldsymbol{a}_i;\ \boldsymbol{b}_i])_{i=1}^n$ be independent copies of $[\boldsymbol{a};\ \boldsymbol{b}] \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_1+d_2})$. Let $\widetilde{\boldsymbol{X}}_0 \in \mathbb{C}^{d_1 \times r}$ and $\widetilde{\boldsymbol{Y}}_0 \in \mathbb{C}^{d_2 \times r}$ be matrices that satisfy $\widetilde{\boldsymbol{X}}_0^* \widetilde{\boldsymbol{X}}_0 = \widetilde{\boldsymbol{Y}}_0^* \widetilde{\boldsymbol{Y}}_0$ and*

$$\left\| \widetilde{\boldsymbol{X}}_0 \widetilde{\boldsymbol{Y}}_0^* - \boldsymbol{M}_0 \right\| \lesssim r^{-1/2} \kappa^{-2} \|\boldsymbol{M}_0\| . \tag{9}$$

If the number of measurements n obeys

$$n \gtrsim \max\{\kappa r(d_1 + d_2), \log(1/\delta)\} , \tag{10}$$

then with probability at least $1 - \delta$ the estimates $\widehat{\boldsymbol{X}}$ and $\widehat{\boldsymbol{Y}}$ obtained by the anchored regression satisfy $\widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}^* = \boldsymbol{M}_0$.

The result by Theorem 1 is comparable to the analogous result for the lifted convex optimization by nuclear norm minimization [14, 9]. However, the dependence on the condition number, which does not appear in the lifted case, is the cost we need to pay to save the computation through explicit factorization.

We also present the sample complexity for the success of the spectral initialization under the same model.

Proposition 1. *Let $([\boldsymbol{a}_i;\ \boldsymbol{b}_i])_{i=1}^n$ be independent copies of $[\boldsymbol{a};\ \boldsymbol{b}] \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_1+d_2})$. Then the estimate $(\widetilde{\boldsymbol{X}}_0, \widetilde{\boldsymbol{Y}}_0)$ by the spectral initialization satisfies* (9) *with probability at least $1 - (d_1 + d_2)^{-\alpha}$, if*

$$n \gtrsim \alpha^3 \kappa^4 r^2 (d_1 + d_2) \log^3(d_1 + d_2) .$$

As shown in Theorem 1 and Proposition 1, the number of samples enough for the success of the spectral initialization dominates that for the estimator. Although the spectral initialization is just one approach to obtain an initial estimate satisfying (9), it has not been shown any alternative practical method providing the same accuracy from fewer measurements.

Next we present the corresponding results for partially derandomized sketching below.

Theorem 2 (Partially derandomized desketching)**.** *Let $\boldsymbol{a}$ and $\boldsymbol{b}$ are independent random vectors uniformly distributed over the corresponding t-design sets according to* (5) *and* (7) *with $t \geq 2$. Let $([\boldsymbol{a}_i;\ \boldsymbol{b}_i])_{i=1}^n$ be independent copies of $[\boldsymbol{a};\ \boldsymbol{b}]$. Let $\widetilde{\boldsymbol{M}}_0$, $\widetilde{\boldsymbol{X}}_0$, and $\widetilde{\boldsymbol{Y}}_0$ as in Theorem 1 satisfying* (9)*. If the number of measurements n satisfies*

$$n \gtrsim \kappa r(d_1 + d_2) \max\{\log(d_1 + d_2), \log(1/\delta)\} , \tag{11}$$

then the anchored regression exactly recovers $\boldsymbol{M}_0$ with probability at least $1 - \delta$.

Proposition 2. *Under the sketching model in Theorem 2, the spectral initialization provides $(\widetilde{\boldsymbol{X}}_0, \widetilde{\boldsymbol{Y}}_0)$ satisfying* (9) *with probability at least $1 - (d_1 + d_2)^{-\alpha}$ provided*

$$n \gtrsim t\kappa^4 r^2 (d_1 + d_2)^{1+1/2t+\alpha/t} \vee \left(t^2 \kappa^2 r^{3/2} (d_1 + d_2)^{1+\alpha/2t} \right)^{2t/(2t-1)} . \tag{12}$$

Particularly, if $t \geq (\alpha + 1/2) \ln(d_1 + d_2)$, then the condition in (12) *simplifies to*

$$n \gtrsim \alpha^2 \kappa^4 r^2 (d_1 + d_2) \ln^4(d_1 + d_2) .$$

Compared to the real Gaussian sketching model, the derandomized case is guaranteed by slightly more measurements (larger by a logarithmic factor).

3 Numerical Results

A set of Monte Carlo numerical results are provided to illustrate that the empirical behavior of the estimator is consistent with the main theoretical results. We first discuss how the convex program in (2) can be solved by a practical numerical algorithm. Recall that the estimator in (2) is equivalently written as

$$(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}) \in \operatorname*{argmin}_{\boldsymbol{X} \in \mathbb{C}^{d_1 \times r}, \boldsymbol{Y} \in \mathbb{C}^{d_2 \times r}} f(\boldsymbol{X}, \boldsymbol{Y}), \tag{13}$$

where the convex objective function is given by

$$f(\boldsymbol{X}, \boldsymbol{Y}) = -\langle \widetilde{\boldsymbol{X}_0}, \boldsymbol{X} \rangle - \langle \widetilde{\boldsymbol{Y}_0}, \boldsymbol{Y} \rangle - \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \|\boldsymbol{X}^* \boldsymbol{a}_i\|^2 + \frac{1}{2} \|\boldsymbol{Y}^* \boldsymbol{b}_i\|^2 + |\boldsymbol{a}_i^* \boldsymbol{X} \boldsymbol{Y}^* \boldsymbol{b}_i - m_i| \right).$$

A simple *subgradient method* can be used to find a minimizer to (13). The estimator is refined by

$$\begin{bmatrix} \boldsymbol{X}_{t+1} \\ \boldsymbol{Y}_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix} - \eta_t \boldsymbol{G}_t,$$

where η_t denotes the step size at the tth iteration and $\boldsymbol{G}_t \in \mathbb{C}^{(d_1+d_2) \times r}$ is a subgradient of f at $[\boldsymbol{X}_t \ \boldsymbol{Y}_t]$ specifically given by

$$\boldsymbol{G}_t = -\begin{bmatrix} \widetilde{\boldsymbol{X}_0} \\ \widetilde{\boldsymbol{Y}_0} \end{bmatrix} + \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \boldsymbol{a}_i \boldsymbol{a}_i^* & \phi_i^* \boldsymbol{a}_i \boldsymbol{b}_i^* \\ \phi_i \boldsymbol{b}_i \boldsymbol{a}_i^* & \boldsymbol{b}_i \boldsymbol{b}_i^* \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix},$$

where

$$\phi_i = \exp\left(\imath \operatorname{Arg}(\boldsymbol{a}_i^* \boldsymbol{X}_t \boldsymbol{Y}_t^* \boldsymbol{b}_i - m_i) \right),$$

with $\imath = \sqrt{-1}$ denoting imaginary unit, and $\operatorname{Arg}(z)$ denoting the principal argument of $z \in \mathbb{C}$. The per-iteration-cost of this subgradient method is comparable to that of the nonconvex gradient descent. We use the diminishing step size rule for $(\eta_t)_t$. We chose the simplest algorithm to solve the convex program in (13). However, we believe that more sophisticated optimization algorithms are also applicable to our problem. For example, in expression of $\ell_i(\cdot, \cdot)$ as (3) the constraint $|\phi| = 1$ can be relaxed to $|\phi| \leq 1$ without affecting the function value. Then, we can show that the proposed estimator can be equivalently formulated as a convex-concave saddle-point problem, which can be solved using algorithms based on *mirror descent* and *mirror prox* [45, Chs. 5 and 6], [8, Ch. 5].

The first set of simulation provides the empirical phase transition as a function of the rank while the other parameters are fixed ($d_1 = d_2 = 128$ and $n = d_1 d_2 / 4$). The measurements are generated by the standard complex normal distribution. Theorem 1 shows that, if $\widetilde{\boldsymbol{X}_0} \widetilde{\boldsymbol{Y}_0}^*$ is close enough to the ground-truth matrix $\boldsymbol{M}_0$, the maximum rank that leads to the exact recovery is determined by (10). In order to consider the effect of the accuracy of the anchor, we introduce a parameter $\alpha \in [0, 1]$ to linearly interpolate between the spectral initialization, corresponding to $\alpha = 1$, and the ground truth, corresponding to $\alpha = 0$. Figure 1 illustrates the empirical phase transitions, with the left and right panels respectively showing the median and 90th percentile of the relative error over 100 trials. Exact recovery is achieved in most cases for $r \leq 2$. The top rows in Figure 1 shows the empirical recovery phase transition when $(\widetilde{\boldsymbol{X}_0}, \widetilde{\boldsymbol{Y}_0})$ is given by the spectral initialization, whereas the bottom rows correspond to the ground truth chosen as the anchor. Figure 1 suggests that at larger r the phase transition occurs at smaller α, thereby requiring a more accurate initial estimate $\widetilde{\boldsymbol{X}_0} \widetilde{\boldsymbol{Y}_0}^*$, which is consistent with the requirement in (9).

7

(a) Median   (b) 90th percentile
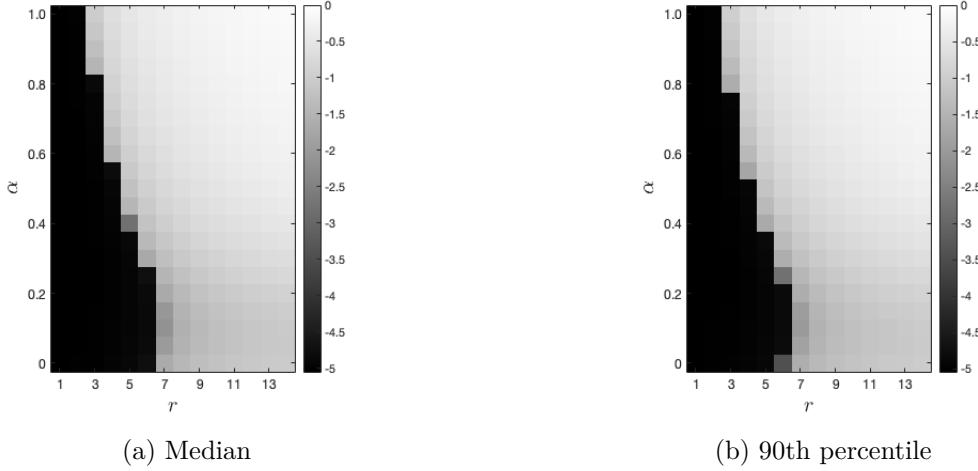
Figure 1: Empirical phase transition in the noiseless case. The logarithm base 10 of percentiles of the normalized estimation error $\|\widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}^* - \boldsymbol{M}_0\|_{\mathrm{F}}/\|\boldsymbol{M}_0\|_{\mathrm{F}}$ is plotted. The size of matrix and the number of measurements are set to $d_1 = d_2 = 128$ and $n = d_1 d_2/4$, respectively. The anchor matrix is computed from a convex combination of the rank-r matrix in the spectral initialization and the ground truth respectively with weights α and $1 - \alpha$. Thus $\alpha = 1$ denotes the case with the spectral initialization.

The second simulation illustrates how the additive noise to measurements propagates to the estimation error. We consider the regime of parameters where the convex estimator provides exact recovery in the noiseless case. We set $d_1 = d_2 = 128$, $r = 2$, and $n = d_1 d_2/4$. The 90th percentile of 100 realizations was observed while the signal-to-noise ratio (SNR) varies over 5 to 50 dB. Figure 2 shows that the estimation error decreases gradually as SNR increases. In other words, the estimation error in the presence of measurement noise scales smoothly as a function of SNR. This stable performance of the estimator is due to the nice geometry of the convex program in (13).
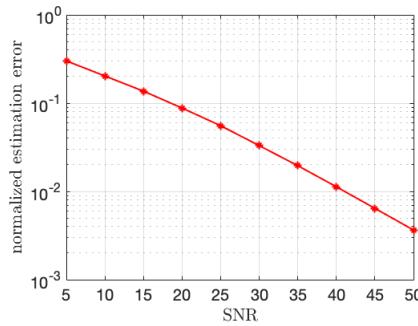
Figure 2: Estimation error for varying SNR. The observation is corrupted with additive Gaussian noise so that $m_i = \boldsymbol{a}_i^* \boldsymbol{M}_0 \boldsymbol{b}_i + \zeta_i$ for $i = 1, \ldots, n$ with $\zeta_1, \ldots, \zeta_n$ being i.i.d. $\mathcal{N}(0, \sigma^2)$. SNR is defined as $10 \log_{10}(\sum_{i=1}^n m_i^2/\sigma^2)$. The 90th percentile of the normalized estimation error is plotted. The size of matrix, rank, and number of measurements are set to $d_1 = d_2 = 128$, $r = 2$, and $n = d_1 d_2/4$, respectively. The anchor matrix is computed by the spectral initialization.

4 Proof of the Main Theorems

We prove Theorems 1 and 2 in two steps. First, we derive a set of deterministic conditions to guarantee exact recovery for the proposed estimator. Then, we show that these conditions hold, with high probability, for the sketching models introduced in Section 1.2.

4.1 A deterministic sufficient condition for exact recovery

Recall that $\boldsymbol{M}_0 = \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 \boldsymbol{V}_0^*$ is the compact singular decomposition of the ground truth matrix $\boldsymbol{M}_0$, where both $\boldsymbol{U}_0 \in \mathbb{C}^{d_1 \times r}$ and $\boldsymbol{V}_0 \in \mathbb{C}^{d_2 \times r}$ have orthonormal columns, and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{r \times r}$ is the diagonal matrix of the singular values. The support space of $\boldsymbol{M}_0$, denoted by $\mathcal{T}$, is defined as

$$\mathcal{T} \stackrel{\text{def}}{=} \left\{ \boldsymbol{\Delta}_1 \boldsymbol{V}_0^* + \boldsymbol{U}_0 \boldsymbol{\Delta}_2^* \ : \ \boldsymbol{\Delta}_i \in \mathbb{C}^{d_i \times r}, \ i = 1, 2 \right\} .$$

With these notations, the following proposition provides a set of deterministic sufficient conditions for the estimator (2) to exactly recover $\boldsymbol{M}_0$. The proof is deferred to Section 5.

Proposition 3. *Let r and κ respectively denote the rank and the condition number of $\boldsymbol{M}_0$, and $(\widetilde{\boldsymbol{X}}_0, \widetilde{\boldsymbol{Y}}_0)$ be a given pair of matrices that satisfy $\widetilde{\boldsymbol{X}}_0^* \widetilde{\boldsymbol{X}}_0 = \widetilde{\boldsymbol{Y}}_0^* \widetilde{\boldsymbol{Y}}_0$. Furthermore, suppose that there exist $\rho \in (0, 1]$, and absolute constants $C_1, C_2 \geq 1$ such that*

$$\frac{1}{n} \sum_{i=1}^{n} |\langle \boldsymbol{a}_i \boldsymbol{b}_i^*, \boldsymbol{H} \rangle| \geq \rho \|\boldsymbol{H}\|_{\mathrm{F}}, \quad \text{for all } \boldsymbol{H} \in \mathcal{T}, \tag{14}$$

$$\left\| \boldsymbol{I}_{d_1} - \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{a}_k \boldsymbol{a}_k^* \right\| \leq \frac{\rho}{C_1 \sqrt{r\kappa}}, \tag{15}$$

$$\left\| \boldsymbol{I}_{d_2} - \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{b}_k \boldsymbol{b}_k^* \right\| \leq \frac{\rho}{C_1 \sqrt{r\kappa}}, \tag{16}$$

and

$$\left\| \widetilde{\boldsymbol{X}}_0 \widetilde{\boldsymbol{Y}}_0^* - \boldsymbol{M}_0 \right\| \leq \frac{\rho \|\boldsymbol{M}_0\|}{C_2 \sqrt{r} \kappa^2} . \tag{17}$$

Then, the maximizer $(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}})$ to (2) is unique and satisfies $\widehat{\boldsymbol{X}} \widehat{\boldsymbol{Y}}^* = \boldsymbol{M}_0$.

4.2 Verifying the sufficient condition under the random models

We demonstrate that under the random measurement models introduced in Section 2 the assumptions made in Proposition 3 hold with high probability.

4.2.1 Small-ball method

Our analysis is based on the *small-ball method* [31, 39]. A simple exposition and some applications of this method can be found in [48] and [21]. We provide the proofs for the manuscript to be self-contained as well as addressing some subtle but important differences due to operation in the complex domain.

We first show in the following proposition that (14) is satisfied with high probability.

Proposition 4 (Lower-tail via the small-ball method). *Let $\mathcal{T}$ be a subset of $\mathbb{C}^{d_1 \times d_2}$ that is invariant under multiplication by unit-modulus scalars. For i.i.d. and isotropic random vectors $[\boldsymbol{a};\ \boldsymbol{b}]$, $[\boldsymbol{a}_1;\ \boldsymbol{b}_1], \ldots, [\boldsymbol{a}_n;\ \boldsymbol{b}_n] \in \mathbb{C}^{d_1+d_2}$ define*

$$p_\tau(\mathcal{T}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{H} \in \mathcal{T} \setminus \{\boldsymbol{0}\}} \mathbb{P}\left\{ |\boldsymbol{a}^* \boldsymbol{H} \boldsymbol{b}| \geq \tau \|\boldsymbol{H}\|_{\text{F}} \right\},$$

$$\mathfrak{C}_n(\mathcal{T}) \stackrel{\text{def}}{=} \mathbb{E} \sup_{\boldsymbol{H} \in \mathcal{T} \setminus \{\boldsymbol{0}\}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\varepsilon_i \langle \boldsymbol{a}_i \boldsymbol{b}_i^*, \boldsymbol{H} \rangle}{\|\boldsymbol{H}\|_{\text{F}}},$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables independent of everything else. Then, for any $\tau > 0$ and $\delta \in (0,1)$, with probability at least $1 - \delta$, we have

$$\inf_{\boldsymbol{H} \in \mathcal{T} \setminus \{\boldsymbol{0}\}} \frac{1}{n} \sum_{i=1}^n \frac{|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|}{\|\boldsymbol{H}\|_{\text{F}}} \geq \tau p_\tau(\mathcal{T}) - \frac{\pi \, \mathfrak{C}_n(\mathcal{T})}{\sqrt{n}} - \tau \sqrt{\frac{\log(1/\delta)}{n}}. \tag{18}$$

Proof. Using $[z]_{\leq t} \stackrel{\text{def}}{=} \min\{z, t\}$ to denote the "saturation" at t, for any $\tau > 0$, we have

$$\frac{1}{n} \sum_{i=1}^n |\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i| \geq \frac{1}{n} \sum_{i=1}^n [|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|]_{\leq \tau \|\boldsymbol{H}\|_{\text{F}}}, \tag{19}$$

for every $\boldsymbol{H} \in \mathbb{R}^{d_1 \times d_2}$. By normalizing by $\|\boldsymbol{H}\|_{\text{F}}$, it suffices to find a lower bound for the right-hand side of (19) for all $\boldsymbol{H} \in \mathcal{T} \bigcap \mathbb{S}$, where $\mathbb{S}$ denotes the unit sphere of the Frobenius norm in $\mathbb{C}^{d_1 \times d_2}$.

Adding and subtracting $\mathbb{E}\left(|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|_{\leq \tau}\right)$, and using the fact that $\mathbb{E}\left([|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|]_{\leq \tau}\right) \geq \tau \, \mathbb{P}\left(|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i| \geq \tau\right)$, we can write

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n [|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|]_{\leq \tau} &\geq \frac{1}{n} \sum_{i=1}^n \tau \, \mathbb{P}\left(|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i| \geq \tau\right) \\
&\quad - \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left([|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|]_{\leq \tau}\right) - [|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|]_{\leq \tau}.
\end{aligned} \tag{20}$$

The function $F : (\mathbb{C}^d)^n \to \mathbb{R}_{\geq 0}$ defined as

$$F\left([\boldsymbol{a}_1;\ \boldsymbol{b}_1], \ldots, [\boldsymbol{a}_n;\ \boldsymbol{b}_n]\right) \stackrel{\text{def}}{=} \sup_{\boldsymbol{H} \in \mathcal{T} \bigcap \mathbb{S}} \frac{1}{n} \left| \sum_{i=1}^n \mathbb{E}\left([|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|]_{\leq \tau}\right) - [|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|]_{\leq \tau} \right|$$

has the bounded difference property. Therefore, invoking the *bounded difference inequality* [38], with probability at least $1 - \delta$, we have

$$F\left([\boldsymbol{a}_1;\ \boldsymbol{b}_1], \ldots, [\boldsymbol{a}_n;\ \boldsymbol{b}_n]\right) \leq \mathbb{E} F\left([\boldsymbol{a}_1;\ \boldsymbol{b}_1], \ldots, [\boldsymbol{a}_n;\ \boldsymbol{b}_n]\right) + \tau \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{21}$$

Using the standard Giné-Zinn symmetrization argument (e.g., see [49, Lemma 2.3.1]), the expectation on the right-hand side of (21) can be bounded as

$$\mathbb{E} F\left([\boldsymbol{a}_1;\ \boldsymbol{b}_1], \ldots, [\boldsymbol{a}_n;\ \boldsymbol{b}_n]\right) \leq \frac{2}{\sqrt{n}} \mathbb{E}\left(\sup_{\boldsymbol{H} \in \mathcal{T} \bigcap \mathbb{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i [|\boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i|]_{\leq \tau} \right),$$

where the expectation on the right-hand side is with respect to $[\boldsymbol{a}_1;\ \boldsymbol{b}_1],\ldots,[\boldsymbol{a}_n;\ \boldsymbol{b}_n]$ as well as the i.i.d. Rademacher random variables $\varepsilon_1,\ldots,\varepsilon_n$. Since the function $z \mapsto [|z|]_{\leq\tau}$ is 1-Lipschitz, invoking the *Rademacher contraction principle* [34, Theorem 4.12] yields

$$\mathbb{E}\left(\sup_{\boldsymbol{H}\in\mathcal{T}\bigcap\mathbb{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i[|\boldsymbol{a}_i^*\boldsymbol{H}\boldsymbol{b}_i|]_{\leq\tau}\right) \leq \mathbb{E}\left(\sup_{\boldsymbol{H}\in\mathcal{T}\bigcap\mathbb{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\,|\boldsymbol{a}_i^*\boldsymbol{H}\boldsymbol{b}_i|\right). \tag{22}$$

Let ϕ be a unit-modulus scalar in $\mathbb{C}$ that is selected uniformly at random, and $\mathbb{E}_\phi$ denote the expectation with respect to ϕ conditioned on everything else. Straightforward calculus shows that for any $z \in \mathbb{C}$ we have $|z| = (\pi/2)\mathbb{E}_\phi\left(|\mathrm{Re}(\phi^*z)|\right)$. Applying this identity in (22) then yields

$$\begin{aligned}
\mathbb{E}\left(\sup_{\boldsymbol{H}\in\mathcal{T}\bigcap\mathbb{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i[|\boldsymbol{a}_i^*\boldsymbol{H}\boldsymbol{b}_i|]_{\leq\tau}\right) &\leq \frac{\pi}{2}\,\mathbb{E}\left(\sup_{\boldsymbol{H}\in\mathcal{T}\bigcap\mathbb{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\mathbb{E}_\phi\,|\langle\boldsymbol{a}_i\boldsymbol{b}_i^*,\phi\boldsymbol{H}\rangle|\right) \\
&\leq \frac{\pi}{2}\,\mathbb{E}\mathbb{E}_\phi\left(\sup_{\boldsymbol{H}\in\mathcal{T}\bigcap\mathbb{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\,|\langle\boldsymbol{a}_i\boldsymbol{b}_i^*,\phi\boldsymbol{H}\rangle|\right) \\
&\leq \frac{\pi}{2}\,\mathbb{E}\left(\sup_{\boldsymbol{H}\in\mathcal{T}\bigcap\mathbb{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\,|\langle\boldsymbol{a}_i\boldsymbol{b}_i^*,\boldsymbol{H}\rangle|\right) \\
&\leq \frac{\pi}{2}\,\mathfrak{C}_n(\mathcal{T}),
\end{aligned}$$

where the second, third, and fourth lines follow respectively from the Jensen's inequality, the assumption that $\mathcal{T}$ is invariant under multiplication by unit-modulus scalars, and applying the Rademacher contraction principle once more.

Furthermore, since $[\boldsymbol{a}_1;\ \boldsymbol{b}_1],\ldots,[\boldsymbol{a}_n;\ \boldsymbol{b}_n] \in \mathbb{C}^{d_1+d_2}$ are identically distributed, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\left(|\boldsymbol{a}_i^*\boldsymbol{H}\boldsymbol{b}_i| \geq \tau\right) = \mathbb{P}\left(|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}| \geq \tau\right). \tag{23}$$

Therefore, in view of (20), (21), and (23), with probability at least $1-\delta$, for all $\boldsymbol{H} \in \mathcal{T}\cap\mathbb{S}$ we have

$$\frac{1}{n}\sum_{i=1}^{n}[|\boldsymbol{a}_i^*\boldsymbol{H}\boldsymbol{b}_i|]_{\leq\tau} \geq \tau\,\mathbb{P}\left(|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}| \geq \tau\right) - \frac{\pi\,\mathfrak{C}_n(\mathcal{T})}{\sqrt{n}} - \tau\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Recalling the definition of $p_\tau(\mathcal{T})$ is enough to complete the proof. $\qquad\square$

We apply Proposition 4 under the assumptions in either of Theorems 1 and 2. Then (14) is satisfied with high probability provided that the right-hand side of (18) is lower bounded by a nonnegative scalar ρ. The following lemmas provides a lower (resp. upper) bound on $p_\tau(\mathcal{T})$ (resp. $\mathfrak{C}_n(\mathcal{T})$). The proofs are provided in Appendix sections B.1 and B.2.

Lemma 1 (Lower bound on probability). *Let $[\boldsymbol{a};\ \boldsymbol{b}]$ to be a random vector drawn either according to (4), or the pair (5) and (7). Then*

$$p_\tau(\mathcal{T}) \geq c(1-\tau^2)^2$$

for an absolute constant $c > 0$.

Lemma 2 (Upper bound on Rademacher complexity)**.** *Let* $[\boldsymbol{a}; \boldsymbol{b}]$ *satisfy that i)* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *are independent; ii) each of* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *is isotropic. Then*

$$\mathfrak{C}_n(\mathcal{T}) \leq \sqrt{(d_1 + d_2)r}\,,$$

By plugging in the results of the above lemmas, a sufficient condition for satisfying (14) with probability $1 - \delta$ is given by

$$c(1 - \tau^2)^2 - \frac{4\sqrt{(d_1 + d_2)r}}{\sqrt{n}} - \tau\sqrt{\frac{\log(1/\delta)}{2n}} \geq \rho\,. \tag{24}$$

Given ρ, by choosing C in the assumption $n \geq C \max\{r(d_1 + d_2), \log(1/\delta)\}$ large enough and by choosing τ small we obtain that (24) holds with probability $1 - \delta$.

4.2.2 Approximate Isotropy

Next we show that (15) and (16) are satisfied with high probability for both the Gaussian and t-design cases. To simplify the notation, let η denote the right-hand side of (15), which coincides with that of (16), i.e., $\eta = \rho/C_1\sqrt{r\kappa}$. We are interested in the regime where $0 < \eta < 1$.

In the Gaussian case, the concentration of extreme singular values of a Wishart matrix has been well studied in the literature (e.g., see [17, Theorem II.13], which is summarized as Theorem 3 in Appendix). If $[\boldsymbol{a}; \boldsymbol{b}]$ is a standard Gaussian random vector, then we have

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}_i\boldsymbol{a}_i^* - \boldsymbol{I}_{d_1}\right\| > 3\max\left(\sqrt{\frac{4d_1}{n}}, \frac{4d_1}{n}\right)\right\} \leq 2\exp(-d_1/2)$$

and

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{b}_i\boldsymbol{b}_i^* - \boldsymbol{I}_{d_2}\right\| > 3\max\left(\sqrt{\frac{4d_2}{n}}, \frac{4d_2}{n}\right)\right\} \leq 2\exp(-d_2/2)\,.$$

Therefore, (15) and (16) are satisfied with probability $1 - \delta$ if

$$n \geq \max\{36\eta^{-2}(d_1 + d_2), 2\log(4/\delta)\} = \max\{C_1^2\rho^{-2}\kappa r(d_1 + d_2), 2\log(4/\delta)\}\,,$$

which is implied by (10) in Theorem 1.

In the t-design case with $t \geq 2$, we obtain a similar result via the matrix Bernstein inequality [47, Theorem 1.6], summarized as Theorem 4 in the appendix. If $[\boldsymbol{a}; \boldsymbol{b}]$ satisfy (5) and (7), then we have

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}_i\boldsymbol{a}_i^* - \boldsymbol{I}_{d_1}\right\| > \eta\right\} \leq 2d_1\exp\left(\frac{-\eta^2 n}{4d_1}\right)$$

and

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{b}_i\boldsymbol{b}_i^* - \boldsymbol{I}_{d_2}\right\| > \eta\right\} \leq 2d_2\exp\left(\frac{-\eta^2 n}{4d_2}\right)\,.$$

Therefore, (15) and (16) are satisfied with probability $1 - \delta$ if

$$n \geq \frac{4(d_1 + d_2)}{\eta^2}\cdot\log\left(\frac{4(d_1 + d_2)}{\delta}\right) = \frac{4C_1^2\kappa r(d_1 + d_2)}{\rho^2}\cdot\log\left(\frac{4(d_1 + d_2)}{\delta}\right)\,,$$

which is implied by (11) in Theorem 2.

5   Proof of Proposition 3

For conciseness, we introduce the following shorthand notations. Let

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \ldots & \boldsymbol{a}_n \end{bmatrix}, \qquad\qquad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \ldots & \boldsymbol{b}_n \end{bmatrix},$$

and define

$$\eta \overset{\text{def}}{=} \frac{\rho}{C_1 \sqrt{r\kappa}}. \tag{25}$$

Then, (15) and (16) are equivalent to

$$\left\| \boldsymbol{I}_{d_1} - \frac{1}{n} \boldsymbol{A}\boldsymbol{A}^* \right\| \leq \eta \quad \text{and} \quad \left\| \boldsymbol{I}_{d_2} - \frac{1}{n} \boldsymbol{B}\boldsymbol{B}^* \right\| \leq \eta. \tag{26}$$

First, through the following lemma we establish a sufficient optimality condition needed to prove Proposition 3.

Lemma 3. *Let* $\boldsymbol{X}_0 \in \mathbb{C}^{n_1 \times r}$ *and* $\boldsymbol{Y}_0 \in \mathbb{C}^{n_2 \times r}$ *satisfy* $\boldsymbol{X}_0 \boldsymbol{Y}_0^* = \boldsymbol{M}_0$. *Then* $[\boldsymbol{X}_0; \ \boldsymbol{Y}_0]$ *is the unique maximizer of* (2), *if for every* $\boldsymbol{\Delta}_1 \in \mathbb{C}^{d_1 \times r}$ *and* $\boldsymbol{\Delta}_2 \in \mathbb{C}^{d_2 \times r}$ *we have*

$$\langle \widetilde{\boldsymbol{X}}_0 - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}_0, \boldsymbol{\Delta}_1 \rangle + \langle \widetilde{\boldsymbol{Y}}_0 - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y}_0, \boldsymbol{\Delta}_2 \rangle \leq \frac{1}{n}\sum_{i=1}^{n} |\langle \boldsymbol{a}_i \boldsymbol{b}_i^*, \boldsymbol{X}_0\boldsymbol{\Delta}_2^* + \boldsymbol{\Delta}_1 \boldsymbol{Y}_0^* \rangle| \tag{27}$$

with equality occurring only when both $\boldsymbol{\Delta}_1$ *and* $\boldsymbol{\Delta}_2$ *are zero.*

Proof. Let $\boldsymbol{X}_0 \in \mathbb{C}^{d_1 \times r}$ and $\boldsymbol{Y}_0 \in \mathbb{C}^{d_2 \times r}$ satisfy $\boldsymbol{X}_0 \boldsymbol{Y}_0^* = \boldsymbol{M}_0$. Note that $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ would be the *unique* maximizer of (2) if for any $\boldsymbol{\Delta}_1 \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{\Delta}_2 \in \mathbb{R}^{d_2 \times r}$

$$\langle \widetilde{\boldsymbol{X}}_0, \boldsymbol{\Delta}_1 \rangle + \langle \widetilde{\boldsymbol{Y}}_0, \boldsymbol{\Delta}_2 \rangle \leq \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{X}_0 + \boldsymbol{\Delta}_1, \boldsymbol{Y}_0 + \boldsymbol{\Delta}_2) - \ell_i(\boldsymbol{X}_0, \boldsymbol{Y}_0), \tag{28}$$

with equality holding only for $\boldsymbol{\Delta}_1 = \boldsymbol{0}$ and $\boldsymbol{\Delta}_2 = \boldsymbol{0}$.

Since $\boldsymbol{a}_i^* \boldsymbol{X}_0 \boldsymbol{Y}_0^* \boldsymbol{b}_i = m_i$, for each i, we obtain

$$\begin{aligned}
&\ell_i(\boldsymbol{X}_0 + \boldsymbol{\Delta}_1, \boldsymbol{Y}_0 + \boldsymbol{\Delta}_2) - \ell_i(\boldsymbol{X}_0, \boldsymbol{Y}_0) \\
&= \frac{1}{2}\|\boldsymbol{\Delta}_1^* \boldsymbol{a}_i\|^2 + \langle \boldsymbol{X}_0^* \boldsymbol{a}_i, \boldsymbol{\Delta}_1^* \boldsymbol{a}_i \rangle + \frac{1}{2}\|\boldsymbol{\Delta}_2^* \boldsymbol{b}_i\|^2 + \langle \boldsymbol{Y}_0^* \boldsymbol{b}_i, \boldsymbol{\Delta}_2^* \boldsymbol{b}_i \rangle \\
&\quad + |\boldsymbol{a}_i^*(\boldsymbol{X}_0\boldsymbol{\Delta}_2^* + \boldsymbol{\Delta}_1 \boldsymbol{Y}_0^* + \boldsymbol{\Delta}_1 \boldsymbol{\Delta}_2^*)\boldsymbol{b}_i| \\
&\geq \langle \boldsymbol{a}_i \boldsymbol{a}_i^* \boldsymbol{X}_0, \boldsymbol{\Delta}_1 \rangle + \langle \boldsymbol{b}_i \boldsymbol{b}_i^* \boldsymbol{Y}_0, \boldsymbol{\Delta}_2 \rangle + |\boldsymbol{a}_i^*(\boldsymbol{X}_0\boldsymbol{\Delta}_2^* + \boldsymbol{\Delta}_1 \boldsymbol{Y}_0^*)\boldsymbol{b}_i| \\
&\quad - |\boldsymbol{a}_i^* \boldsymbol{\Delta}_1 \boldsymbol{\Delta}_2^* \boldsymbol{b}_i| + \frac{1}{2}\|\boldsymbol{\Delta}_1^* \boldsymbol{a}_i\|^2 + \frac{1}{2}\|\boldsymbol{\Delta}_2^* \boldsymbol{b}_i\|^2 \\
&\geq \langle \boldsymbol{a}_i \boldsymbol{a}_i^* \boldsymbol{X}_0, \boldsymbol{\Delta}_1 \rangle + \langle \boldsymbol{b}_i \boldsymbol{b}_i^* \boldsymbol{Y}_0, \boldsymbol{\Delta}_2 \rangle + |\boldsymbol{a}_i^*(\boldsymbol{X}_0\boldsymbol{\Delta}_2^* + \boldsymbol{\Delta}_1 \boldsymbol{Y}_0^*)\boldsymbol{b}_i|,
\end{aligned} \tag{29}$$

where the first lower bound is obtained by the triangle inequality and the next lower bound follows from the Cauchy-Schwarz inequality.

Using (29), for $i = 1, \ldots, n$, the right-hand side of (27) can be bounded from above. This bound shows that if (27) holds with equality occurring only at $[\boldsymbol{\Delta}_1; \ \boldsymbol{\Delta}_2] = \boldsymbol{0}$, then (28) holds and the claim is proved. $\qquad\square$

For any $\boldsymbol{X}_0 \in \mathbb{C}^{d_1 \times r}$ and $\boldsymbol{Y}_0 \in \mathbb{C}^{d_2 \times r}$ that satisfy $\boldsymbol{X}_0 \boldsymbol{Y}_0^* = \boldsymbol{M}_0$, we have $\boldsymbol{X}_0 \boldsymbol{\Delta}_2^* + \boldsymbol{\Delta}_1 \boldsymbol{Y}_0^* \in \mathcal{T}$. Therefore, by Lemma 3 and (14), it suffices to show that

$$\langle \widetilde{\boldsymbol{X}}_0 - \frac{1}{n} \boldsymbol{A} \boldsymbol{A}^* \boldsymbol{X}_0, \boldsymbol{\Delta}_1 \rangle + \langle \overline{\widetilde{\boldsymbol{Y}}}_0 - \frac{1}{n} \overline{\boldsymbol{B}} \boldsymbol{B}^\mathsf{T} \overline{\boldsymbol{Y}}_0, \overline{\boldsymbol{\Delta}}_2 \rangle \leq \rho \| \boldsymbol{X}_0 \boldsymbol{\Delta}_2^* + \boldsymbol{\Delta}_1 \boldsymbol{Y}_0^* \|_\mathrm{F} \tag{30}$$

for all $\boldsymbol{\Delta}_1 \in \mathbb{C}^{d_1 \times r}$ and $\boldsymbol{\Delta}_2 \in \mathbb{C}^{d_2 \times r}$ with the equality only when $[\boldsymbol{\Delta}_1; \; \boldsymbol{\Delta}_2] = \boldsymbol{0}$.

Define the linear operator $\mathcal{L} : \mathbb{C}^{(d_1+d_2) \times r} \to \mathbb{C}^{d_1 \times d_2}$ by

$$\mathcal{L}\left(\left[\boldsymbol{\Delta}_1; \; \overline{\boldsymbol{\Delta}}_2 \right] \right) = \boldsymbol{X}_0 \boldsymbol{\Delta}_2^* + \boldsymbol{\Delta}_1 \boldsymbol{Y}_0^*, \quad \text{for all } \boldsymbol{\Delta}_1 \in \mathbb{C}^{d_1 \times r}, \boldsymbol{\Delta}_2 \in \mathbb{C}^{d_2 \times r},$$

whose adjoint operator is

$$\mathcal{L}^*(\boldsymbol{Z}) = \left[\boldsymbol{Z} \boldsymbol{Y}_0; \; \boldsymbol{Z}^\mathsf{T} \overline{\boldsymbol{X}}_0 \right], \quad \text{for all } \boldsymbol{Z} \in \mathbb{C}^{d_1 \times d_2}.$$

With

$$\boldsymbol{\Delta} = \left[\boldsymbol{\Delta}_1; \; \overline{\boldsymbol{\Delta}}_2 \right],$$

and

$$\boldsymbol{E} = \left[\widetilde{\boldsymbol{X}}_0 - \frac{1}{n} \boldsymbol{A} \boldsymbol{A}^* \boldsymbol{X}_0; \; \overline{\widetilde{\boldsymbol{Y}}}_0 - \frac{1}{n} \overline{\boldsymbol{B}} \boldsymbol{B}^\mathsf{T} \overline{\boldsymbol{Y}}_0 \right], \tag{31}$$

we can rewrite (30) as

$$\langle \boldsymbol{E}, \boldsymbol{\Delta} \rangle \leq \rho \| \mathcal{L}(\boldsymbol{\Delta}) \|_\mathrm{F}.$$

Note that $\mathcal{L}$ generally has a nontrivial nullspace, particularly, if $(d_1 + d_2)r < d_1 d_2$. Therefore, in view of the inequality above, it is necessary to have $\langle \boldsymbol{E}, \boldsymbol{\Delta} \rangle = 0$ for all $\boldsymbol{\Delta}$ in the nullspace of $\mathcal{L}$. Fortunately, for a certain choice of $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ the corresponding matrix $\boldsymbol{E}$ satisfies the required condition, as shown by the following lemma, which is proved in Appendix B.3.

Lemma 4. *Let $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ be the solution to*

$$\max_{\boldsymbol{X} \in \mathbb{C}^{d_1 \times r}, \boldsymbol{Y} \in \mathbb{C}^{d_2 \times r}} \langle \widetilde{\boldsymbol{X}}_0, \boldsymbol{X} \rangle + \langle \widetilde{\boldsymbol{Y}}_0, \boldsymbol{Y} \rangle - \frac{1}{2n} \| \boldsymbol{X}^* \boldsymbol{A} \|_\mathrm{F}^2 - \frac{1}{2n} \| \boldsymbol{Y}^* \boldsymbol{B} \|_\mathrm{F}^2 \tag{32}$$
$$\text{subject to } \boldsymbol{X} \boldsymbol{Y}^* = \boldsymbol{M}_0.$$

For the operator $\mathcal{L}$ and the matrix $\boldsymbol{E}$, defined by (31) in terms of the particular solution $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ above, we have

$$\boldsymbol{E} \in \mathbb{V} \overset{\text{def}}{=} \mathrm{range}(\mathcal{L}^*).$$

Furthermore, if (26) holds, then

$$\left\| \begin{bmatrix} \boldsymbol{X}_0 \\ \boldsymbol{Y}_0 \end{bmatrix} - \begin{bmatrix} \widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0 \end{bmatrix} \right\|_\mathrm{F}^2 \leq \frac{1+\eta}{1-\eta} \cdot \left\| \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} - \begin{bmatrix} \widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0 \end{bmatrix} \right\|_\mathrm{F}^2 + \frac{2\eta}{1-\eta} \cdot \left\| \begin{bmatrix} \boldsymbol{X}_0 \\ \boldsymbol{Y}_0 \end{bmatrix} - \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \right\|_\mathrm{F} \left\| \begin{bmatrix} \widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0 \end{bmatrix} \right\|_\mathrm{F} \tag{33}$$

for all $\boldsymbol{X} \in \mathbb{C}^{d_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{C}^{d_2 \times r}$ satisfying $\boldsymbol{X} \boldsymbol{Y}^* = \boldsymbol{M}_0$.

Hereafter, the pair $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ is chosen as in Lemma 4. The subspace $\mathbb{V}$ can be described explicitly as

$$\mathbb{V} = \mathrm{range}(\mathcal{L}^*) = \left\{ \left[\boldsymbol{Z} \boldsymbol{Y}_0; \; \boldsymbol{Z}^\mathsf{T} \overline{\boldsymbol{X}}_0 \right] : \boldsymbol{Z} \in \mathbb{C}^{d_1 \times d_2} \right\}. \tag{34}$$

By the fundamental theorem of linear algebra, we also have $\mathbb{V} = \text{null}(\mathcal{L})^{\perp}$. Thus, with $\mathcal{P}_{\mathbb{V}}$ denoting the orthogonal projection onto the subspace $\mathbb{V}$, Lemma 4 implies that $\boldsymbol{E} = \mathcal{P}_{\mathbb{V}}\boldsymbol{E}$. Consequently, to guarantee (30), it suffices to have

$$\|\boldsymbol{E}\|_{\mathrm{F}}\|\mathcal{P}_{\mathbb{V}}\boldsymbol{\Delta}\|_{\mathrm{F}} \leq \rho\|\mathcal{L}(\mathcal{P}_{\mathbb{V}}\boldsymbol{\Delta})\|_{\mathrm{F}}, \tag{35}$$

because by the Cauchy-Schwarz inequality

$$\langle \boldsymbol{E}, \boldsymbol{\Delta} \rangle = \langle \mathcal{P}_{\mathbb{V}}\boldsymbol{E}, \boldsymbol{\Delta} \rangle = \langle \boldsymbol{E}, \mathcal{P}_{\mathbb{V}}\boldsymbol{\Delta} \rangle \leq \|\boldsymbol{E}\|_{\mathrm{F}}\|\mathcal{P}_{\mathbb{V}}\boldsymbol{\Delta}\|_{\mathrm{F}}.$$

Furthermore, the following technical lemma provides a lower bound for $\|\mathcal{L}(\mathcal{P}_{\mathbb{V}}\boldsymbol{\Delta})\|_{\mathrm{F}}/\|\mathcal{P}_{\mathbb{V}}\boldsymbol{\Delta}\|_{\mathrm{F}}$.

Lemma 5. *The linear operator $\mathcal{L}$ satisfies*

$$\|\mathcal{L}(P_{\mathbb{V}}\boldsymbol{\Delta})\|_{\mathrm{F}} \geq \min\{\sigma_{\min}(\boldsymbol{X}_0), \sigma_{\min}(\boldsymbol{Y}_0)\}\|\mathcal{P}_{\mathbb{V}}\boldsymbol{\Delta}\|_{\mathrm{F}}, \quad \forall \boldsymbol{\Delta} \in \mathbb{C}^{(d_1+d_2)\times r}. \tag{36}$$

Proof. Let $[\boldsymbol{\Delta}_1; \overline{\boldsymbol{\Delta}}_2]$ belong to $\mathbb{V} = \text{range}(\mathcal{L}^*)$. Then there exists $\boldsymbol{Z} \in \mathbb{C}^{d_1 \times d_2}$ such that $\boldsymbol{\Delta}_1 = \boldsymbol{Z}\boldsymbol{Y}_0$ and $\boldsymbol{\Delta}_2 = \boldsymbol{Z}^*\boldsymbol{X}_0$. Thus

$$\begin{aligned}
\left\|\mathcal{L}([\boldsymbol{\Delta}_1; \overline{\boldsymbol{\Delta}}_2])\right\|_{\mathrm{F}}^2 &= \|\boldsymbol{X}_0\boldsymbol{\Delta}_2^* + \boldsymbol{\Delta}_1\boldsymbol{Y}_0^*\|_{\mathrm{F}}^2 \\
&= \|\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{Y}_0\boldsymbol{Y}_0^*\|_{\mathrm{F}}^2 \\
&= \|\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{Z}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\boldsymbol{Y}_0\boldsymbol{Y}_0^*\|_{\mathrm{F}}^2 + 2\langle \boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{Z}, \boldsymbol{Z}\boldsymbol{Y}_0\boldsymbol{Y}_0^* \rangle \\
&= \|\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{Z}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\boldsymbol{Y}_0\boldsymbol{Y}_0^*\|_{\mathrm{F}}^2 + 2\|\boldsymbol{X}_0^*\boldsymbol{Z}\boldsymbol{Y}_0\|_{\mathrm{F}}^2 \\
&\geq \sigma_{\min}^2(\boldsymbol{X}_0)\|\boldsymbol{X}_0^*\boldsymbol{Z}\|_{\mathrm{F}}^2 + \sigma_{\min}^2(\boldsymbol{Y}_0)\|\boldsymbol{Z}\boldsymbol{Y}_0\|_{\mathrm{F}}^2 \\
&\geq \min\{\sigma_{\min}^2(\boldsymbol{X}_0), \sigma_{\min}^2(\boldsymbol{Y}_0)\}\left\|[\boldsymbol{\Delta}_1; \overline{\boldsymbol{\Delta}}_2]\right\|_{\mathrm{F}}^2.
\end{aligned}$$

$\square$

Note that

$$\max_{\boldsymbol{X}\in\mathbb{C}^{d_1\times r}, \boldsymbol{Y}\in\mathbb{C}^{d_2\times r}} \min\{\sigma_{\min}(\boldsymbol{X}), \sigma_{\min}(\boldsymbol{Y})\} = \sqrt{\sigma_r(\boldsymbol{M}_0)}$$
$$\text{subject to } \boldsymbol{X}\boldsymbol{Y}^* = \boldsymbol{M}_0.$$

Indeed, the assumptions of the proposition implies that $\min\{\sigma_{\min}(\boldsymbol{X}_0), \sigma_{\min}(\boldsymbol{Y}_0)\}$ is larger than $\sqrt{\sigma_r(\boldsymbol{M}_0)}$ divided by a numerical constant. In order to show this, we introduce another pair $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$ with $\boldsymbol{X}_1\boldsymbol{Y}_1^* = \boldsymbol{M}_0$ so that $[\boldsymbol{X}_1; \boldsymbol{Y}_1]$ approximates $[\widetilde{\boldsymbol{X}}_0; \widetilde{\boldsymbol{Y}}_0]$. The following lemma provides an upper bound on the approximation error; the proof is provided in Appendix C.1.

Lemma 6. *Suppose that the rank-r matrices $\boldsymbol{M}_0$ and $\widetilde{\boldsymbol{M}}_0$, whose compact SVDs are respectively $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^*$ and $\widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{\Sigma}}_0\widetilde{\boldsymbol{V}}_0^*$, satisfy $\|\widetilde{\boldsymbol{M}}_0 - \boldsymbol{M}_0\| < \sigma_r(\boldsymbol{M}_0)$. Then*

$$\min_{\boldsymbol{Q}\in\mathbb{C}^{r\times r}: \boldsymbol{Q}^{-1}=\boldsymbol{Q}^*} \left\|\begin{bmatrix}\boldsymbol{U}_0\\\boldsymbol{V}_0\end{bmatrix}\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{Q} - \begin{bmatrix}\widetilde{\boldsymbol{U}}_0\\\widetilde{\boldsymbol{V}}_0\end{bmatrix}\widetilde{\boldsymbol{\Sigma}}_0^{1/2}\right\|_{\mathrm{F}} \leq \frac{8\sqrt{2}\,\sigma_1(\boldsymbol{\Sigma}_0)}{\sqrt{\sigma_r(\boldsymbol{\Sigma}_0)}} \cdot \frac{\|\widetilde{\boldsymbol{M}}_0 - \boldsymbol{M}_0\|_{\mathrm{F}}}{\sigma_r(\boldsymbol{\Sigma}_0) - \|\widetilde{\boldsymbol{M}}_0 - \boldsymbol{M}_0\|}.$$

Let $\widetilde{\boldsymbol{U}}_0$, $\widetilde{\boldsymbol{\Sigma}}_0$ and $\widetilde{\boldsymbol{V}}_0$ be as in Lemma 6. Let $\boldsymbol{Q}$ be the minimizer in Lemma 6. Let

$$\boldsymbol{X}_1 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{Q} \quad \text{and} \quad \boldsymbol{Y}_1 = \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{Q}.$$

15

Then (17) implies

$$\frac{\left\|[\boldsymbol{X}_1;\ \boldsymbol{Y}_1] - [\widetilde{\boldsymbol{X}}_0;\ \widetilde{\boldsymbol{Y}}_0]\right\|_{\mathrm{F}}}{\sqrt{\sigma_r(\boldsymbol{M}_0)}} \leq \frac{8\sqrt{2}\rho}{C_2 - \rho/\sqrt{r}\kappa} \leq \frac{8\sqrt{2}\rho}{C_2 - 1}. \tag{37}$$

Choosing $C_2 \geq 8\sqrt{2} + 1$ yields

$$\left\|[\boldsymbol{X}_1;\ \boldsymbol{Y}_1] - [\widetilde{\boldsymbol{X}}_0;\ \widetilde{\boldsymbol{Y}}_0]\right\|_{\mathrm{F}} \leq \rho\sqrt{\sigma_r(\boldsymbol{M}_0)}. \tag{38}$$

It follows from (38) via the triangle inequality that

$$\left\|\begin{bmatrix}\widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0\end{bmatrix}\right\|_{\mathrm{F}} \leq \left\|\begin{bmatrix}\boldsymbol{X}_1 \\ \boldsymbol{Y}_1\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0\end{bmatrix}\right\|_{\mathrm{F}} + \left\|\begin{bmatrix}\boldsymbol{X}_1 \\ \boldsymbol{Y}_1\end{bmatrix}\right\|_{\mathrm{F}} \leq \rho\sqrt{\sigma_r(\boldsymbol{M}_0)} + \sqrt{r\sigma_1(\boldsymbol{M}_0)} \leq 2\sqrt{r\sigma_1(\boldsymbol{M}_0)}. \tag{39}$$

Plugging in (39) to (33) with $\boldsymbol{X} = \boldsymbol{X}_1$ and $\boldsymbol{Y} = \boldsymbol{Y}_1$ gives

$$\left\|\begin{bmatrix}\boldsymbol{X}_0 \\ \boldsymbol{Y}_0\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0\end{bmatrix}\right\|_{\mathrm{F}}^2 \leq \frac{1+\eta}{1-\eta} \cdot \left\|\begin{bmatrix}\boldsymbol{X}_1 \\ \boldsymbol{Y}_1\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0\end{bmatrix}\right\|_{\mathrm{F}}^2$$
$$+ \frac{4\eta\sqrt{r\sigma_1(\boldsymbol{M}_0)}}{1-\eta}\left(\left\|\begin{bmatrix}\boldsymbol{X}_0 \\ \boldsymbol{Y}_0\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0\end{bmatrix}\right\|_{\mathrm{F}} + \left\|\begin{bmatrix}\boldsymbol{X}_1 \\ \boldsymbol{Y}_1\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0\end{bmatrix}\right\|_{\mathrm{F}}\right).$$

After some simplification, the above inequality and (37) imply

$$\frac{\left\|[\boldsymbol{X}_0;\ \boldsymbol{Y}_0] - [\widetilde{\boldsymbol{X}}_0;\ \widetilde{\boldsymbol{Y}}_0]\right\|_{\mathrm{F}}}{\rho\sqrt{\sigma_r(\boldsymbol{M}_0)}} \leq \frac{1}{C_1-1}\left(2 + \sqrt{\frac{128(C_1+1)}{(C_1-1)(C_2-1)^2}} + \frac{32\sqrt{2}}{(C_1-1)(C_2-1)} + \frac{4}{(C_1-1)^2}\right).$$

By choosing both C_1 and C_2 large enough, we obtain

$$\left\|\begin{bmatrix}\boldsymbol{X}_1 \\ \boldsymbol{Y}_1\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0\end{bmatrix}\right\|_{\mathrm{F}} \leq \frac{\rho\sqrt{\sigma_r(\boldsymbol{M}_0)}}{10} \quad \text{and} \quad \left\|\begin{bmatrix}\boldsymbol{X}_0 \\ \boldsymbol{Y}_0\end{bmatrix} - \begin{bmatrix}\widetilde{\boldsymbol{X}}_0 \\ \widetilde{\boldsymbol{Y}}_0\end{bmatrix}\right\|_{\mathrm{F}} \leq \frac{\rho\sqrt{\sigma_r(\boldsymbol{M}_0)}}{10}.$$

Then by the triangle inequality we have

$$\left\|\begin{bmatrix}\boldsymbol{X}_0 \\ \boldsymbol{Y}_0\end{bmatrix} - \begin{bmatrix}\boldsymbol{X}_1 \\ \boldsymbol{Y}_1\end{bmatrix}\right\|_{\mathrm{F}} \leq \frac{\rho\sqrt{\sigma_r(\boldsymbol{M}_0)}}{5}.$$

By the singular value perturbation theory by Weyl [6] and because $\rho \in (0,1)$, it follows that

$$\sigma_r(\boldsymbol{X}_0) \geq \sigma_r(\boldsymbol{X}_1) - \|\boldsymbol{X}_0 - \boldsymbol{X}_1\| \geq \frac{4\sqrt{\sigma_r(\boldsymbol{M}_0)}}{5}$$

and

$$\sigma_r(\boldsymbol{Y}_0) \geq \sigma_r(\boldsymbol{Y}_1) - \|\boldsymbol{Y}_0 - \boldsymbol{Y}_1\| \geq \frac{4\sqrt{\sigma_r(\boldsymbol{M}_0)}}{5}.$$

Then (36) is implied by

$$\|\mathcal{L}(P_{\mathbb{V}}\boldsymbol{\Delta})\|_2 \geq \frac{\sqrt{\sigma_r(\boldsymbol{M}_0)}}{2} \cdot \|\mathcal{P}_{\mathbb{V}}\boldsymbol{\Delta}\|_{\mathrm{F}}.$$

Therefore, by combinig the above estimates, we obtain a sufficient condition for (35) given by

$$\|\boldsymbol{E}\|_{\mathrm{F}} \leq \frac{4\rho\sqrt{\sigma_r(\boldsymbol{M}_0)}}{5}. \tag{40}$$

The following lemma, which is proved in Appendix C.2, provides an upper bound for $\|\boldsymbol{E}\|_{\mathrm{F}}$ that can be used to ensure the sufficient condition (40).

Lemma 7. *Let $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ and $\boldsymbol{E}$ be as in Lemma 4. Suppose that (26) holds. Then for all $(\boldsymbol{X}, \boldsymbol{Y})$ satisfying $\boldsymbol{X}\boldsymbol{Y}^* = \boldsymbol{M}_0$, we have*

$$\|\boldsymbol{E}\|_{\mathrm{F}} \leq \frac{3+\eta}{1-\eta} \cdot \left(\left\| [\widetilde{\boldsymbol{X}}_0 - \boldsymbol{X}; \ \widetilde{\boldsymbol{Y}}_0 - \boldsymbol{Y}] \right\|_{\mathrm{F}} + \eta \|[\boldsymbol{X}; \ \boldsymbol{Y}]\|_{\mathrm{F}} \right).$$

By applying Lemma 7 with $\boldsymbol{X} = \boldsymbol{X}_1$ and $\boldsymbol{Y} = \boldsymbol{Y}_1$, and by (25), we obtain

$$\|\boldsymbol{E}\|_{\mathrm{F}} \leq \frac{3C_1 + 1}{C_1 - 1} \left(\frac{1}{10} + \frac{1}{C_1} \right) \rho \sqrt{\sigma_r(\boldsymbol{M}_0)} \,.$$

Finally we choose C_1 large enough so that

$$\frac{3C_1 + 1}{C_1 - 1} \left(\frac{1}{10} + \frac{1}{C_1} \right) \leq \frac{4}{5} \,.$$

This completes the proof.

6 Analysis of Spectral Initialization

Note that

$$\mathcal{A}^* \mathcal{A}(\boldsymbol{M}_0) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{a}_i^* \boldsymbol{M}_0 \boldsymbol{b}_i \boldsymbol{b}_i^* \,.$$

Then by the independence between $\boldsymbol{a}$ and $\boldsymbol{b}$ together with the isotropy of each of them implies

$$\mathbb{E} \mathcal{A}^* \mathcal{A}(\boldsymbol{M}_0) = \boldsymbol{M}_0 \,.$$

Since the spectral initialization computes $\widetilde{\boldsymbol{M}}_0$ as the best rank-r approximation of $\mathcal{A}^* \mathcal{A}(\boldsymbol{M}_0)$, by the optimality, we have

$$\|\widetilde{\boldsymbol{M}}_0 - \boldsymbol{M}_0\| \leq 2\|(\mathcal{A}^* \mathcal{A} - \mathrm{Id})\boldsymbol{M}_0\| \,.$$

Our goal is to show that

$$\|(\mathcal{A}^* \mathcal{A} - \mathrm{Id})\boldsymbol{M}_0\| \leq \frac{C\|\boldsymbol{M}_0\|}{\sqrt{r}\kappa^2}$$

for an absolute constant C. Then it will imply (9).

Let

$$\boldsymbol{Z}_i = \frac{1}{n} \boldsymbol{a}_i \boldsymbol{a}_i^* \boldsymbol{M}_0 \boldsymbol{b}_i \boldsymbol{b}_i^* \,, \quad i = 1, \ldots, n \,.$$

Then we will show that

$$\left\| \sum_{i=1}^{n} \boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i \right\| \leq \frac{C\|\boldsymbol{M}_0\|}{\sqrt{r}\kappa^2}$$

holds with high probability respectively in the cases of Gaussian and t-design measurements.

In the following derivations we use the shorthand

$$\xi \stackrel{\mathrm{def}}{=} \left\| \sum_{i=1}^{n} \boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i \right\|_{S_p} \tag{41}$$

for compact notation.

6.1 Proof of Proposition 1

By the triangle inequality we have

$$n\|\boldsymbol{Z}_i\|_{S_p} = \left\|\sum_{k=1}^{r} \sigma_k \boldsymbol{a}_i \boldsymbol{a}_i^* \boldsymbol{u}_k \boldsymbol{v}_k^* \boldsymbol{b}_i \boldsymbol{b}_i^*\right\|_{S_p} \leq \sum_{k=1}^{r} \sigma_k \|\boldsymbol{a}_i\|_2 |\boldsymbol{a}_i^* \boldsymbol{u}_k| |\boldsymbol{v}_k^* \boldsymbol{b}_i| \|\boldsymbol{b}_i^*\|_2.$$

It follows that

$$
\begin{aligned}
\left(\mathbb{E}\|\boldsymbol{Z}_i\|_{S_p}^p\right)^{1/p} &\leq \frac{1}{n} \sum_{k=1}^{r} \sigma_k \left(\mathbb{E}\|\boldsymbol{a}_i\|_2^p |\boldsymbol{a}_i^* \boldsymbol{u}_k|^p |\boldsymbol{v}_k^* \boldsymbol{b}_i|^p \|\boldsymbol{b}_i^*\|_2^p\right)^{1/p} \\
&\overset{(a)}{\leq} \frac{1}{n} \sum_{k=1}^{r} \sigma_k \left(\mathbb{E}\|\boldsymbol{a}_i\|_2^p |\boldsymbol{a}_i^* \boldsymbol{u}_k|^p\right)^{1/p} \left(\mathbb{E}|\boldsymbol{v}_k^* \boldsymbol{b}_i|^p \|\boldsymbol{b}_i^*\|_2^p\right)^{1/p} \\
&\overset{(b)}{\leq} \frac{1}{n} \sum_{k=1}^{r} \sigma_k \left(\mathbb{E}\|\boldsymbol{a}_i\|_2^{2p}\right)^{1/2p} \left(\mathbb{E}|\boldsymbol{a}_i^* \boldsymbol{u}_k|^{2p}\right)^{1/2p} \left(\mathbb{E}|\boldsymbol{v}_k^* \boldsymbol{b}_i|^{2p}\right)^{1/2p} \left(\mathbb{E}\|\boldsymbol{b}_i^*\|_2^{2p}\right)^{1/2p} \\
&\overset{(c)}{\lesssim} \frac{p^2 \sqrt{d_1 d_2}}{n} \sum_{k=1}^{r} \sigma_k \\
&\lesssim \frac{\|\boldsymbol{M}_0\| p^2 r(d_1 + d_2)}{n},
\end{aligned}
$$

where (a) holds by the independence between $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$; (b) follows from Cauchy-Schwarz inequality; and (c) holds by [50, Eq. (2.15) and Lemma 2.7.6] since $\|\boldsymbol{a}_i\|_2^2$, $|\boldsymbol{a}_i^* \boldsymbol{u}_k|^2$, $|\boldsymbol{v}_k^* \boldsymbol{b}_i|^2$, and $\|\boldsymbol{b}_i\|_2^2$ are subexponential random variables. Then we deduce that

$$\left(\mathbb{E}\|\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i\|_{S_p}^p\right)^{1/p} \leq 2 \left(\mathbb{E}\|\boldsymbol{Z}_i\|_{S_p}^p\right)^{1/p} \lesssim \frac{\|\boldsymbol{M}_0\| p^2 r(d_1 + d_2)}{n}.$$

Furthermore, the second-order moments are computed as

$$\mathbb{E}(\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i)(\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i)^* = \frac{d_2 + 2}{n^2} \|\boldsymbol{M}_0\|_{\mathrm{F}}^2 \boldsymbol{I}_{d_1} + \frac{2d_2 + 3}{n^2} \boldsymbol{M}_0 \boldsymbol{M}_0^*$$

and

$$\mathbb{E}(\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i)^*(\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i) = \frac{d_1 + 2}{n^2} \|\boldsymbol{M}_0\|_{\mathrm{F}}^2 \boldsymbol{I}_{d_2} + \frac{2d_1 + 3}{n^2} \boldsymbol{M}_0^* \boldsymbol{M}_0.$$

Then we obtain

$$\left(\left\|\sum_{i=1}^{n} \mathbb{E}(\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i)(\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i)^*\right\|_{S_p} \vee \left\|\sum_{i=1}^{n} \mathbb{E}(\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i)^*(\boldsymbol{Z}_i - \mathbb{E}\boldsymbol{Z}_i)\right\|_{S_p}\right) \lesssim \frac{\|\boldsymbol{M}_0\|^2 r(d_1 + d_2)^{1+1/p}}{n}.$$

Now we are ready to apply Theorem 5. Recalling (41), Theorem 5 provides

$$\frac{(\mathbb{E}\xi^p)^{1/p}}{\|\boldsymbol{M}_0\|} \lesssim \left(\frac{pr(d_1 + d_2)^{1+1/p}}{n}\right)^{1/2} \vee \frac{p^3 n^{1/p} r(d_1 + d_2)}{n}.$$

By Markov's inequality, it implies

$$\mathbb{P}\{\xi \geq (\mathbb{E}\xi^p)^{1/p} \eta^{-1/p}\} \leq \eta. \tag{42}$$

We choose $\eta = (d_1 + d_2)^{-\alpha}$. Then there exists an absolute constant C_1 such that if

$$n \geq C_1 \left[\epsilon^{-2} prd^{1+1/p+2\alpha/p} \vee \left(\epsilon^{-1} p^3 rd^{1+\alpha/p} \right)^{p/(p-1)} \right] , \tag{43}$$

then

$$\mathbb{P}\{\xi \geq \epsilon \|\boldsymbol{M}_0\|\} \leq (d_1 + d_2)^{-\alpha} . \tag{44}$$

Since the Shatten-p norm is larger than the spectral norm, it follows that (44) implies

$$\mathbb{P}\left\{ \|\mathcal{A}^* \mathcal{A} \boldsymbol{M}_0 - \mathbb{E}\mathcal{A}^* \mathcal{A} \boldsymbol{M}_0\| \geq \epsilon \|\boldsymbol{M}_0\| \right\} \leq (d_1 + d_2)^{-\alpha} .$$

We choose $p = (2\alpha + 1)\ln(d_1 + d_2)$ so that the condition in (43) simplifies to

$$n \geq C_2 r(d_1 + d_2) \left(\epsilon^{-2} \alpha \ln^2(d_1 + d_2) \vee \epsilon^{-1} \alpha^3 \ln^3(d_1 + d_2) \right)$$

for another an absolute constant C_2. We obtain a sufficient condition by choosing $\epsilon = C/\kappa^2 \sqrt{r}$. This compltes the proof.

6.2 Proof of Proposition 2

Since $\|\boldsymbol{a}_i\|_2 = \sqrt{d_1}$ and $\|\boldsymbol{b}_i\|_2 = \sqrt{d_2}$, we have

$$\|\boldsymbol{Z}_i\|_{S_{2t}} = \frac{\sqrt{d_1 d_2}}{n} \left| \sum_{k=1}^r \sigma_k \boldsymbol{a}_i^* \boldsymbol{u}_k \boldsymbol{v}_k^* \boldsymbol{b}_i \right| .$$

Note that we have

$$\left| \sum_{k=1}^r \sigma_k \boldsymbol{a}_i^* \boldsymbol{u}_k \boldsymbol{v}_k^* \boldsymbol{b}_i \right| \leq \sum_{k=1}^r \sigma_k \left| \boldsymbol{a}_i^* \boldsymbol{u}_k \boldsymbol{v}_k^* \boldsymbol{b}_i \right| .$$

Thus, by the triangle inequality in L_{2t} together with the homogeneity of the L_{2p} norm, we obtain

$$\begin{aligned}
\left(\mathbb{E}\|\boldsymbol{Z}_i\|_{S_{2t}}^{2t} \right)^{1/2t} &\leq \frac{\sqrt{d_1 d_2}}{n} \sum_{k=1}^r \sigma_k \left(\mathbb{E} \left| \boldsymbol{a}_i^* \boldsymbol{u}_k \boldsymbol{v}_k^* \boldsymbol{b}_i \right|^{2t} \right)^{1/2t} \\
&\leq \frac{d_1 d_2}{n} \sum_{k=1}^r \sigma_k \left(\mathbb{E} \left| d_1^{-1/2} \boldsymbol{a}_i^* \boldsymbol{u}_k \right|^{2t} \right)^{1/2t} \left(\mathbb{E} \left| d_2^{-1/2} \boldsymbol{b}_i^* \boldsymbol{v}_k \right|^{2t} \right)^{1/2t} ,
\end{aligned} \tag{45}$$

where the second inequality holds since $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are independent.

Recall that, by the construction of a t-design set, $\boldsymbol{a}_i$ satisfies

$$\mathbb{E} \left(d_1^{-1} \boldsymbol{a}_i \boldsymbol{a}_i^* \right)^{\otimes t} = \binom{d+t-1}{t}^{-1} \mathcal{P}_{\mathrm{Sym}^t} .$$

Therefore we have

$$\begin{aligned}
\mathbb{E} \left| d_1^{-1/2} \boldsymbol{a}_i^* \boldsymbol{u}_k \right|^{2t} &= \mathrm{tr} \left(\mathbb{E} \left(d_1^{-1} \boldsymbol{a}_i \boldsymbol{a}_i^* \right)^{\otimes t} \left(\boldsymbol{u}_k \boldsymbol{u}_k^* \right)^{\otimes t} \right) \\
&= \binom{d_1+t-1}{t}^{-1} \mathrm{tr} \left(\mathcal{P}_{\mathrm{Sym}^t} (\boldsymbol{u}_k \boldsymbol{u}_k^*)^{\otimes t} \right) \\
&= \binom{d_1+t-1}{t}^{-1} \mathrm{tr} \left((\boldsymbol{u}_k \boldsymbol{u}_k^*)^{\otimes t} \right) \leq d_1^{-t} t! ,
\end{aligned}$$

19

where the third identity follows since $(\boldsymbol{u}_k \boldsymbol{u}_k^*)^{\otimes t}$ is invariant under the factor-wise permutation. Then we obtain

$$\left(\mathbb{E} \left| d_1^{-1/2} \boldsymbol{a}_i^* \boldsymbol{u}_k \right|^{2t} \right)^{1/2t} \le \left(d_1^{-t} t! \right)^{1/2t} \le \sqrt{\frac{t}{d_1}} \, . \tag{46}$$

Similarly we also have

$$\left(\mathbb{E} \left| d_2^{-1/2} \boldsymbol{b}_i^* \boldsymbol{v}_k \right|^{2t} \right)^{1/2t} \le \left(d_2^{-t} t! \right)^{1/2t} \le \sqrt{\frac{t}{d_2}} \, . \tag{47}$$

Then by plugging in the upper bounds in (46) and (47) into (45), we obtain

$$\left(\mathbb{E} \| \boldsymbol{Z}_i \|^{2t} \right)^{1/2t} \le \frac{t \sigma_1 r \sqrt{d_1 d_2}}{n} \le \frac{t \sigma_1 r (d_1 + d_2)}{2n} \, .$$

Furthermore, the triangle inequality and Jensen's inequality yield

$$\begin{aligned}
\left(\mathbb{E} \| \boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i \|_{S_{2t}}^{2t} \right)^{1/2t} &\le \left(\mathbb{E} \| \boldsymbol{Z}_i \|_{S_{2t}}^{2t} \right)^{1/2t} + \left(\mathbb{E} \| \mathbb{E} \boldsymbol{Z}_i \|_{S_{2t}}^{2t} \right)^{1/2t} \\
&= \left(\mathbb{E} \| \boldsymbol{Z}_i \|_{S_{2t}}^{2t} \right)^{1/2t} + \| \mathbb{E} \boldsymbol{Z}_i \|_{S_{2t}} \\
&\le 2 \left(\mathbb{E} \| \boldsymbol{Z}_i \|_{S_{2t}}^{2t} \right)^{1/2t} \, .
\end{aligned}$$

Putting these bounds together we obtain

$$\left(\mathbb{E} \| \boldsymbol{Z}_i \|^{2t} \right)^{1/2t} \le \frac{2 t \sigma_1 r (d_1 + d_2)}{2n} \, .$$

Furthermore, the second moments are computed as

$$\mathbb{E} (\boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i)(\boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i)^* = \frac{1}{n^2} \left(\frac{d_1 d_2 \| \boldsymbol{M}_0 \|_{\mathrm{F}}^2 \boldsymbol{I}_{d_1}}{d_1 + 1} + \left(\frac{d_1 d_2}{d_1 + 1} - 1 \right) \boldsymbol{M}_0 \boldsymbol{M}_0^* \right)$$

and

$$\mathbb{E} (\boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i)^* (\boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i) = \frac{1}{n^2} \left(\frac{d_1 d_2 \| \boldsymbol{M}_0 \|_{\mathrm{F}}^2 \boldsymbol{I}_{d_2}}{d_2 + 1} + \left(\frac{d_1 d_2}{d_2 + 1} - 1 \right) \boldsymbol{M}_0^* \boldsymbol{M}_0 \right) \, .$$

Then we have

$$\left(\left\| \sum_{i=1}^{n} \mathbb{E} (\boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i)(\boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i)^* \right\|_{S_{2t}} \vee \left\| \sum_{i=1}^{n} \mathbb{E} (\boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i)^* (\boldsymbol{Z}_i - \mathbb{E} \boldsymbol{Z}_i) \right\|_{S_{2t}} \right) \le \frac{4 \sigma_1^2 r (d_1 + d_2)^{1+1/2t}}{n} \, .$$

With ξ defined in (41), Theorem 5 provides

$$\left(\mathbb{E} \xi^{2t} \right)^{1/2t} \lesssim \left(\frac{t \sigma_1^2 r (d_1 + d_2)^{1+1/2t}}{n} \right)^{1/2} \vee \frac{t^2 n^{1/2t} \sigma_1 r (d_1 + d_2)}{n} \, .$$

Then invoking (42) with $\eta = (d_1 + d_2)^{-\alpha}$, we can show that there exists an absolute constant C_1 such that if

$$n \ge C_1 \left[\epsilon^{-2} tr (d_1 + d_2)^{1+1/2t+\alpha/t} \vee \left(\epsilon^{-1} t^2 r (d_1 + d_2)^{1+\alpha/2t} \right)^{2t/(2t-1)} \right] \, ,$$

then

$$\mathbb{P} \left\{ \| \mathcal{A}^* \mathcal{A} \boldsymbol{M}_0 - \mathbb{E} \mathcal{A}^* \mathcal{A} \boldsymbol{M}_0 \| \ge \epsilon \| \boldsymbol{M}_0 \| \right\} \le d^{-\alpha} \, .$$

The desired result follows by choosing $\epsilon = C / \kappa^2 \sqrt{r}$.

A Tools from Random Matrix Theory

Theorem 3 ([17, Theorem II.13]). *Let $\boldsymbol{G} \in \mathbb{R}^{m \times n}$ be a random matrix whose entries are independent copies of $g \sim \mathcal{N}(0,1)$. Then*

$$\mathbb{P}\left\{\left\|\frac{1}{m}\boldsymbol{G}^\mathsf{T}\boldsymbol{G} - \boldsymbol{I}_n\right\| > 3\max\left(\sqrt{\frac{4n}{m}}, \frac{4n}{m}\right)\right\} \le 2\exp(-n/2).$$

Theorem 4 (Matrix Bernstein inequality [47, Theorem 1.6]). *Let $(\boldsymbol{Y}_k) \subset \mathbb{C}^{m \times n}$ be a finite sequence of independent zero-mean random matrices such that $\|\boldsymbol{Y}_k\| \le R$ almost surely for all k. Let*

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

Then for all $t > 0$

$$\mathbb{P}\left\{\left\|\sum_k \boldsymbol{Y}_k\right\| \ge t\right\} \le (m+n)\cdot\exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Theorem 5 (Noncommutative Rosenthal inequality [29, Theorem 0.4], [20, Theorem 3.8]). *Let $(\boldsymbol{Y}_k) \subset \mathbb{C}^{d_1 \times d_2}$ be a finite sequence of independent zero-mean random matrices. Then there exists an absolute constant $C > 0$ such that for all $2 \le p < \infty$*

$$\left(\mathbb{E}\left\|\sum_k \boldsymbol{Y}_k\right\|_{S_p}^p\right)^{1/p}$$

$$\le C\left[\sqrt{p}\left(\left\|\sum_k \mathbb{E}\boldsymbol{Y}_k\boldsymbol{Y}_k^*\right\|_{S_p} \vee \left\|\sum_k \mathbb{E}\boldsymbol{Y}_k^*\boldsymbol{Y}_k\right\|_{S_p}\right)^{1/2} \vee p\left(\mathbb{E}\sum_k \|\boldsymbol{Y}_k\|_{S_p}^p\right)^{1/p}\right].$$

B Proofs of the main lemmas

B.1 Proof of Lemma 1

Let $\tau > 0$ be fixed. Since $\boldsymbol{a}$ and $\boldsymbol{b}$ are independent and isotropic, they satisfy

$$\mathbb{E}\boldsymbol{a}\boldsymbol{a}^* \otimes \boldsymbol{b}\boldsymbol{b}^* = \boldsymbol{I}_{d_1 d_2}, \tag{48}$$

which implies

$$\mathbb{E}\,|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^2 = \|\boldsymbol{H}\|_\mathrm{F}^2, \quad \forall \boldsymbol{H} \in \mathbb{R}^{d_1 \times d_2}. \tag{49}$$

Therefore, the Paley-Zygmund inequality [40] (see also [18, Corollary 3.3.2]), we have

$$\mathbb{P}\left(|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}| \ge \tau\|\boldsymbol{H}\|_\mathrm{F}\right) = \mathbb{P}\left(|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^2 \ge \tau^2\mathbb{E}\,|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^2\right)$$

$$\ge \frac{(1-\tau^2)^2\left(\mathbb{E}\,|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^2\right)^2}{\mathbb{E}|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^4}. \tag{50}$$

Then it suffices to show that the fourth order moment $\mathbb{E}|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^4$ is upper-bounded by $\left(\mathbb{E}\,|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^2\right)^2 = \|\boldsymbol{H}\|_\mathrm{F}^4$ within a constant factor. We first show this for the real Gaussian case. Since $[\boldsymbol{a};\ \boldsymbol{b}] \sim$

$\mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_1+d_2})$, we have

$$
\begin{aligned}
\mathbb{E}|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^4 &= \mathbb{E}_{\boldsymbol{b}}(\boldsymbol{b}^*\boldsymbol{H}^* \otimes \boldsymbol{b}^*\boldsymbol{H}^*)\mathbb{E}_{\boldsymbol{a}}(\boldsymbol{a}\boldsymbol{a}^* \otimes \boldsymbol{a}\boldsymbol{a}^*)(\boldsymbol{H}\boldsymbol{b} \otimes \boldsymbol{H}\boldsymbol{b}) \\
&= \mathbb{E}_{\boldsymbol{b}}\|\boldsymbol{b}^*\boldsymbol{H}^*\|_{\mathrm{F}}^4 + 2\,\|\boldsymbol{b}^*\boldsymbol{H}^*\|_{S_4}^4 \\
&= 3\,\mathbb{E}_{\boldsymbol{b}}\|\boldsymbol{b}^*\boldsymbol{H}^*\|_2^4 \\
&= 3\,\mathrm{tr}\left[(\boldsymbol{H} \otimes \boldsymbol{H})\mathbb{E}_{\boldsymbol{b}}(\boldsymbol{b}\boldsymbol{b}^* \otimes \boldsymbol{b}\boldsymbol{b}^*)(\boldsymbol{H}^* \otimes \boldsymbol{H}^*)\right] \\
&= 3\|\boldsymbol{H}\|_{\mathrm{F}}^4 + 6\,\|\boldsymbol{H}\|_{S_4}^4 \\
&\leq 9\|\boldsymbol{H}\|_{\mathrm{F}}^4\,,
\end{aligned}
\tag{51}
$$

where $\|\cdot\|_{S_4}$ denotes the Schatten-4 norm.

By plugging in (49) and (51) to (50), we obtain

$$
\mathbb{P}\left(|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}| \geq \tau\|\boldsymbol{H}\|_{\mathrm{F}}\right) \geq \frac{(1-\tau^2)^2}{18}\,.
$$

We obtain an analogous upper bound in the t-design case. With the isotropic normalization, $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy (6) and (8). Therefore,

$$
\begin{aligned}
\mathbb{E}|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}|^4 &= \mathbb{E}_{\boldsymbol{b}}(\boldsymbol{b}^*\boldsymbol{H}^* \otimes \boldsymbol{b}^*\boldsymbol{H}^*)\mathbb{E}_{\boldsymbol{a}}(\boldsymbol{a}\boldsymbol{a}^* \otimes \boldsymbol{a}\boldsymbol{a}^*)(\boldsymbol{H}\boldsymbol{b} \otimes \boldsymbol{H}\boldsymbol{b}) \\
&= \frac{d_1}{d_1+1}\left(\mathbb{E}_{\boldsymbol{b}}\|\boldsymbol{b}^*\boldsymbol{H}^*\|_{\mathrm{F}}^4 + \|\boldsymbol{b}^*\boldsymbol{H}^*\|_{S_4}^4\right) \\
&= \frac{2d_1}{d_1+1} \cdot \mathbb{E}_{\boldsymbol{b}}\|\boldsymbol{b}^*\boldsymbol{H}^*\|_2^4 \\
&= \frac{2d_1}{d_1+1} \cdot \mathrm{tr}\left[(\boldsymbol{H} \otimes \boldsymbol{H})\mathbb{E}_{\boldsymbol{b}}(\boldsymbol{b}\boldsymbol{b}^* \otimes \boldsymbol{b}\boldsymbol{b}^*)(\boldsymbol{H}^* \otimes \boldsymbol{H}^*)\right] \\
&= \frac{2d_1 d_2}{(d_1+1)(d_2+1)} \cdot \left(\|\boldsymbol{H}\|_{\mathrm{F}}^4 + \|\boldsymbol{H}\|_{S_4}^4\right) \\
&\leq \frac{4d_1 d_2}{(d_1+1)(d_2+1)} \cdot \|\boldsymbol{H}\|_{\mathrm{F}}^4\,.
\end{aligned}
\tag{52}
$$

By plugging in (49) and (52) to (50), we obtain

$$
\mathbb{P}\left(|\boldsymbol{a}^*\boldsymbol{H}\boldsymbol{b}| \geq \tau\|\boldsymbol{H}\|_{\mathrm{F}}\right) \geq \frac{(1-\tau^2)^2}{8}\,.
$$

B.2 Proof of Lemma 2

The independence and isotropy assumptions imply (48). Note that any $\boldsymbol{H} \in \mathcal{T}$ can be written as $\boldsymbol{H} = \boldsymbol{\Delta}_1\boldsymbol{V}_0^* + \boldsymbol{U}_0\boldsymbol{\Delta}_2^*$. Furthermore, without loss of generality, we may assume

$$
\langle \boldsymbol{\Delta}_1, \boldsymbol{U}_0 \rangle = 0\,,
\tag{53}
$$

which implies

$$
\|\boldsymbol{H}\|_{\mathrm{F}}^2 = \|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}_2\|_{\mathrm{F}}^2\,.
$$

Since

$$
\boldsymbol{a}_i^*\boldsymbol{H}\boldsymbol{b}_i = \mathrm{tr}\left((\boldsymbol{a}_i\boldsymbol{b}_i^*\boldsymbol{V}_0)^* \boldsymbol{\Delta}_1\right) + \mathrm{tr}\left((\boldsymbol{U}_0^*\boldsymbol{a}_i\boldsymbol{b}_i^*)^* \boldsymbol{\Delta}_2^*\right)\,,
$$

by [(53)] we have

$$\mathfrak{C}_n(\mathcal{T}) \le \mathbb{E} \sup_{\boldsymbol{H} \in \mathcal{T} \cap \mathbb{S}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \boldsymbol{a}_i^* \boldsymbol{H} \boldsymbol{b}_i \right|$$

$$\le \mathbb{E} \sup_{\|[\boldsymbol{\Delta}_1;\, \boldsymbol{\Delta}_2]\|_{\mathrm{F}}=1} \|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \cdot \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \boldsymbol{a}_i \boldsymbol{b}_i^* \boldsymbol{V}_0 \right\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_2\|_{\mathrm{F}} \cdot \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \boldsymbol{U}_0^* \boldsymbol{a}_i \boldsymbol{b}_i^* \right\|_{\mathrm{F}}$$

$$= \mathbb{E} \left(\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \boldsymbol{a}_i \boldsymbol{b}_i^* \boldsymbol{V}_0 \right\|_{\mathrm{F}}^2 + \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \boldsymbol{U}_0^* \boldsymbol{a}_i \boldsymbol{b}_i^* \right\|_{\mathrm{F}}^2 \right)^{1/2}$$

$$\le \left(\mathbb{E} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \boldsymbol{a}_i \boldsymbol{b}_i^* \boldsymbol{V}_0 \right\|_{\mathrm{F}}^2 + \mathbb{E} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \boldsymbol{U}_0^* \boldsymbol{a}_i \boldsymbol{b}_i^* \right\|_{\mathrm{F}}^2 \right)^{1/2}$$

$$= \sqrt{(d_1 + d_2)r} \,,$$

where the second step follows by the Cauchy-Schwarz inequality, the fourth step is obtained by Jensen's inequality, and the last step holds since $(\varepsilon_i)_{i=1}^n$ is a Rademacher sequence and $(\boldsymbol{a}, \boldsymbol{b})$ satisfies [(48)].

B.3 Proof of Lemma [4]

This lemma basically follows from requiring stationarity at $\boldsymbol{Q} = \boldsymbol{I}$ which is away from singularities of the objective. However, to avoid complications arising from derivatives with respect to complex-valued variables, we provide a "lower level" proof. To show this claim, first observe that any $(\boldsymbol{X}, \boldsymbol{Y})$ satisfying $\boldsymbol{X}\boldsymbol{Y}^* = \boldsymbol{M}_0$ can be parameterized by an *invertible $r \times r$ matrix $\boldsymbol{Q}$* as $(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X}_0 \boldsymbol{Q}, \boldsymbol{Y}_0 \boldsymbol{Q}^{-*})$. Let $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ be the maximizer considered in the statement of the lemma. Therefore, for any invertible $\boldsymbol{Q}$ in $\mathrm{GL}_r(\mathbb{C})$, the set of invertible $r \times r$ complex matrices, we should necessarily have

$$\langle \widetilde{\boldsymbol{X}}_0, \boldsymbol{X}_0 \boldsymbol{Q} \rangle + \langle \widetilde{\boldsymbol{Y}}_0, \boldsymbol{Y}_0 \boldsymbol{Q}^{-*} \rangle - \frac{1}{2n} \|\boldsymbol{Q}^* \boldsymbol{X}_0^* \boldsymbol{A}\|_{\mathrm{F}}^2 - \frac{1}{2n} \left\| \boldsymbol{Q}^{-1} \boldsymbol{Y}_0^* \boldsymbol{B} \right\|_{\mathrm{F}}^2$$

$$\le \langle \widetilde{\boldsymbol{X}}_0, \boldsymbol{X}_0 \rangle + \langle \widetilde{\boldsymbol{Y}}_0, \boldsymbol{Y}_0 \rangle - \frac{1}{2n} \|\boldsymbol{X}_0^* \boldsymbol{A}\|_{\mathrm{F}}^2 - \frac{1}{2n} \|\boldsymbol{Y}_0^* \boldsymbol{B}\|_{\mathrm{F}}^2 \,,$$

which is equivalent to

$$\langle \widetilde{\boldsymbol{X}}_0, \boldsymbol{X}_0 \boldsymbol{Q} - \boldsymbol{X}_0 \rangle + \langle \widetilde{\boldsymbol{Y}}_0, \boldsymbol{Y}_0 \boldsymbol{Q}^{-*} - \boldsymbol{Y}_0 \rangle - \frac{1}{n} \langle (\boldsymbol{Q} - \boldsymbol{I}_r)^* \boldsymbol{X}_0^* \boldsymbol{A}, \boldsymbol{X}_0^* \boldsymbol{A} \rangle$$

$$- \frac{1}{n} \langle (\boldsymbol{Q}^{-1} - \boldsymbol{I}_r) \boldsymbol{Y}_0^* \boldsymbol{B}, \boldsymbol{Y}_0^* \boldsymbol{B} \rangle - \frac{1}{2n} \|(\boldsymbol{Q} - \boldsymbol{I}_r)^* \boldsymbol{X}_0^* \boldsymbol{A}\|_{\mathrm{F}}^2 - \frac{1}{2n} \left\| (\boldsymbol{Q}^{-1} - \boldsymbol{I}_r) \boldsymbol{Y}_0^* \boldsymbol{B} \right\|_{\mathrm{F}}^2 \le 0 \,.$$

To simplify the notation let us define the short-hands

$$\boldsymbol{\Theta}_X = \boldsymbol{X}_0^* \left(\widetilde{\boldsymbol{X}}_0 - \frac{1}{n} \boldsymbol{A} \boldsymbol{A}^* \boldsymbol{X}_0 \right) ,$$

and

$$\boldsymbol{\Theta}_Y = \boldsymbol{Y}_0^* \left(\widetilde{\boldsymbol{Y}}_0 - \frac{1}{n} \boldsymbol{B} \boldsymbol{B}^* \boldsymbol{Y}_0 \right) .$$

The necessary inequality can be expressed as

$$\langle \boldsymbol{\Theta}_X, \boldsymbol{Q} - \boldsymbol{I}_r \rangle + \langle \boldsymbol{\Theta}_Y, (\boldsymbol{Q}^{-1} - \boldsymbol{I}_r)^* \rangle$$
$$\le \frac{1}{2n} \|(\boldsymbol{Q} - \boldsymbol{I}_r)^* \boldsymbol{X}_0^* \boldsymbol{A}\|_{\mathrm{F}}^2 + \frac{1}{2n} \left\| (\boldsymbol{Q}^{-1} - \boldsymbol{I}_r) \boldsymbol{Y}_0^* \boldsymbol{B} \right\|_{\mathrm{F}}^2. \tag{54}$$

Choosing $\boldsymbol{Q}$ within an arbitrarily small neighborhood of $\boldsymbol{I}_r$ allows us to use the identity

$$\boldsymbol{Q}^{-1} - \boldsymbol{I}_r = \sum_{k=1}^{\infty} (-1)^k (\boldsymbol{Q} - \boldsymbol{I}_r)^k.$$

Applying this identity in (54) yields

$$\langle \boldsymbol{\Theta}_X - \boldsymbol{\Theta}_Y^*, \boldsymbol{Q} - \boldsymbol{I}_r \rangle + \sum_{k=2}^{\infty} (-1)^k \langle \boldsymbol{\Theta}_Y^*, (\boldsymbol{Q} - \boldsymbol{I}_r)^k \rangle$$
$$\leq \frac{1}{2n} \|(\boldsymbol{Q} - \boldsymbol{I}_r)^* \boldsymbol{X}_0^* \boldsymbol{A}\|_{\mathrm{F}}^2 + \frac{1}{2n} \left\|(\boldsymbol{Q}^{-1} - \boldsymbol{I}_r) \boldsymbol{Y}_0^* \boldsymbol{B}\right\|_{\mathrm{F}}^2$$
$$\leq \frac{1}{2n} \|\boldsymbol{Q} - \boldsymbol{I}_r\|^2 \|\boldsymbol{X}_0^* \boldsymbol{A}\|_{\mathrm{F}}^2 + \frac{1}{2n} \|\boldsymbol{Q} - \boldsymbol{I}_r\|^2 \left\|\boldsymbol{Q}^{-1}\right\|^2 \|\boldsymbol{Y}_0^* \boldsymbol{B}\|_{\mathrm{F}}^2.$$

As $\boldsymbol{Q} \to \boldsymbol{I}_r$, the terms linear in $\boldsymbol{Q} - \boldsymbol{I}_r$ dominate and the terms with superlinear dependence on $\|\boldsymbol{Q} - \boldsymbol{I}_r\|$ vanish faster. Therefore, the inequality above implies

$$\boldsymbol{\Theta}_X = \boldsymbol{\Theta}_Y^*. \tag{55}$$

Let us express (31) as $\boldsymbol{E} = [\boldsymbol{E}_X; \boldsymbol{E}_Y]$ with $\boldsymbol{E}_X = \widetilde{\boldsymbol{X}_0} - \frac{1}{n} \boldsymbol{A} \boldsymbol{A}^* \boldsymbol{X}_0$ and $\boldsymbol{E}_Y = \overline{\widetilde{\boldsymbol{Y}_0}} - \frac{1}{n} \overline{\boldsymbol{B}} \boldsymbol{B}^\mathsf{T} \overline{\boldsymbol{Y}_0}$. Then (55) implies that $[\boldsymbol{E}_X;\ \boldsymbol{E}_Y]$ belongs to the subspace

$$\mathbb{W} \stackrel{\text{def}}{=} \left\{ \left[\boldsymbol{E}_1;\ \boldsymbol{E}_2\right] : \boldsymbol{E}_1 \in \mathbb{C}^{d_1 \times r}, \boldsymbol{E}_2 \in \mathbb{C}^{d_2 \times r}, \boldsymbol{X}_0^* \boldsymbol{E}_1 = \boldsymbol{E}_2^\mathsf{T} \boldsymbol{Y}_0 \right\}.$$

It only remains to show that $\mathbb{W}$ coincides with $\mathbb{V} = \mathrm{null}(\mathcal{L})^\perp$ defined in (34).

We first verify that $\mathbb{V} \subseteq \mathbb{W}$. By the definition of $\mathcal{T}$, we can express $\mathbb{V}$ equivalently as

$$\mathbb{V} = \left\{ \left[\boldsymbol{Z} \boldsymbol{Y}_0;\ \boldsymbol{Z}^\mathsf{T} \overline{\boldsymbol{X}_0}\right] : \boldsymbol{Z} \in \mathcal{T} \right\}.$$

Suppose that $[\boldsymbol{E}_1;\ \boldsymbol{E}_2]$ belongs to $\mathbb{V}$. Then there exists $\boldsymbol{Z} \in \mathcal{T}$ such that $\boldsymbol{E}_1 = \boldsymbol{Z} \boldsymbol{Y}_0$ and $\boldsymbol{E}_2 = \boldsymbol{Z}^\mathsf{T} \overline{\boldsymbol{X}_0}$. Therefore,

$$\boldsymbol{X}_0^* \boldsymbol{E}_1 = \boldsymbol{X}_0^* \boldsymbol{Z} \boldsymbol{Y}_0 = (\boldsymbol{Z}^\mathsf{T} \overline{\boldsymbol{X}_0})^\mathsf{T} \boldsymbol{Y}_0 = \boldsymbol{E}_2^\mathsf{T} \boldsymbol{Y}_0,$$

which implies $\left[\boldsymbol{E}_1;\ \boldsymbol{E}_2\right]$ is contained in $\mathbb{W}$. Since $[\boldsymbol{E}_1;\ \boldsymbol{E}_2]$ can be chosen arbitrarily in $\mathbb{V}$, we have shown $\mathbb{V} \subseteq \mathbb{W}$.

Furthermore, the orthogonal complement of $\mathbb{W}$ within $\mathbb{C}^{(d_1+d_2) \times r}$ can be written as

$$\mathbb{W}^\perp = \left\{ \left[\boldsymbol{X}_0 \boldsymbol{S}; -\overline{\boldsymbol{Y}_0} \boldsymbol{S}^\mathsf{T}\right] : \boldsymbol{S} \in \mathbb{C}^{r \times r} \right\}.$$

Since both $\boldsymbol{X}_0$ and $\boldsymbol{Y}_0$ are full column rank (otherwise, the rank of $\boldsymbol{M}_0 = \boldsymbol{X}_0 \boldsymbol{Y}_0^*$ would be smaller than r), we deduce that the dimension of $\mathbb{W}^\perp$ is r^2, thereby the dimension of $\mathbb{W}$ is $r(d_1 + d_2 - r)$. In view of the inclusion $\mathbb{V} \subseteq \mathbb{W}$, it only remains to show that the dimension of $\mathbb{V}$ is $r(d_1 + d_2 - r)$. We do so by arguing that the linear function $\boldsymbol{Z} \mapsto [\boldsymbol{Z} \boldsymbol{Y}_0;\ \boldsymbol{Z}^\mathsf{T} \overline{\boldsymbol{X}_0}]$ is a *bijection* from $\mathcal{T}$ to $\mathbb{V}$, or equivalently if $\boldsymbol{Z} \in \mathcal{T}$ is mapped to $\boldsymbol{0} \in \mathbb{V}$, then $\boldsymbol{Z} = \boldsymbol{0}$. If for some $\boldsymbol{Z} = \boldsymbol{\Delta}_1 \boldsymbol{Y}_0^* + \boldsymbol{X}_0 \boldsymbol{\Delta}_2^* \in \mathcal{T}$ we have $\boldsymbol{X}_0^* \boldsymbol{Z} = \boldsymbol{0}$ and $\boldsymbol{Z} \boldsymbol{Y}_0 = \boldsymbol{0}$, then we should have $\|\boldsymbol{Z}\|_{\mathrm{F}}^2 = \mathrm{tr}(\boldsymbol{Z}^* \boldsymbol{Z}) = 0$, and consequently $\boldsymbol{Z} = \boldsymbol{0}$. Therefore, we have shown that $\mathbb{V} = \mathbb{W}$, and particularly

$$\boldsymbol{E} = [\boldsymbol{E}_X;\ \boldsymbol{E}_Y] \in \mathbb{W} = \mathbb{V}.$$

Next we prove the second part of the lemma. By (26), we have

$$
(1-\eta)\left\|\boldsymbol{X}_0 - \widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}^2 + (1-\eta)\left\|\boldsymbol{Y}_0 - \widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}^2
$$
$$
\leq \frac{1}{n}\left\|\boldsymbol{A}^*(\boldsymbol{X}_0 - \widetilde{\boldsymbol{X}}_0)\right\|_{\mathrm{F}}^2 + \frac{1}{n}\left\|\boldsymbol{B}^*(\boldsymbol{Y}_0 - \widetilde{\boldsymbol{Y}}_0)\right\|_{\mathrm{F}}^2
$$
$$
= \frac{1}{n}\left\|\boldsymbol{A}^*\boldsymbol{X}_0\right\|_{\mathrm{F}}^2 - 2\left\langle \boldsymbol{X}_0, \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\widetilde{\boldsymbol{X}}_0 \right\rangle + \frac{1}{n}\left\|\boldsymbol{A}^*\widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}^2
$$
$$
+ \frac{1}{n}\left\|\boldsymbol{B}^*\boldsymbol{Y}_0\right\|_{\mathrm{F}}^2 - 2\left\langle \boldsymbol{Y}_0, \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\widetilde{\boldsymbol{Y}}_0 \right\rangle + \frac{1}{n}\left\|\boldsymbol{B}^*\widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}^2
$$
$$
= \frac{1}{n}\left\|\boldsymbol{A}^*\boldsymbol{X}_0\right\|_{\mathrm{F}}^2 - 2\langle \boldsymbol{X}_0, \widetilde{\boldsymbol{X}}_0\rangle + 2\left\langle \boldsymbol{X}_0, \left(\boldsymbol{I}_{d_1} - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\right)\widetilde{\boldsymbol{X}}_0 \right\rangle + \frac{1}{n}\left\|\boldsymbol{A}^*\widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}^2
$$
$$
+ \frac{1}{n}\left\|\boldsymbol{B}^*\boldsymbol{Y}_0\right\|_{\mathrm{F}}^2 - 2\langle \boldsymbol{Y}_0, \widetilde{\boldsymbol{Y}}_0\rangle + 2\left\langle \boldsymbol{Y}_0, \left(\boldsymbol{I}_{d_2} - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\right)\widetilde{\boldsymbol{Y}}_0 \right\rangle + \frac{1}{n}\left\|\boldsymbol{B}^*\widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}^2.
$$

Since $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ is the maximizer, for any $\boldsymbol{X} \in \mathbb{C}^{d_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{C}^{d_2 \times r}$ satisfying $\boldsymbol{X}\boldsymbol{Y}^* = \boldsymbol{M}_0$, we can continue by

$$
(1-\eta)\left\|\boldsymbol{X}_0 - \widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}^2 + (1-\eta)\left\|\boldsymbol{Y}_0 - \widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}^2
$$
$$
\leq \frac{1}{n}\left\|\boldsymbol{A}^*\boldsymbol{X}\right\|_{\mathrm{F}}^2 - 2\langle \boldsymbol{X}, \widetilde{\boldsymbol{X}}_0\rangle + 2\left\langle \boldsymbol{X}_0, \left(\boldsymbol{I}_{d_1} - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\right)\widetilde{\boldsymbol{X}}_0 \right\rangle + \frac{1}{n}\left\|\boldsymbol{A}^*\widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}^2
$$
$$
+ \frac{1}{n}\left\|\boldsymbol{B}^*\boldsymbol{Y}\right\|_{\mathrm{F}}^2 - 2\langle \boldsymbol{Y}, \widetilde{\boldsymbol{Y}}_0\rangle + 2\left\langle \boldsymbol{Y}_0, \left(\boldsymbol{I}_{d_2} - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\right)\widetilde{\boldsymbol{Y}}_0 \right\rangle + \frac{1}{n}\left\|\boldsymbol{B}^*\widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}^2
$$
$$
= \frac{1}{n}\left\|\boldsymbol{A}^*(\boldsymbol{X} - \widetilde{\boldsymbol{X}}_0)\right\|_{\mathrm{F}}^2 + 2\left\langle \boldsymbol{X}_0 - \boldsymbol{X}, \left(\boldsymbol{I}_{d_1} - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\right)\widetilde{\boldsymbol{X}}_0 \right\rangle
$$
$$
+ \frac{1}{n}\left\|\boldsymbol{B}^*(\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}_0)\right\|_{\mathrm{F}}^2 + 2\left\langle \boldsymbol{Y}_0 - \boldsymbol{Y}, \left(\boldsymbol{I}_{d_2} - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\right)\widetilde{\boldsymbol{Y}}_0 \right\rangle
$$
$$
\leq (1+\eta)\left\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}^2 + 2\eta\left\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}\left\|\widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}
$$
$$
+ (1+\eta)\left\|\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}^2 + 2\eta\left\|\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}\left\|\widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}},
$$

where the last step follows from (26). Finally note that

$$
\left\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}\left\|\widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}} + \left\|\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}\left\|\widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}
$$
$$
\leq \left(\left\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}^2\right)^{1/2}\left(\left\|\widetilde{\boldsymbol{X}}_0\right\|_{\mathrm{F}}^2 + \left\|\widetilde{\boldsymbol{Y}}_0\right\|_{\mathrm{F}}^2\right)^{1/2}.
$$

This completes the proof.

C Proofs of the supporting lemmas

C.1 Proof of Lemma 6

It suffices to show the upper bound for some $\boldsymbol{Q}$ in the *orthogonal group* $\mathcal{O}_r(\mathbb{C})$. Let $\boldsymbol{Q}$ be given by

$$
\boldsymbol{Q} \in \operatorname*{argmin}_{\boldsymbol{R} \in \mathbb{R}^{r \times r}}\left\{\left\|\begin{bmatrix}\boldsymbol{U}_0 \\ \boldsymbol{V}_0\end{bmatrix}\boldsymbol{R} - \begin{bmatrix}\widetilde{\boldsymbol{U}}_0 \\ \widetilde{\boldsymbol{V}}_0\end{bmatrix}\right\|_{\mathrm{F}} : \boldsymbol{R}^*\boldsymbol{R} = \boldsymbol{I}_r\right\}.
$$

Then it follows that

$$
\begin{aligned}
\left\| \begin{bmatrix} \widetilde{U}_0 \\ \widetilde{V}_0 \end{bmatrix} \widetilde{\Sigma}_0^{1/2} - \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} \Sigma_0^{1/2} Q \right\|_{\mathrm{F}} &= \left\| \begin{bmatrix} \widetilde{U}_0 \\ \widetilde{V}_0 \end{bmatrix} \widetilde{\Sigma}_0^{1/2} - \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} QQ^* \Sigma_0^{1/2} Q \right\|_{\mathrm{F}} \\
&\le \left\| \left(\begin{bmatrix} \widetilde{U}_0 \\ \widetilde{V}_0 \end{bmatrix} - \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} Q \right) \widetilde{\Sigma}_0^{1/2} \right\|_{\mathrm{F}} + \left\| \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} Q(\widetilde{\Sigma}_0^{1/2} - Q^* \Sigma_0^{1/2} Q) \right\|_{\mathrm{F}} \\
&\le \sqrt{\sigma_1(\widetilde{\Sigma}_0)} \left\| \begin{bmatrix} \widetilde{U}_0 \\ \widetilde{V}_0 \end{bmatrix} - \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} Q \right\|_{\mathrm{F}} + \left\| \widetilde{\Sigma}_0^{1/2} - Q^* \Sigma_0^{1/2} Q \right\|_{\mathrm{F}}.
\end{aligned}
$$

Since $\widetilde{\Sigma}_0^{1/2}$ (resp. $Q^* \Sigma_0^{1/2} Q$) is the matrix square root of $\widetilde{\Sigma}_0$ (resp. $Q^* \Sigma_0 Q$), by the perturbation bound by Schmitt [43, equation (1.3)], we obtain

$$
\left\| \widetilde{\Sigma}_0^{1/2} - Q^* \Sigma_0^{1/2} Q \right\|_{\mathrm{F}} \le \frac{\left\| \widetilde{\Sigma}_0 - Q^* \Sigma_0 Q \right\|_{\mathrm{F}}}{\sqrt{\sigma_r(\widetilde{\Sigma}_0)} + \sqrt{\sigma_r(\Sigma_0)}} \le \frac{\left\| \widetilde{\Sigma}_0 - Q^* \Sigma_0 Q \right\|_{\mathrm{F}}}{\sqrt{\sigma_r(\Sigma_0)}}.
$$

On the other hand, by the triangle inequality, we have

$$
\begin{aligned}
\| \widetilde{M}_0 - M_0 \|_{\mathrm{F}} &= \| \widetilde{U}_0 \widetilde{\Sigma}_0 \widetilde{V}_0^* - U_0 QQ^* \Sigma_0 QQ^* V_0^* \|_{\mathrm{F}} \\
&\ge \| U_0 Q(\widetilde{\Sigma}_0 - Q^* \Sigma_0 Q) \widetilde{V}_0^* \|_{\mathrm{F}} - \| (\widetilde{U}_0 - U_0 Q) \widetilde{\Sigma}_0 \widetilde{V}_0^* \|_{\mathrm{F}} \\
&\quad - \| U_0 QQ^* \Sigma_0 Q(\widetilde{V}_0 - V_0 Q)^* \|_{\mathrm{F}} \\
&\ge \| \widetilde{\Sigma}_0 - Q^* \Sigma_0 Q \|_{\mathrm{F}} - \sigma_1(\widetilde{\Sigma}_0) \| \widetilde{U}_0 - U_0 Q \|_{\mathrm{F}} - \sigma_1(\Sigma_0) \| \widetilde{V}_0 - V_0 Q \|_{\mathrm{F}}.
\end{aligned}
$$

By rearranging the above inequality, we obtain

$$
\| \widetilde{\Sigma}_0 - Q^* \Sigma_0 Q \|_{\mathrm{F}} \le \| \widetilde{M}_0 - M_0 \|_{\mathrm{F}} + \sigma_1(\widetilde{\Sigma}_0) \| \widetilde{U}_0 - U_0 Q \|_{\mathrm{F}} + \sigma_1(\Sigma_0) \| \widetilde{V}_0 - V_0 Q \|_{\mathrm{F}}.
$$

By combining the results with Weyl's inequality, we obtain

$$
\begin{aligned}
\left\| \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} \Sigma_0^{1/2} Q - \begin{bmatrix} \widetilde{U}_0 \\ \widetilde{V}_0 \end{bmatrix} \widetilde{\Sigma}_0^{1/2} \right\|_{\mathrm{F}} &\le \sqrt{\sigma_1(\widetilde{\Sigma}_0)} \left\| \begin{bmatrix} \widetilde{U}_0 \\ \widetilde{V}_0 \end{bmatrix} - \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} Q \right\|_{\mathrm{F}} \\
&\quad + \frac{\| \widetilde{M}_0 - M_0 \|_{\mathrm{F}} + \sigma_1(\widetilde{\Sigma}_0) \| \widetilde{U}_0 - U_0 Q \|_{\mathrm{F}} + \sigma_1(\Sigma_0) \| \widetilde{V}_0 - V_0 Q \|_{\mathrm{F}}}{\sqrt{\sigma_r(\Sigma_0)}} \\
&\le \frac{\| \widetilde{M}_0 - M_0 \|_{\mathrm{F}}}{\sqrt{\sigma_r(\Sigma_0)}} + \sqrt{2\sigma_1(\Sigma_0)} \left(1 + 2\sqrt{\frac{\sigma_1(\Sigma_0)}{\sigma_r(\Sigma_0)}} \right) \cdot \underbrace{\left\| \begin{bmatrix} \widetilde{U}_0 \\ \widetilde{V}_0 \end{bmatrix} - \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} Q \right\|_{\mathrm{F}}}_{(\sharp)}.
\end{aligned}
$$

Finally, $(\sharp)$ is bounded by a variant of the Davis-Kahan theorem by Dopico [22, Theorem 2.1] as follows:

$$
\left\| \begin{bmatrix} \widetilde{U}_0 \\ \widetilde{V}_0 \end{bmatrix} - \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} Q \right\|_{\mathrm{F}} \le \frac{\sqrt{2(\| (\widetilde{M}_0 - M_0) \widetilde{V}_0 \|_{\mathrm{F}}^2 + \| \widetilde{U}_0^* (\widetilde{M}_0 - M_0) \|_{\mathrm{F}}^2)}}{\sigma_r(\Sigma_0) - \| \widetilde{M}_0 - M_0 \|}.
$$

C.2   Proof of Lemma 7

Since $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ is a maximizer to (32), we have

$$
\langle \widetilde{\boldsymbol{X}}_0, \boldsymbol{X}_0 - \boldsymbol{X} \rangle + \langle \widetilde{\boldsymbol{Y}}, \boldsymbol{Y}_0 - \boldsymbol{Y} \rangle
$$
$$
\geq \frac{1}{2n} \left(\|\boldsymbol{A}^*\boldsymbol{X}_0\|_{\mathrm{F}}^2 - \|\boldsymbol{A}^*\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{B}^*\boldsymbol{Y}_0\|_{\mathrm{F}}^2 - \|\boldsymbol{B}^*\boldsymbol{Y}\|_{\mathrm{F}}^2 \right) \tag{56}
$$
$$
\geq \frac{1}{n} \langle \boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}, \boldsymbol{X}_0 - \boldsymbol{X} \rangle + \frac{1-\eta}{2} \|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{1}{n} \langle \boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y}, \boldsymbol{Y}_0 - \boldsymbol{Y} \rangle + \frac{1-\eta}{2} \|\boldsymbol{Y}_0 - \boldsymbol{Y}\|_{\mathrm{F}}^2,
$$

where the second inequality follows from the strong convexity of $\|\boldsymbol{A}^*\boldsymbol{X}_0\|_{\mathrm{F}}^2$ and $\|\boldsymbol{B}^*\boldsymbol{Y}_0\|_{\mathrm{F}}^2$, which are quadratic functions of $\boldsymbol{X}_0$ and $\boldsymbol{Y}_0$, respectively. Then (56) is rearranged to

$$
\frac{1-\eta}{2} \|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{1-\eta}{2} \|\boldsymbol{Y}_0 - \boldsymbol{Y}\|_{\mathrm{F}}^2
$$
$$
\leq \langle \widetilde{\boldsymbol{X}} - \frac{1}{n} \boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}, \boldsymbol{X}_0 - \boldsymbol{X} \rangle + \langle \widetilde{\boldsymbol{Y}}_0 - \frac{1}{n} \boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y}, \boldsymbol{Y}_0 - \boldsymbol{Y} \rangle
$$
$$
\leq \sqrt{\|\boldsymbol{X}_0 - \boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}_0 - \boldsymbol{Y}\|_{\mathrm{F}}^2} \cdot \sqrt{\left\|\widetilde{\boldsymbol{X}}_0 - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}\right\|_{\mathrm{F}}^2 + \left\|\widetilde{\boldsymbol{Y}}_0 - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y}\right\|_{\mathrm{F}}^2},
$$

where the last step follows from the Cauchy-Schwarz inequality.

Therefore it follows that

$$
\left\| \left[\widetilde{\boldsymbol{X}}_0 - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}; \ \widetilde{\boldsymbol{Y}}_0 - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y} \right] \right\|_{\mathrm{F}} \geq \frac{1-\eta}{2} \left\| [\boldsymbol{X}_0 - \boldsymbol{X}; \ \boldsymbol{Y}_0 - \boldsymbol{Y}] \right\|_{\mathrm{F}}.
$$

Finally by the triangle inequality we obtain

$$
\left\| \left[\widetilde{\boldsymbol{X}}_0 - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}_0; \ \widetilde{\boldsymbol{Y}}_0 - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y}_0 \right] \right\|_{\mathrm{F}}
$$
$$
\leq \left\| \left[\widetilde{\boldsymbol{X}}_0 - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}; \ \widetilde{\boldsymbol{Y}}_0 - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y} \right] \right\|_{\mathrm{F}} + \left\| \left[\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*(\boldsymbol{X}_0 - \boldsymbol{X}); \ \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*(\boldsymbol{Y}_0 - \boldsymbol{Y}) \right] \right\|_{\mathrm{F}}
$$
$$
\leq \left\| \left[\widetilde{\boldsymbol{X}}_0 - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}; \ \widetilde{\boldsymbol{Y}}_0 - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y} \right] \right\|_{\mathrm{F}} + (1+\eta) \left\| [\boldsymbol{X}_0 - \boldsymbol{X}; \ \boldsymbol{Y}_0 - \boldsymbol{Y}] \right\|_{\mathrm{F}}
$$
$$
\leq \left\| \left[\widetilde{\boldsymbol{X}} - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}; \ \widetilde{\boldsymbol{Y}}_0 - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y} \right] \right\|_{\mathrm{F}} + \frac{2(1+\eta)}{1-\eta} \left\| \left[\widetilde{\boldsymbol{X}}_0 - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}; \ \widetilde{\boldsymbol{Y}} - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y} \right] \right\|_{\mathrm{F}}
$$
$$
\leq \left(1 + \frac{2(1+\eta)}{1-\eta} \right) \left\{ \left\| \left[\widetilde{\boldsymbol{X}}_0 - \boldsymbol{X}; \ \widetilde{\boldsymbol{Y}}_0 - \boldsymbol{Y} \right] \right\|_{\mathrm{F}} + \left\| \left[\boldsymbol{X} - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}; \ \boldsymbol{Y} - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y} \right] \right\|_{\mathrm{F}} \right\}.
$$

Finally note that

$$
\left\| \left[\boldsymbol{X} - \frac{1}{n}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{X}; \ \boldsymbol{Y} - \frac{1}{n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{Y} \right] \right\|_{\mathrm{F}} \leq \eta \left\| [\boldsymbol{X}; \ \boldsymbol{Y}] \right\|_{\mathrm{F}},
$$

which completes the proof.

References

[1] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2013.

[2] S. Bahmani. Estimation from nonlinear observations via convex programming with application to bilinear regression. *Electronic Journal of Statistics*, 13(1):1978–2011, 2019.

[3] S. Bahmani and J. Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. In *Artificial Intelligence and Statistics*, pages 252–260, 2017.

[4] S. Bahmani and J. Romberg. Solving equations of random convex functions via anchored regression. *Foundations of Computational Mathematics*, 19(4):813–841, 2019.

[5] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin. Painless reconstruction from magnitudes of frame coefficients. *Journal of Fourier Analysis and Applications*, 15(4):488–501, 2009.

[6] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[7] J. Bourgain, S. Dirksen, and J. Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. *Geometric and Functional Analysis*, 25(4):1009–1088, 2015.

[8] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. ISSN 1935-8237. doi: 10.1561/2200000050.

[9] T. T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[10] V. Cambareri and L. Jacques. Through the haze: a non-convex approach to blind gain calibration for linear random sensing models. *Information and Inference: A Journal of the IMA*, 8(2):205–271, 04 2018. ISSN 2049-8772. doi: 10.1093/imaiai/iay004.

[11] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053, 2010.

[12] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Review*, 57(2):225–251, 2015.

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

[14] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[15] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[16] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[17] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

[18] V. de la P̃ena and E. Giné. *Decoupling: from dependence to independence.* Springer Science & Business Media, 2012.

[19] P. Delsarte, J.-M. Goethals, and J. J. Seidel. Spherical codes and designs. In *Geometry and Combinatorics*, pages 68–93. Elsevier, 1991.

[20] S. Dirksen. *Noncommutative and vector-valued Rosenthal inequalities.* PhD thesis, Delft University of Technology, 2011.

[21] S. Dirksen, G. Lecué, and H. Rauhut. On the gap between restricted isometry properties and sparse recovery conditions. *IEEE Transactions on Information Theory*, 64(8):5478–5487, 2016.

[22] F. M. Dopico. A note on sin θ theorems for singular subspace variations. *BIT Numerical Mathematics*, 40(2):395–403, 2000.

[23] S. Foucart and S. Subramanian. Iterative hard thresholding for low-rank recovery from rank-one projections. *Linear Algebra and its Applications*, 572:117–134, 2019. ISSN 0024-3795.

[24] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, 2018.

[25] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[26] D. Gross, F. Krahmer, and R. Kueng. A partial derandomization of phaselift using spherical designs. *Journal of Fourier Analysis and Applications*, 21(2):229–266, 2015.

[27] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.

[28] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[29] M. Junge and Q. Zeng. Noncommutative Bennett and Rosenthal inequalities. *The Annals of Probability*, 41(6):4287–4316, 2013.

[30] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[31] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[32] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

[33] J. M. Landsberg. Tensors: Geometry and applications. *Representation theory*, 381(402):3, 2012.

[34] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

[35] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. *arXiv preprint arXiv:1802.06286*, 2018.

[36] M. Lin and J. Ye. A non-convex one-pass framework for generalized factorization machine and rank-one matrix sensing. In *Advances in Neural Information Processing Systems*, pages 1633–1641, 2016.

[37] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[38] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.

[39] S. Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[40] R. E. A. C. Paley and A. Zygmund. A note on analytic functions in the unit circle. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 266–272. Cambridge University Press, 1932.

[41] J. T. Parker, P. Schniter, and V. Cevher. Bilinear generalized approximate message passing–Part I: Derivation. *IEEE Transactions on Signal Processing*, 62(22):5839–5853, 2014.

[42] S. Sanghavi, R. Ward, and C. D. White. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, 71(3-4):569–608, 2017.

[43] B. A. Schmitt. Perturbation bounds for matrix square roots and Pythagorean sums. *Linear algebra and its applications*, 174:215–227, 1992.

[44] M. Soltani and C. Hegde. Improved algorithms for matrix recovery from rank-one projections. *arXiv preprint arXiv:1705.07469*, 2017.

[45] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. Neural Information Processing Series. MIT Press, 2011. ISBN 9780262537766.

[46] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

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

[48] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[49] A. W. van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, 1996.

[50] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[51] K. Zhong, P. Jain, and I. S. Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *International conference on algorithmic learning theory*, pages 3–18. Springer, 2015.