Fast Exact Recovery of Noisy Matrix from Few Entries: an Infinity Norm Approach

December 2, 2024

Abstract

The matrix completion problem is to recover a matrix A from a limited sample of revealed entries. While such a task is impossible in general, a series of works by Candes, Recht and Tao (2009 and 2010) proved that a nuclear norm minimization approach can recover A exactly, with high probability, from a random sample of observations, under three (basic) assumptions: (1) the rank of A is very small compared to its dimensions (low rank), (2) A has delocalized singular vectors (incoherence), and (3) the sample size is sufficiently large.

Spectral algorithms, based on low rank approximation, is another way to tackle the problem. This approach is usally faster in practice, but at the cost of extra assumptions. Most notably, one need to require A to have a small condition number (Keshavan-Montanari-Oh (2010) or the gaps between consecutive eigenvalues are sufficiently large Bhardwaj-Vu (2023)).

In this paper, we propose a simple low-rank approximation algorithm that fully recovers every entry of A up to an arbitrarily small error with high probability. Once this error is smaller than (half of) the discretization level of input, one obtains a perfect recovery by rounding off each entry.

We assume only the three basic conditions above. In particular, we do not need a requirement on the condition number or the gaps. Our incoherence condition is also more relaxed than all previous works, and the sample size is optimal up to a polylogarithmic factor.

The mathematics behind this algorithm is quite different from other approaches. Using the contour integral method from operator theory combined with combinatorial ideas, we show that under some mild conditions, the best rank p approximations of two matrices A and $A+E$ (where E represents noise) are close in the infinity norm.

1 Introduction: the matrix completion problem

1.1 Problem description

A large matrix $A \in \mathbb{R}^{m\times n}$ is hidden, except for a few revealed entries in a set $\Omega \in [m] \times [n]$. We call Ω the set of *observations* or *samples*. The matrix A_Ω, defined by

$$(A_\Omega)_{ij} = A_{ij} \quad \text{for} \quad (i,j) \in \Omega, \quad \text{and 0 otherwise.} \tag{1}$$

is called the *observed* or *sample* matrix. The problem of recovering A, given Ω and A_Ω, is the *matrix completion* problem. In many real-life data, A_Ω is corrupted by an additive noise ξ before observation, resulting in an even harder problem. We reserve the discussion of this "noisy" version for the end of this introduction and focus on the noiseless case for now.

There are many types of recovery/completion:

- *Full:* Every entry of A has to be determined. This recovery can be *exact*, meaning each entry has to be computed exactly; or *approximate* up to some additive error $\varepsilon > 0$.
- *Partial:* The output has to agree with A in at least a $(1-\varepsilon)$ portion of the entries. Typically, a partial recovery algorithm cannot choose the entries to recover (otherwise one can run it repeatedly to achieve full recovery). Again, the computation can be either exact or approximate.
- *Norm:* Given a matrix norm f, there are two version of norm recovery. In the *absolute* version, the output $\hat{A}$ should satisfy $f(\hat{A} - A) \leq \varepsilon$ for an absolute error margin ε. In the *relative* version, one replaces the right-hand side with $\varepsilon f(A)$, where ε is the relative error margin. The desirable norm, version and error margin are determined by the specific application.

Remark 1.1. As our data (and our computers as well) have finite precision, to recover an entry exactly and with very small error are actually the same problem. If all entries of A have the form $k\varepsilon$ for some integer k, and one can recover all of them within an error less than $\varepsilon/2$, then one can obtain the perfect recovery by just rounding off each entry to the nearest multiple of ε.

For instance, in the Netflix problem, the rating of movies are half integers between 1 and 5 (so $\varepsilon = 1/2$). This setting is typical for most recommendation systems.

In this paper, we focus on **full, exact, recovery** for **discrete data**. By Remark 1.1, this is simply equivalent to obtaining an approximate recovery with sufficiently small error term.

The rest of this introduction is organized as follows: In section 1.2, we will introduce the common setting used in most existing literature and discuss why they are desirable for full recovery. In section 1.3, we will describe and compare several approaches to solve the problem and some representative works for each. In section 1.5, we will introduce our own algorithm and our setting, and discuss several advantages, compared to the existing methods. In section 2, we discuss the new mathematical ideas behind our algorithms. We believe that our main mathematical tool (Theorem 2.2) is of independent interest.

1.2 Common settings and assumptions

Before beginning, we define some notation for convenience:

- Let the SVD of A be given by $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, where $r = \operatorname{rank} A$.
- For two real numbers a and b, let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. When discussing A, we denote $N := \max\{m, n\}$.
- The *coherence parameter* of U is given by

$$\mu(U) := \max_{i \in [m]} \frac{m}{r} \|e_i^T U\|^2 = \frac{m\|U\|_{2,\infty}}{r}, \tag{2}$$

  where the 2-to-∞ norm of a matrix M is given by $\|M\|_{2,\infty} := \sup\{\|Mu\|_\infty : \|u\|_2 = 1\}$, and is just the largest row norm of M. Define $\mu(V)$ similarly. In the context of U and V being singular bases of A, we let $\mu_0(A) = \max\{\mu(U), \mu(V)\}$ and simply use μ_0 when A is clear from the context.

  Note that the notion of coherence appears in many fields, and many works use the stronger definition $\mu(U) = m\|U\|_\infty^2$, where the infinity norm is simply the absolute value of the largest entry. We stick to the definition above, which is used widely in matrix completion papers.

- Some papers use the following parameter:

$$\mu_1 = \max_{i\in[m],j\in[n]} \frac{\sqrt{mn}}{\sqrt{r}} |e_i^T U^T V e_j| = \frac{\sqrt{mn}}{\sqrt{r}} \|UV^T\|_\infty. \tag{3}$$

  We could not find any widely used name for this parameter in the literature, and temporarily use the name *joint coherence parameter* in this paper. Note that unless UV^T exhibits a special delocalized structure, the best upper bound for μ_1 is $\mu_0\sqrt{r}$.

If we know only A_Ω, filling out the missing entries is clearly impossible, unless extra assumptions are given. Most existing works made assumptions of the following types:

- *Low-rank:* It is clear that when A has full rank, one cannot find any hidden entry even when knowing the rest. If $\operatorname{rank} A = r$, the degree of freedom for A is $r(m+n-1)$ [5], so one needs r much smaller than $\min\{m, n\}$ to be able to compute A from a small sample.

  Many papers assume A has *low rank*; specifically $r = O(1)$ (while $m, n \to \infty$) in some. Others have no explicit bounds but their results are only meaningful (better than the trivial choice $|\Omega| = mn$) when r is smaller than some function of m and n.

- *Incoherence:* It is commonly assumed that the rows and columns of A are sufficiently "spread out", so the information does not concentrate in a small set of entries easily overlooked by random sampling. In technical terms, one requires μ_0, and μ_1 for some papers, to be small. Like the low rank condition, some papers require μ_0 (and μ_1 for some) to be $O(1)$, while others have implicit bounds through the *sample size condition*, discussed in more details below.

- *Observations are random:* Broadly speaking, the choice of Ω must be *non-adversarial*. For example, if A_Ω misses an entire column or row, then it is impossible to recover it even when A is low-rank. Most papers assume that Ω is *random*, most commonly one of the two models: (1) Ω is sampled uniformly among subsets with the same size, or (2) that Ω has independently chosen entries, each with the same probability p, which can be known or hidden.

  It is usually not too hard to replace the former model by the latter, using the standard conditioning trick. One samples the entries independently, and condition on the event that the sample size equals a given number.

- *Density assumption:* A coupon collector effect shows that both random sampling models above need at least $N \log N$ observations to avoid empty rows or columns. Another lower bound is given by the degree of freedom: One needs to know $r(m+n-1)$ parameters to compute A exactly. A more elaborate argument in [5] gives the lower bound $|\Omega| \gtrsim rN \log N$. This is equivalent to $p \gtrsim r(m^{-1} + n^{-1}) \log N$ for the independent sampling model.

  Most results prove that some type of recovery is possible under a *sample size* condition of the form $|\Omega| \gtrsim f(\mu_0, \mu_1) r^a N \log^b N$ for constants $a, b \geq 1$ and positive function f. This needs to be smaller than mn to be meaningful, which implicitly gives bounds on the rank and the coherence parameters.

- *Centered noise:* If the sampling is noisy, it is common to assume that the noise has zero mean, so that A_Ω has unbiased entries.

1.3 A brief summary

The literature on matrix completion is huge. In this section, we try to summarize some of the main methods. We will compare their assumptions (most of which are slight variants of the common setting above), recovery types, performances, and robustness to noise. We focus primarily on the three approaches below.

- *Nuclear norm minimization:* This method is based on convexifying the intuitive but NP-hard approach of rank minimization given the observations. This method guaranteed to achieve full exact recovery under perhaps the most general assumption. However, the time complexity includes high powers of N and the calculation may be sensitive to noise.
- *Alternating projections:* This is based on another intuitive but NP-hard approach of fixing the rank then minimizing the difference with the observations. The algorithm switches between optimizing the column and row spaces, given the other, in alternating steps. Similar to the iterative method above, it runs well in practice but only provides Frobenius norm recovery.
- *Low-rank approximation with SVD:* One looks at the sample matrix A_Ω as a rescaled random perturbation of A, and attempts to recover A by taking a truncated SVD of A_Ω. Methods following this idea typically do not aim for exact recovery after the SVD, but an extra cleaning step with iterative optimization similar to the two above can achieve Frobenius norm recovery. Besides being robust to noise, these methods save time because the heavy computation (SVD in this case) only takes one step, as opposed to being repeated over many iterations in the three methods above. One may even achieve full approximate recovery with some extra but common assumptions. Our algorithm in Section 1.5 also belongs to this category.

1.3.1 Nuclear norm minimization

This approach starts from the intuitive idea that if A is mathematically recoverable, it has to be the matrix with the lowest rank agreeing with the observations at the revealed entries. Formally, one solves the following optimization problem:

$$\text{minimize} \quad \operatorname{rank} X \quad \text{subject to} \quad X_\Omega = A_\Omega. \tag{4}$$

Despite being theoretically guaranteed to achive full exact recovery, this approach is impractical since the problem is NP-hard, and all existing algorithms take doubly exponential time in terms of the dimensions of A [9]. To overcome the NP-hardness, Candes and Recht [6] proposed a heuristic version by replacing the rank with the nuclear norm of X. Formally, one solves the following:

$$\text{minimize} \quad \|X\|_* \quad \text{subject to} \quad X_\Omega = A_\Omega. \tag{5}$$

The authors of [6] were motivated by an idea from the *sparse signal recovery* problem in the field of *compressed sensing* [7, 12]. One has to find a hidden sparse vector $\mathbf{x} \in \mathbb{R}^n$ from a sample $\mathbf{y} = \mathbf{\Phi x} \in \mathbb{R}^k$, where $\Phi \in \mathbb{R}^{k \times n}$. Instead of finding the vector with the smallest support satisfying the sample, one can find the vector with the smallest ℓ_1-norm. The analogy is clear if one looks at the singular values of A as entries of a vector with an r-sized support.

The original paper [6] was shortly followed by Candes and Tao [5] with both improvements and trade-offs, and ultimately by Recht [22], who improved both previous results, proving that A is the unique solution to (5) given the sampling size bound

$$|\Omega| \geq C \max\{\mu_0, \mu_1^2\} r N \log^2 N, \tag{6}$$

for the coherence parameters μ_0 and μ_1 defined previously. With no special structure on UV^T, one replaces μ_1 with $\mu_0\sqrt{r}$ to obtain $C\mu_0^2 r^2 N \log^2 N$. This attains the optimal power of N while missing slightly from the optimal powers of r and $\log N$. With a structure on UV^T that allows μ_1 to be constant, the power of r is optimal.

The advantage of replacing the rank in Problem (4) with the nuclear norm is that Problem (5) is a convex program, which can be further translated into a semidefinite program [6, 5], solvable in polynomial time by a number of algorithms. However, convex optimization program usually runs slowly in practice. The survey [20] mentioned the interior point-based methods SDPT3 [24] and SeDuMi [23], which can take up to $O(|\Omega|^2 N^2)$ floating point operations (FLOPs) assuming (6), even if one takes advantage of the sparsity of A_Ω. An *iterative singular value thresholding* method aiming to solve a regularized version of nuclear norm minimization, trading exactness for performance, has been proposed [3].

Another potential issue is that the semidefinite program may be sensitive to noise [5]. If the sampling is noisy, the constraint $X_\Omega = A_\Omega$ may be corrupted, leading to an optimal solution no longer close to the original A. We will provide more details when discussing solutions to the noisy version in Section 1.4.

1.3.2 Modified alternating projections

The intuition behind this approach is to fix the rank, then attempt to match the observations at much as possible. If the hidden matrix has a known rank r, and is mathematically recoverable, it should be the only matrix of rank at most r satisfying the observation, so full exact recovery is theoretically guaranteed. Formally, one solves the optimization problem below:

$$\text{minimize} \quad \|(A - XY^T)_\Omega\|_F^2 \quad \text{over } X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{r\times n}. \tag{7}$$

This unfortunately, like (4), is NP-hard [10]. There have been many studies proposing variants of *alternating projections*, all of which involve the following basic idea: suppose one already obtains an approximate $X^{(l)}$ at iteration l, then $Y^{(l)}$ and $X^{(l+1)}$ are defined by

$$Y^{(l)} := \operatorname*{argmin}_{Y\in\mathbb{R}^{r\times n}} \|(A - X^{(l)}Y^T)_\Omega\|_F^2, \qquad X^{(l+1)} := \operatorname*{argmin}_{X\in\mathbb{R}^{r\times n}} \|(A - X(Y^{(l)})^T)_\Omega\|_F^2.$$

These methods tend to outperform nuclear norm minimization in practice [10], however there are few rigorous guarantees for recovery. The convergence and final output of the basic algorithm above also depends highly on the choice of $X^{(0)}$ [10].

Jain, Netrapalli and Sanghavi (2012) [17] developed one of the first few alternating projections variants for matrix completion with rigorous recovery guarantees. They proved that, under the same setting in Section 1.2 and the sample size condition

$$|\Omega| \geq C\mu_0 r^{4.5}\left(\frac{\sigma_1}{\sigma_r}\right)^4 N\log N \log\frac{r}{\varepsilon},$$

the AP algorithm in [17] recovers A within a relative Frobenius norm error ε in $O(|\Omega| r^2 \log(1/\varepsilon))$ time with high probability. Interestingly, their initialization step uses a rescaled SVD of a randomly chosen part of M_Ω, an idea similar to the SVD approach in the next section. Another notable feature is the *condition number* σ_1/σ_r, which can be large if σ_r is small, making the bound less efficient.

The condition number factor was reduced to quadratic by Hardt [14] and again by Hardt and Wooters [15] to logarithmic, at the cost of an increase in the powers of r, μ_0 and $\log N$.

1.3.3 Spectral algorithms

Compared to the previous approach, spectral algorithms do not aim to solve an optimization problem whose unique solution should be A. Instead, the intuition is to treat A_Ω, with a proper rescaling, as a random matrix whose expection is A, in other words, a random *unbiased perturbation* of A. Let us assume the independent sampling model with probability p, then $\tilde{A} := p^{-1}A_\Omega$ is such a rescaling. We can write $\tilde{A} = A + E$, where E has independent, mean-zero entries. From the perspective of perturbation theory, one views E as a type of *noise* that needs to be cleansed to recover the *signal* A. The "denoising" step computes the truncated SVD of $\tilde{A}$ at a proper index, which may not be exactly r. In some works, the cutoff point is the last index with a large enough singular value gap, so that one can apply a Davis-Kahan type result to show that the truncated SVDs of $\tilde{A}$ and A are very close. If one simply uses the r-rank SVD of $\tilde{A}$, typically one needs to assume σ_r is large enough, either explicitly or implicitly in the sample size bound, which may lead to the appreence of the *condition number* σ_1/σ_r. One does not expect the truncated SVD to remove all the noise, so the aim is approximate or norm recovery, rather than full exact recovery.

Keshavan, Montanari and Oh [19] combined a rank-r SVD step with two cleaning steps in the following algorithm:

1. *Trimming:* first zero out all columns in A_Ω with more than $2|\Omega|/m$ entries, then zero out all rows with more than $2|\Omega|/n$ entries, producing a matrix $\widetilde{A_\Omega}$.

2. *Low-rank approximation:* Compute the best rank-r approximation of $\widetilde{A_\Omega}$ via truncated SVD. Let $\mathsf{T}_r(\widetilde{A_\Omega}) = \tilde{U}_r\tilde{\Sigma}_r\tilde{V}_r^T$ be the output.

3. *Cleaning:* Solve for X, Y, S in the following optimization problem:

$$\text{minimize} \quad \left\| A_\Omega - (XSY^T)_\Omega \right\|_F^2 \quad \text{for} \quad X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}, S \in \mathbb{R}^{r\times r}, \tag{8}$$

   using a gradient descent variant [19], starting with $X_0 = \tilde{U}_r$, $Y_0 = \tilde{V}_r$ and S_0 be the $r \times r$ matrix minimizing the objective function above given X_0 and Y_0.

   Let (X_*, Y_*, S_*) be the solution. Output $X_*S_*Y_*^T$.

The last cleaning step resembles the optimization problem in alternating projections methods. The authors [19] showed that the algorithm returns an output close to A to arbitrary precision in the Frobenius norm, given enough iterations in the last step, provided the following sampling size condition:

$$|\Omega| \gtrsim \max\left\{ \mu_0\sqrt{mn}\left(\frac{\sigma_1}{\sigma_r}\right)^2 r\log N, \quad \max\{\mu_0,\mu_1\}^2 r\min\{m,n\}\left(\frac{\sigma_1}{\sigma_r}\right)^6 \right\}. \tag{9}$$

Besides the clear dependence on the condition number, the powers of r and $\log N$ are optimal by the coupon-collector limit. This shows that, at least in the case where the singular values of A are closely tied, SVD-based approaches can achieve arbitrarily accurate norm recovery with the optimal sampling size [5]. A potential drawback is that the authors of [19] did not include a full time complexity analysis of their gradient descent technique, only briefly mentioning that quadratic convergence is possible. This final cleaning step can be heavy depending on the desired margin of error, presenting a potential challenge.

In a recent paper, Bhardwaj and Vu [2] proposed the following simple spectral algorithm to recover a square $N \times N$ matrix A [2, Algorithm 22], without the cleaning step:

1. Let $\tilde{A} := p^{-1}A_\Omega$ and compute the SVD: $\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^T = \sum_{i=1}^{m\wedge n}\tilde{\sigma}_i\tilde{u}_i\tilde{v}_i^T$.

2. Let $\tilde{s}$ be the last index such that $\tilde{\sigma}_i \geq \frac{N}{8r\mu^2}$, where $\mu := N \max\{\|U\|_\infty^2, \|V\|_\infty^2\}$ is known.

3. Let $\hat{A} := \sum_{i=1}^{\tilde{s}} \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^T$ and return $\hat{A}$.

Regarding the common setting, the authors assume that r and μ are both $O(1)$ (note that their μ is larger than μ_0), and the sampling density condition $p \geq N^{-1} \log^{4.03} N$.

They showed that $\|\hat{A} - A\| \leq \log^{-c} N$ for a constant $c > 0$, achieving full approximate recovery, with probability $1 - O(N^{-1})$, provided the following extra conditions hold:

- *Bounded entries:* $\|A\|_\infty \leq K$ for a known constant K.

- *Large signal-to-noise ratio:* $\sigma_s \geq Cr^3 K \sqrt{N/p} \log^{2.01} N$, where s is the last index such that $\sigma_s \geq \frac{N}{16r\mu^2}$ and C is a sufficiently large constant.

  The name "signal-to-noise" comes from the fact that typically one has $\|E\| = O(\sqrt{N/p})$.

- *Large gaps:* $\min_{i \in [s]}(\sigma_i - \sigma_{i+1}) \geq Cp^{-1} \log N$.

When A has integer entries, rounding the entries of $\hat{A}$ to the nearest integer will recover A exactly.

The most important advantage of the above algorithm is that it runs the SVD only once, on a sparse matrix with $|\Omega|$ nonzero entries, truncated at rank at most r, thus taking only $O(|\Omega| r)$ FLOPs [27], which does not grow with $1/\varepsilon$ as opposed to the previous two iterative methods. The implementation is also much simpler, thus having fewer potential performance bottlenecks in real-life runs. There are several shortcomings, including

1. The sampling density p, the rank r of A and the infinity norms of the singular vectors are assumed to be known, which may be hidden in real-life data.

2. There are signal-to-noise and gap assumptions that are rather restrictive, as opposed to a simple sampling size condition in the guarantees of the optimization-based methods above. In fact, while the sample size bound does not directly involve the condition number like the alternating projections papers [14, 15], one can rewrite the signal-to-noise condition as

$$p \geq C^2 r^6 N \log^{-4.02} N \left(\frac{K\sqrt{N}}{\sigma_s} \right)^2 .$$

   Observe that the last factor on the right-hand side satisfies

$$\frac{K\sqrt{N}}{\sigma_s} \geq \frac{\|A\|_F}{\sigma_s} \geq \frac{\sigma_1}{\sigma_s},$$

   thus the condition number at index s (the cutoff point for the SVD) should be small for the bound to be effective.

3. The incoherence parameter μ is larger than μ_0 from the common setting, meaning the sampling density condition is slightly worse. The polylog(N) factor is also slightly suboptimal to the coupon-collector limit.

One can view the SVD-based methods as trading off guaranteed full exact recovery for significant speedup and potentially simpler implementation in practice, compared to nuclear norm minimization and alternating projections. Another advantage of the former lies in the case where the observations are corrupted by noise that are typically random. We will discuss this in the next section.

1.4 The noisy version

In practice, data are often not perfectly low-rank. Even if only a few factors determine the entire dataset, some form of additive noise may enter through natural sampling errors, hidden factors not captured by the data collector, or artificial means to achieve some desired effect. For example, some numerical computations artificially add noise to the data for stability, or to preserve privacy [13]. Data that are *approximately low-rank*, namely there is a sharp drop in the singular values at some index $r \in \min\{m, n\}$, can be modeled as $A_\xi = A + \xi$, where A is a low-rank matrix as before and ξ is a noise matrix. The approximate low-rank effect is captured if one assumes $\|\xi\|_{op} < \sigma_r$, but we do not need it for the general case. We can denote the sample matrix as $A_{\Omega,\xi}$, given by

$$(A_{\Omega,\xi})_{kl} = A_{kl} + \xi_{kl} \text{ for } (k, l) \in \Omega, \text{ and 0 otherwise.}$$

For nuclear norm minimization, Candes and Plan [4] adapted to the noisy situation by relaxing the constraint on the observations, leading to the following problem:

$$\text{minimize} \quad \|X\|_* \quad \text{subject to} \quad \|X_\Omega - M_\Omega\|_F \leq \delta, \tag{10}$$

where δ is a known upper bound on $\|\xi_\Omega\|_F$. The authors showed that, under the same sample size condition in [22], with probability $1 - o(1)$, the optimal solution $\hat{A}$ is close to A in the Frobenius norm, namely

$$\frac{1}{\sqrt{mn}}\|\hat{A} - A\|_F \lesssim \|\xi_\Omega\|_F \sqrt{\frac{\min\{m, n\}}{|\Omega|}}. \tag{11}$$

This result is deterministic and holds for all noise matrices ξ. If one would like the $\|\hat{A} - A\| \leq \varepsilon\sqrt{mn}$ for a chosen ε, this implicitly gives an extra sample size condition:

$$|\Omega| \geq \frac{\|\xi\|_F^2 \min\{m, n\}}{\varepsilon^2}, \tag{12}$$

which grows quadratically with $1/\varepsilon$.

For alternating projections, extensions for the noisy case was done by Hardt [14] and Hardt and Wooters [15]. They assumed that A is an approximately low-rank matrix, with a large singular value gap $\delta_r = \sigma_r - \sigma_{r+1}$. Partition the left singular subspace at index r into two subspaces and let A_r and $A_{>r}$ be the projections of A onto them. They treated $A_{>r}$ as noise and aimed to recover A_r from the sample A_Ω. The latter result [15] showed that $\|\hat{A} - A\|_{op} \leq \varepsilon\sigma_1 + (1 + o(1))\sigma_{r+1}$ (an operator norm recovery) with the following sample size lower bound:

$$|\Omega| \geq C\left(\frac{r\sigma_r}{\delta_r}\right)^c \mu_0^2 \log\frac{\sigma_1}{\sigma_r} N\left(\log^2\frac{N}{\varepsilon} + \frac{\|A_{>r}\|_F^2}{\varepsilon^2\|A\|_F^2}\right). \tag{13}$$

A notable feature, besides the condition number, is the factor σ_r/δ_r, which shrinks closer to 1 the larger the gap is, meaning fewer observations are required the larger the signal is compared to the noise, and vice versa. However, this noise model is highly structured (it is orthogonal to A_r, the matrix to be recovered) and is not the focus of our paper.

Keshavan, Montanari and Oh [18] also extended their result with the spectral algorithm for the noisy version, using the model $A_{\Omega,\xi}$ mentioned above, with no assumptions on ξ. They proved that with the same sample size condition as (9), the output satisfies w.h.p.

$$\|\hat{A} - A\|_F \lesssim \left(\frac{\sigma_1}{\sigma_r}\right)^2 \frac{r^{1/2}mn}{|\Omega|}\|\xi_\Omega\|_{op}. \tag{14}$$

This is better in general than the bound (11) in [4], unless the condition number is very large. If one would like $\|\hat{A} - A\|_F \leq \varepsilon\sqrt{mn}$, this translates to another sample size condition:

$$|\Omega| \geq \frac{\sigma^2 r N}{\varepsilon^2} \left(\frac{\sigma_1}{\sigma_r}\right)^2. \tag{15}$$

Again, the inverse quadratic relationship between the number of samples required and the magnitude of the noise appears. One should thus expect this is the case for most noisy situations.

Bhardwaj and Vu [2] did not consider the noisy case, but it is possible to adapt their technique to handle random, independent noise. In fact, our algorithm, which we describe in the next section, is partially based on their idea, and works well with this noise model. In our chosen setting, which largely overlaps with [2], our algorithm is guaranteed to achieve full approximate recovery with arbitrary precision in only one SVD step, with a lower sample size requirement than [2], without involving the condition number.

1.5 A simple spectral algorithm

Let us clarify the setting in this paper. We combine everything in the common setting (Section 1.2) with two extra assumptions about A and one about the noise ξ.

Setting 1.2 (Matrix completion with noise)**.** Given the objects A, Ω, $A_{\Omega,\xi}$, and ξ, we assume:

1. *Known bound on entries:* We assume $\|A\|_\infty \leq K_A$ for some known constant K_A. This is the case for many applications, such as the Netflix challenge.

2. *Known bound on rank:* We do not assume the knowledge of the rank r, but assume that we know some small value $r_{\max}$ such as $r = \operatorname{rank} A \leq r_{\max}$ (upper bound).

3. *Independent, bounded, centered noise:* ξ has independent entries satisfying $\mathbf{E}\left[\xi_{ij}\right] = 0$ and $\mathbf{E}\left[|\xi_{ij}|^l\right] \leq K_\xi^l$ for all $l \in \mathbb{N}$ and $i \in [m]$, $j \in [n]$.

The quantitative parameters of the model are r, $r_{\max}$, p, K_A, K_ξ and κ_ξ.

The second assumption is slightly more general than the known rank assumptions common in papers using the spectral method [2, 19, 18] and alternating projections [17, 14, 15]. In some papers, e.g. [2], it is not trivial to relax the assumption to $r_{\max}$, since their algorithm takes r to determine the cutoff index for the SVD. In [19, 18], if one knows only $r_{\max}$, one needs to repeat the computation $r_{\max}$ times to get the best output.

We propose the following algorithm for matrix completion under the assumptions above. For simplicity, let us first consider the case when the entries of A are integer.

Algorithm 1.3 (Matrix completion)**.** Given an $m \times n$ matrix $A_{\Omega,\xi}$ and the set $\Omega \subset [m] \times [n]$, return an estimate $\hat{A}$ of A. One proceeds in the steps below:

1. *Sampling density estimation:* Let $\hat{p} := (mn)^{-1}|\Omega|$.

2. *Rescaling:* Let $\hat{A} := \hat{p}^{-1} A_{\Omega,\xi}$.

3. *Low-rank approximation:* Compute the SVD of the sample matrix: $\hat{A} = \sum_i \hat{\sigma}_i \hat{u}_i \hat{v}_i^T$. Take the largest s such that $\hat{\sigma}_s - \hat{\sigma}_{s+1} \geq 20(K_A + K_\xi)\sqrt{r_{\max}(m+n)/\hat{p}}$, then let $\hat{A}_s := \sum_{i \leq s} \hat{\sigma}_i \hat{u}_i \hat{v}_i^T$.

4. *Cleaning and output:* Round each entry of $\hat{A}_s$ to the nearest integer. Return $\hat{A}_s$.

Theorem 1.4. *There is a universal constant $C > 0$ such that the following holds. Let $N = m + n$ and let $\varepsilon > 0$ be an arbitrary error margin. Under the model 1.2, assume the following:*

- Signal-to-noise: $\sigma_1 \geq 100 r K_{A,\xi} \sqrt{\frac{r_{\max} N}{p}}$ *for* $K_{A,\xi} := K_A + K_\xi$ *and* $N := m + n$.
- Sampling density:

$$p \geq C \left(\tfrac{1}{m} + \tfrac{1}{n}\right) \max \left\{ \log^{10} N,\ \frac{r^4 r_{\max} \mu^2 K_{A,\xi}^2}{\varepsilon^2}, \frac{r^3 K_{A,\xi}^2}{\varepsilon^2} \left(1 + \frac{\mu^2}{\log^2 N}\right) \left(1 + \frac{r^3 \log N}{N}\right) \log^6 N \right\}. \quad (16)$$

Then with probability $1 - O(N^{-1})$, the first three steps of Algorithm 1.3 recovers every entry of A within an absolute error ε. Consequently, if one assumes Eq. (16) for $\varepsilon = 0.4$ and A has integer entries, then the cleaning step recovers A exactly.

Remark 1.5. The condition (16) looks very complicated, but in the common case where A has constant rank, constantly bounded entries, and uniformly random singular vectors, we have $\mu = O(\log N)$, $K_A, K_\xi = O(1)$ and $r_{\max} = O(r) = O(1)$, it reduces to

$$p \geq C \max \left\{ \log^4 N,\ \varepsilon^{-2} \right\} \left(m^{-1} + n^{-1}\right) \log^6 N, \quad (17)$$

where $N = m + n$.

Remark 1.6. Suppose the situation in the remark above for the parameters other than N and ε holds. If one considers the general case when the entries of A are multiples of $\varepsilon > 0$, the number of sample needed for exact recovery after rounding stays at $O(N \log^{10} N)$ until $\varepsilon < \log^{-2} N$, then it grows quadratically with $1/\varepsilon$. In fact, we see that this quadratic grow factor is also present in the papers [4, 2, 19, 18] when we compare our results with theirs.

In the next section, we will make further comparisons with the papers of other methods.

1.5.1 Summary of the techniques and main contributions

We give a brief summary of the strongest results of the approaches discussed so far, and highlight some improvements we have over them in our setting. Note that we consider the two lines of results in the spectral methods section separately, since they have substantially different goals for recovery and assumptions. We omit the comparison of the powers of r and $\log N$ in the sample size conditions, since they are suboptimal in ours and most of these papers.

Table 1 below summarizes the comparison. Since the aim is comparing our work with the others, we simplified many details. For the text to fit in the table, we use the following shorthands:

- AltProj stands for alternating projections.
- Spectral stands for spectral methods.
- NuclMin stands for nuclear norm minimization.
- $\omega_{1/\varepsilon}$ stands for an asymptotic that grows to infinity as $\varepsilon \to 0$.
- The recovery type ε-Frobenius means the output has to satisfies $\|\hat{A} - A\|_F \leq \varepsilon$.
- The recovery type ε-Infinity means the output has to satisfies $\|\hat{A} - A\|_\infty \leq \varepsilon$, or equivalently, full approximate recovery with error margin ε.

Feature/Method	NuclMin [22, 4]	AltProj [15]	Spectral [19, 18]	Spectral [2]	Spectral (ours)
Recovery type	Full exact	ε-Frobenius	ε-Frobenius with noise	ε-Infinity	ε-Infinity
Time complexity asymptotics	$\|\Omega\|^2 N^2$	$\|\Omega\|^2 r\omega_{1/\varepsilon}$	$\|\Omega\| r + N r\omega_{1/\varepsilon}$	$\|\Omega\| r$	$\|\Omega\| r$
Recovery w.h.p.	Guaranteed	Guaranteed	Guaranteed	Guaranteed	Guaranteed
Condition number factor in $\|\Omega\|$ lower bound	None	$\log \frac{\sigma_1}{\sigma_r}$	$\left(\frac{\sigma_1}{\sigma_r}\right)^6$	Implicitly $\left(\frac{\sigma_1}{\sigma_s}\right)^2$ at cutoff s	None
Needs bounded entries	No	No	No	Yes	Yes
Needs large σ_r to be effective (noisy case)	No	Yes	Yes	Yes	No
Needs large σ_1 to be effective	N/A	N/A	N/A	N/A	Yes

Table 1: Summary of MC techniques

In summary, at the cost of three extra assumptions, namely sufficiently large $\|A\|_{op}$, A has bounded entries, and the noise is independently random, which are satisfied in many applications, our algorithm gains the following advantages:

- We can recover every entry of A to arbitrary precision, which can be desirable for recommendation systems, as opposed to the Frobenius norm recovery in [4, 15]. The recovery in [19] is possibly full approximate, but it is also Frobenius in their noisy adaptation [18].

- The execution is fast, as SVD is the only subroutine one needs to use. The implementation is also very simple. Observe that the running time grows linearly with $|\Omega|$, so it also grows with $1/\varepsilon^2$ (where ε is the desired error margin). This inverse quadratic growth factor is also present in the noisy adaptations of the other approaches [4, 15, 18]; as seen from (12), (13), and (15).

  One slight drawback of our algorithm is that this factor is also present in our sample density bound in the noiseless case, whereas this is not the case for the other approaches [6, 15, 19]. These method use their optimization steps to get arbitrarily close to the hidden matrix when there is no noise, without increasing the sampling size. We do not have such step, opting to trade off some accuracy for simplicity.

  This is not a serious issue in the noiseless case, since for constant ε (in fact, for $\varepsilon \gg \log^{-2} N$), the polylog($N$) factor in (17) dominates ε^{-2}, so it does not matter.

- The sampling size/density bounds in [19, 18, 14, 15] also grow when σ_r becomes small (keeping everything else fixed). This is implicit in [2], which requires a large singular value and gaps up to the cutoff point for the SVD step.

  In contrast, our bound is not affected by a small σ_r. We require only that σ_1 be larger than $\sqrt{(m+n)/p}$ times a factor depending only on K_A, K_ξ and r. If one assume that every of A is in the range $[1, K_A]$, which is the case for many real-life datasets such as the Netflix data, this is automatic as $\sigma_1 \geq r^{-1/2}\|A\|_F \geq \sqrt{mn/r}$, which should be much larger than $\sqrt{(m+n)/p}$ if the density requirement (31) is satisfied.

- Our coherence parameter is μ_0 from the common setting, which is better than the parameter μ from Bhardwaj and Vu [2], and is comparable to the other papers discussed here.

- The sample size lower bound does not depend on the condition number, as oppose to [3, 15, 19], and is optimal up to a polylogarithmic term. If we take a closer look, the fact that one needs the condition number to be small so the algorithm can be effective is somewhat counter intuitive. Notice that the low rank perturbation will omit singular values below some threshold, so the smaller these values are, the better for us. Thus, heuristically at least, the large condition number should help the algorithm (as it suggests that what we omit is small), rather than ruin it.

- Our algorithm does not require the knowledge of the rank r; we only need to know an upper bound $r_{\max}$. In contrast, many algorithms require knowing r in advance. This is the case for the algorithm in [2], which needs to cutoff the SVD at a precise point only determinable from the rank. The algorithms in [14, 15, 19, 18] also involve the rank, it is not clear whether its exact value is strictly required.

  This is also an interesting point, mathematically. An argument authors usually make here is the following: if we do not know the rank exactly, but know that it is at most $r_{\max}$, then we just run the algorithm $r_{\max}$ times, with $r = 1, 2, \dots, r_{max}$ as input, respectively. This blows up the running time by a factor $r_{\max}$, which is negligible if $r_{\max}$ is small. From the running time point of view, this is correct. However, there is a subtlety here, as it is not clear that in that case which output must be chosen. There could be two different outputs which agree on the observed entries. We have not seen any rigorous analysis which rules out this scenario.

In the next section, we will describe the intuition behind the low-rank approximation approach above by looking at the problem from the perspective of matrix perturbation theory. We will introduce the main technical theorem, which strengthens and extends the classical Davis-Kahan theorem in matrix perturbation theory and use it to prove Theorem 1.4.

2 Proof of guarantees for recovery

2.1 The matrix perturbation perspective

Consider the sampling and noise models in Setting 1.2. As discussed in [2], the matrix $A_{\Omega,\xi}/p$ is random matrix whose expectation is A (they only mentioned the pure case, but this holds with centered noise too). One can treat $A_{\Omega,\xi}/p$ as an *unbiased* perturbed version of A, with the perturbation $E := A_{\Omega,\xi}/p - A = A_\Omega/p - A + \xi/p$, which is also a random matrix with independent, mean-zero entries. Let $\tilde{A} := A_{\Omega,\xi}/p = A + E$ for convenience. Since the goal is to recover A, which is a low-rank matrix, it makes sense to consider the best low-rank approximations of $\tilde{A}$, motivating the spectral approach. To bound the error, one needs tools from matrix perturbation theory. Let us introduce notations in the perturbation model, putting aside the matrix completion context:

Setting 2.1 (Matrix perturbation). Consider a fixed $m \times n$ matrix A with SVD

$$A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T, \qquad \text{where } \sigma_1 \geq \sigma_2 \geq \dots \sigma_r.$$

Consider a $m \times n$ matrix E, which can be deterministic or random, which we called the *perturbation matrix*. Let $\tilde{A} = A + E$ be the *perturbed matrix* with the following SVD:

$$\tilde{A} = \tilde{U}\Sigma\tilde{V}^T = \sum_{i=1}^{m \wedge n} \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^T \quad \text{where } \tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \tilde{\sigma}_r.$$

Define the following terms related to A and $\tilde{A}$:

1. For $k \in [r]$, $\delta_k := \sigma_k - \sigma_{k+1}$, using $\sigma_{r+1} = 0$, and let $\Delta_k := \delta_k \wedge \delta_{k-1}$.

2. For $S \subset [r]$, let $\sigma_S := \min\{\sigma_i : i \in S\}$ and $\Delta_S := \min\{|\sigma_i - \sigma_j| : i \in S, j \in S^c\}$.

3. For $S \subset [r]$, define the following matrices:

$$V_S := \left[v_i\right]_{i \in S}, \quad U_S := \left[u_i\right]_{i \in S}, \quad A_S := \sum_{i \in S} \sigma_i u_i v_i^T.$$

   When $S = [s]$ for some $s \in [r]$, we also use V_s, U_s, A_s respectively to denote the three above.

Define analogous notations $\tilde{\delta}_k$, $\tilde{\Delta}_k$, $\tilde{\sigma}_S$, $\tilde{\Delta}_S$, $\tilde{V}_S$, $\tilde{U}_S$, and $\tilde{A}_S$ for $\tilde{A}$.

One of the most well-known results in perturbation theory is the **Davis-Kahan** $\sin\Theta$ **theorem** [11], which bounds the change in eigenspace projections by the ratio between the perturbation and the eigenvalue gap. The extension for singular subspaces, proven by Wedin [29], states that:

$$\|\tilde{U}_s\tilde{U}_s^T - U_sU_s^T\| \vee \|\tilde{V}_s\tilde{V}_s^T - V_sV_s^T\| \leq \frac{C\|E\|}{\delta_s} \quad \text{for a universal constant } C. \tag{18}$$

There are three challenges if one wants to apply this theorem:

1. The inequality above only concerns the change in the singular subspace projections, while the change in the low-rank approximation $\tilde{U}_s\tilde{\Sigma}_s\tilde{V}_s^T$ is needed.

2. The bound on the right-hand side requires the spectral gap-to-noise ratio at index s to be large to be useful, which is a strong assumption.

3. The left-hand side is the operator norm, while an infinity norm bound is needed for exact recovery after rounding.

A key observation is that, similarly to the Frobenius norm bound in [4], Eq. (18) works for all perturbation matrices E. Per the discussion in [26], the worst case (equality) only happens when there are special interactions between E and A. A series of papers by Vu and coauthors [21, 26] exploited the improbability of such interactions when E is random and A has low rank, and improved the bound significantly.

O'Rourke, Vu and Wang [21] proved the following:

$$\|\tilde{V}_s\tilde{V}_s^T - V_sV_s^T\| \leq C\sqrt{s}\left(\frac{\|E\|}{\sigma_s} + \frac{\sqrt{r}\|U^TEV\|_\infty}{\delta_s} + \frac{\|E\|^2}{\delta_s\sigma_s}\right),$$

with high probability, effectively turning the *noise-to-gap* on the right-hand side of Eq. (18) into the *noise-to-signal ratio*, which can be much smaller than the former in many cases.

P. Tran and Vu then [26] improved the third term, at the cost of an extra factor of $\sqrt{r}$, which does not matter when A has constant rank. They showed that when

$$\frac{\|E\|}{\sigma_s} \vee \frac{2r\|U^T EV\|_\infty}{\delta_s} \vee \frac{\sqrt{2r}\|E\|}{\sqrt{\delta_s \sigma_s}} \leq \frac{1}{8},$$

then

$$\|\tilde{V}_s\tilde{V}_s^T - V_sV_s^T\| \leq Cr\left(\frac{\|E\|}{\sigma_s} + \frac{2r\|U^T EV\|_\infty}{\delta_s} + \frac{2ry}{\delta_s\sigma_s}.\right),$$

replacing the term $\|E\|^2$ in [21] with the smaller $y := \frac{1}{2}\max_{i\neq j}(|u_i^T EE^T u_j| + |v_i^T E^T Ev_j|)$. This quantity, while not smaller than $\|E\|^2$ in some random matrix models, can be significantly smaller in many cases, notably when E is *regular* [26], meaning there is a common $\overline{\sigma}$ such that:

$$\overline{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m} \mathbf{Var}\left[E_{ij}\right] = \frac{1}{n}\sum_{j=1}^{n} \mathbf{Var}\left[E_{ij}\right] \quad \text{for all } i \in [m],\ j \in [n]. \tag{19}$$

2.2 Contour Integral with Combinatorial expansion

Our approach here is strongly motivated by the developments in [26]. The main idea is to write the difference of two matrix operators as a contour integral, and then use a combinatorial expansion to split this integral into many subsums, each of which can be treated using tools from complex analysis, linear algebra, and combinatorics. In [26], the authors used this idea to obtain new perturbation bounds in matrix spectral (operator) norm; see Section 3 of [26] for a detailed discussion.

In this paper, we adapt and further develop this method to obtain perturbation bound in the infinity norm. As an infinity norm estimate is much usually much harder than an operator norm, the analysis gets significantly more involved, and we need to find several new ideas to deal with the new difficulties.

Our main result is the following:

Theorem 2.2. *Consider the objects in Setting 2.1. Suppose some positive numbers $\mu_0(U)$, $\mu_1(U)$, $\mu_0(V)$, $\mu_1(V)$ and $\mathcal{K}$ satisfy $\|E\| \leq \mathcal{K}$ and the below, uniformly over $0 \leq a \leq 10\log(m+n)$:*

$$\begin{aligned} &\left\|(EE^T)^a U\right\|_{2,\infty} \leq \mu_0(U)\mathcal{K}^{2a}\sqrt{r} \qquad &&\left\|(E^TE)^a E^T U\right\|_{2,\infty} \leq \mu_1(U)\mathcal{K}^{2a+1}\sqrt{r} \\ &\left\|(E^TE)^a V\right\|_{2,\infty} \leq \mu_0(V)\mathcal{K}^{2a}\sqrt{r}, \qquad &&\left\|(EE^T)^a EV\right\|_{2,\infty} \leq \mu_1(V)\mathcal{K}^{2a+1}\sqrt{r}. \end{aligned} \tag{20}$$

Consider an arbitrary subset $S \subset [r]$. Suppose that S satisfies

$$\frac{\mathcal{K}}{\sigma_S} \vee \frac{2r\|U^T EV\|_\infty}{\Delta_S} \vee \frac{\sqrt{2r}\|E\|}{\sqrt{\Delta_S\lambda_S}} \leq \frac{1}{8}, \tag{21}$$

Then there is a universal constant C such that

$$\left\|\tilde{V}_S\tilde{V}_S^T - V_SV_S^T\right\|_\infty \leq C\mu_{UV}^2 r\left(\frac{\mathcal{K}}{\sigma_S} + \frac{2r\|U^T EV\|_\infty}{\Delta_S} + \frac{2ry}{\Delta_S\sigma_S}\right) + \frac{1}{m+n}, \tag{22}$$

$$\left\|\tilde{V}_S\tilde{V}_S^T - V_SV_S^T\right\|_{2,\infty} \leq C\mu_{UV} r\left(\frac{\mathcal{K}}{\sigma_S} + \frac{2r\|U^T EV\|_\infty}{\Delta_S} + \frac{2ry}{\Delta_S\sigma_S}\right) + \frac{1}{m+n}, \tag{23}$$

where

$$\mu_{UV} := \mu_1(U) + \mu_0(V), \quad \mu_{VU} := \mu_1(V) + \mu_0(U), \quad y := \frac{1}{2}\max_{i\neq j}(|u_i^T EE^T u_j| + |v_i^T E^T Ev_j|).$$

Analogous bounds for U and $\tilde{U}$ hold, with U and V swapped. When $S = [s]$ for some $s \in [r]$, we also have

$$\|\tilde{A}_s - A_s\|_\infty \leq C\mu_{UV}\mu_{VU} r\sigma_s \left(\frac{\mathcal{K}}{\sigma_s} + \frac{2r\|U^T E V\|_\infty}{\delta_s} + \frac{2ry}{\delta_s \sigma_s}\right) + \frac{1}{m+n}. \tag{24}$$

The theorem above is deterministic and works for all perturbation matrices E. When E is random, we can plug in high-probability estimates for the terms therein to get a "random" version. Let us interpret the meanings of the terms to get a sense of which expressions to plug in.

The term $\mathcal{K}$ plays the role of $\|E\|$, but defining it as an upper bound offers more flexibility for Eq. (20). Various works on random matrix norms [1, 28] give the estimate $\mathcal{K} \sim \kappa\sqrt{m+n}$.

The terms $\mu_i(U)$ and $\mu_i(V)$ are the *delocalization coefficients*, playing the role of the coherence parameters introduced earlier. , which can be very small when $\|U\|_{2,\infty}$ and $\|V\|_{2,\infty}$ are small, improving substantially from the trivial choices $\mu_i(U) = \mu_i(V) = 1$. Our analysis later will give the estimates

$$\mu_0(U) \lesssim \log(m+n)\frac{\|U\|_{2,\infty}}{\sqrt{r}} \quad \text{and} \quad \mu_1(U) \lesssim \frac{\log^{3/2}(m+n)}{\sqrt{m+n}} + \frac{K\log^3(m+n)}{\sqrt{m+n}} \cdot \frac{\|U\|_{2,\infty}}{\sqrt{r}},$$

and symmetrically for $\mu_i(V)$. These terms differ from the coherence parameters by some polylogarithmic factors.

In fact, replacing $\mathcal{K}$ with $\|E\|$ in the three bounds recovers the term

$$\frac{\|E\|}{\sigma_S} + \frac{2r\|U^T E V\|_\infty}{\Delta_S} + \frac{2ry}{\Delta_S \sigma_S}$$

in the Davis-Kahan bound of [26]. In philosophy, our bounds look like

$$\begin{aligned}
\left\|\tilde{V}_S\tilde{V}_S^T - V_S V_S^T\right\|_\infty &\lesssim \frac{\|V\|_{2,\infty}^2}{r}\left\|\tilde{V}_S\tilde{V}_S^T - V_S V_S^T\right\| \\
\left\|\tilde{V}_S\tilde{V}_S^T - V_S V_S^T\right\|_{2,\infty} &\lesssim \frac{\|V\|_{2,\infty}}{\sqrt{r}}\left\|\tilde{V}_S\tilde{V}_S^T - V_S V_S^T\right\| \\
\left\|\tilde{A}_s - A_s\right\|_\infty &\lesssim \frac{\|U\|_{2,\infty}}{\sqrt{r}}\frac{\|V\|_{2,\infty}}{\sqrt{r}}\left\|\tilde{V}_s\tilde{V}_s^T - V_s V_s^T\right\|,
\end{aligned} \tag{25}$$

with some additional polylogarithmic factors.

The term $\|U^T E V\|_\infty$ is the maximum among sums of independent random variables, whose sizes are concentrated around $\kappa\sqrt{\log(m+n)}$ by the Hoeffding or Chernoff bound [16, 8].

The term y has been analyzed in [26] and mentioned above. We use the trivial upper bound $\|E\|^2 \sim \kappa^2(m+n)$, which is enough to prove Theorem 1.4.

Plugging in the expressions above, we get the following "random" version:

Theorem 2.3. *Consider the objects in Setting 2.1. Let $\varepsilon \in (0,1)$ be arbitrary Suppose E is a random $m \times n$ matrix with independent entries satisfying:*

$$\mathbf{E}\left[E_{ij}\right] = 0, \quad \mathbf{E}\left[|E_{ij}|^2\right] \leq \kappa^2, \quad \mathbf{E}\left[|E_{ij}|^p\right] \leq K^{p-2}\kappa^p \quad \textit{for all} \;\; p \in \mathbb{N}_{\geq 2}, \tag{26}$$

where K and κ are parameters. Define

$$\mu_{UV} := \frac{K\|U\|_{2,\infty}\log^3(m+n)}{\sqrt{r(m+n)}} + \frac{\|V\|_{2,\infty}\log(m+n)}{\sqrt{r}} + \frac{\log^{3/2}(m+n)}{\sqrt{m+n}},$$

and symmetrically for μ_{VU}. For an arbitrary subset $S \subset [r]$, suppose

$$\frac{\kappa\sqrt{m+n}}{\sigma_S} \vee \frac{r\kappa\sqrt{\log(m+n)}}{\Delta_S} \vee \frac{\kappa\sqrt{r(m+n)}}{\sqrt{\Delta_S \sigma_S}} \leq \frac{1}{16}. \tag{27}$$

There are universal constants c and C such that: If $K \leq c(m+n)^{1/2}\log^{-5}(m+n)$, then with probability at least $1 - O((m+n)^{-1})$,

$$\left\|\tilde{V}_S\tilde{V}_S^T - V_S V_S^T\right\|_\infty \leq C\mu_{UV}^2 r\left(\frac{\kappa\sqrt{m+n}}{\sigma_S} + \frac{r\kappa\sqrt{\log(m+n)}}{\Delta_S} + \frac{2r\kappa^2(m+n)}{\Delta_S\sigma_S}\right) + \frac{1}{m+n}, \tag{28}$$

$$\left\|\tilde{V}_S\tilde{V}_S^T - V_S V_S^T\right\|_{2,\infty} \leq C\mu_{UV} r\left(\frac{\kappa\sqrt{m+n}}{\sigma_S} + \frac{r\kappa\sqrt{\log(m+n)}}{\Delta_S} + \frac{2r\kappa^2(m+n)}{\Delta_S\sigma_S}\right) + \frac{1}{m+n}. \tag{29}$$

Analogous bounds for U and $\tilde{U}$ hold, with μ_{VU} replacing μ_{UV}.

When $S = [s]$ for some $s \in [r]$, we also have

$$\|\tilde{A}_s - A_s\|_\infty \leq C\mu_{UV}\mu_{VU} r\sigma_s\left(\frac{\kappa\sqrt{m+n}}{\sigma_s} + \frac{r\kappa\sqrt{\log(m+n)}}{\delta_s} + \frac{2r\kappa^2(m+n)}{\delta_s\sigma_s}\right) + \frac{1}{m+n}. \tag{30}$$

Furthermore, for each $\varepsilon > 0$, if R_2^2 is replaced with

$$\frac{r}{\Delta_S\sigma_S}\inf\left\{t : \mathbf{P}\left(\max_{i\neq j}(|v_i E^T E v_j| + |u_i E E^T u_j|) \leq 2t\right) \geq 1 - \varepsilon\right\},$$

then all three bounds above hold with probability at least $1 - \varepsilon - O((m+n)^{-1})$.

2.3 Proof of Theorem 1.4

Let us use Eq. (30) to prove Theorem 1.4, asserting the correctness w.h.p. of Algorithm 1.3. For convenience, we reuse the shorthand $N = m + n$ and define $K_{A,\xi} := K_A + K_\xi$.

Proof of Theorem 1.4. As discussed, the scaled version of $A_{\Omega,\xi}$ is a mean-zero perturbation of A. Let $\tilde{A} := p^{-1}A_{\Omega,\xi}$ and $E := \tilde{A} - A$. Then E is a random matrix with independent entries where:

$$E_{ij} = \begin{cases} A_{ij}\left(1/p - 1\right) + \xi_{ij}/p & \text{with prob. } p \\ -A_{ij} & \text{with prob. } 1-p. \end{cases}$$

For each i, j, we have $\mathbf{E}\left[E_{ij}\right] = 0$ and for each $l \geq 2$,

$$\begin{aligned}
\mathbf{E}\left[|E_{ij}|^l\right] &= \frac{\mathbf{E}\left[|A_{ij}(1-p) + \xi_{ij}|^l\right]}{p^{l-1}} + (1-p)|A_{ij}|^l \leq \frac{(K_A(1-p) + K_\xi)^l}{p^{l-1}} + (1-p)K_A^l \\
&\leq \frac{1}{p^{l-1}}\left(\sum_{k=0}^{l-1}\binom{l}{k}K_A^k(1-p)^k K_\xi^{l-k} + K_A^l(1-p)^l + p^{l-1}(1-p)K_A^l\right) \\
&\leq \frac{1}{p^{l-1}}\left(\sum_{k=0}^{l-1}\binom{l}{k}K_A^k K_\xi^{l-k} + K_A^l\right) = \frac{(K_A + K_\xi)^l}{p^{l-1}} = \frac{K_{A,\xi}^l}{p^{l-1}}.
\end{aligned}$$

Therefore E fits the noise model (26) with the parameters $K = p^{-1/2}$ and $\kappa = K_{A,\xi}K = p^{-1/2}K_{A,\xi}$. Let $C_2 = 1/c$ for the constant c in Theorem 2.3. We rewrite the assumptions below:

1. *Signal-to-noise:* $\sigma_1 \geq 100 r\kappa\sqrt{r_{\max} N}$.
2. *Sampling density:* this is equivalent to the conjunction of three conditions:

$$p \geq \frac{Cr^4 r_{\max}\mu^2 K_{A,\xi}^2}{\varepsilon^2}\left(\frac{1}{m}+\frac{1}{n}\right), \tag{31}$$

$$p \geq C\left(\frac{1}{m}+\frac{1}{n}\right)\log^{10} N, \tag{32}$$

$$p \geq \frac{Cr^3 K_{A,\xi}^2}{\varepsilon^2}\left(1+\frac{\mu^2}{\log^2 N}\right)\left(1+\frac{r^3\log N}{N}\right)\left(\frac{1}{m}+\frac{1}{n}\right)\log^6 N. \tag{33}$$

Let $\rho := \hat{p}/p$. From the sampling density assumption, a standard application of concentration bounds [16, 8] guarantees that, with probability $1-O(N^{-2})$.

$$0.9 \leq 1-\frac{1}{\sqrt{N}} \leq 1-\frac{\log N}{\sqrt{pmn}} \leq \rho \leq 1+\frac{\log N}{\sqrt{pmn}} \leq 1+\frac{1}{\sqrt{N}} \leq 1.1. \tag{34}$$

Furthermore, an application of well-established bounds on random matrix norms gives

$$\|E\| \leq 2\kappa\sqrt{N}, \tag{35}$$

with probability $1-O(N^{-1})$. See [1, 28], [25, Lemma A.7] or [1] for detailed proofs. Therefore we can assume both Eqs. (34) and (35) at the cost of an $O(N^{-1})$ exceptional probability.

Let $C_0 := 40$. The index s chosen in the SVD step of Algorithm 1.3 is the largest such that

$$\hat{\delta}_s \geq C_0 K_{A,\xi}\sqrt{r_{\max}N/\hat{p}} = C_0\rho^{-1/2}\kappa\sqrt{r_{\max}N}.$$

Firstly, we show that SVD step is guaranteed to choose a valid $s \in [r]$. Choose an index $l \in [r]$ such that $\delta_l \geq \sigma_1/r \geq 100\kappa\sqrt{r_{\max}N}$, we have

$$\hat{\delta}_l \geq \rho^{-1/2}\tilde{\delta}_l \geq \rho^{-1/2}(\delta_l - 2\|E\|) \geq (100 r_{\max}^{1/2} - 4)\rho^{-1/2}\kappa\sqrt{N} \geq 2C_0\rho^{-1/2}\kappa\sqrt{r_{\max}N},$$

so the cutoff point s is guaranteed to exist. To see why $s \in [r]$, note that

$$\hat{\delta}_{r+1} \leq \rho^{-1/2}\tilde{\sigma}_{r+1} \leq \rho^{-1/2}\|E\| \leq 2\rho^{-1/2}\kappa\sqrt{r_{\max}N} < C_0\rho^{-1/2}\kappa\sqrt{r_{\max}N}.$$

To prove that the first three steps recovers every entry of A to within ε, namely $\|\hat{A}_s - A\|_\infty \leq \varepsilon$, we will first show that $\|\tilde{A}_s - A\|_\infty \leq \varepsilon/2$ (with probability $1-O(N^{-1})$. We proceed in two steps:

1. We will show that $\|A_s - A\|_\infty \leq \varepsilon/4$. To this end, we establish this fact:

$$\sigma_{s+1} \leq r\delta_{s+1} \leq r(\tilde{\delta}_{s+1} + 2\|E\|) \leq r(C_0\rho^{-1/2}\sqrt{r_{\max}} + 4)\kappa\sqrt{N} \leq 2rC_0K_{A,\xi}\sqrt{r_{\max}N/p}. \tag{36}$$

For each fixed indices j,k, we have

$$|(A_s - A)_{jk}| = \left|U_{j,\cdot}^T\Sigma_{[s+1,r]}V_{k,\cdot}\right| \leq \sigma_{s+1}\|U\|_{2,\infty}\|V\|_{2,\infty} \leq 2rC_0K_{A,\xi}\sqrt{\frac{r_{\max}N}{p}}\frac{r\mu}{\sqrt{mn}}$$
$$= \sqrt{\frac{4C_0^2 r^4 r_{\max}\mu^2 K_{A,\xi}^2}{p}\left(\frac{1}{m}+\frac{1}{n}\right)} \leq \varepsilon/4.$$

where the last inequality comes from the assumption (31) if C is large enough. Since this holds for all pairs (j,k), we have $\|A_s - A\|_\infty \leq \varepsilon/4$.

2. Secondly, we will show that $\|\tilde{A}_s - A_s\|_\infty \le \varepsilon/4$ with probability $1 - O(N^{-1})$. We aim to use Theorem 2.3, so let us translate its terms into the current context. By the sampling density condition, we have the following lower bounds for δ_s and σ_s:

$$\sigma_s \ge \delta_s \ge \tilde{\delta}_s - 2\|E\| \ge C_0 \rho^{-1/2} \kappa \sqrt{r_{\max} N} - 2\|E\| \ge 0.9 C_0 \kappa \sqrt{r_{\max} N}. \tag{37}$$

Consider the condition (27). By Eq. (37), we have

$$\frac{\kappa\sqrt{N}}{\sigma_s} \vee \frac{r\kappa \log^{1/2} N}{\delta_s} \vee \frac{\kappa\sqrt{rN}}{\sqrt{\delta_s \sigma_s}} \le \frac{4}{C_0 r_{\max}^{1/2}} \vee \frac{4r\log^{1/2} N}{C_0\sqrt{r_{\max}N}} \vee \frac{4\sqrt{r}}{C_0\sqrt{r_{\max}}} < \frac{1}{16},$$

therefore we are guaranteed to be able to apply Theorem 2.3. We will get, for a constant C_1,

$$\|\tilde{A}_s - A_s\|_\infty \le C_1 \mu_{UV}\mu_{VU} r\sigma_s \left(\frac{\kappa\sqrt{N}}{\sigma_s} + \frac{r\kappa\sqrt{\log N}}{\delta_s} + \frac{r\kappa^2 N}{\delta_s\sigma_s}\right) + \frac{1}{N}.$$

Let us consider the first term in the product, $\mu_{UV}\mu_{VU}$.

$$\begin{aligned}\mu_{UV} &= \frac{K\|U\|_{2,\infty}\log^3 N}{\sqrt{rN}} + \frac{\log^{3/2} N}{\sqrt{N}} + \frac{\|V\|_{2,\infty}\log N}{\sqrt{r}}\\ &\le \frac{\sqrt{\mu}\log^3 N}{\sqrt{pmN}} + \frac{\log^{3/2} N}{\sqrt{N}} + \frac{\sqrt{\mu}\log N}{\sqrt{n}} \le \frac{\log^{3/2} N}{\sqrt{N}} + \frac{\sqrt{2\mu}\log N}{\sqrt{n}},\end{aligned}$$

where the first inequality comes from (32) if C is large enough. Similarly,

$$\mu_{VU} \le N^{-1/2}\log^{3/2} N + m^{-1/2}\sqrt{2\mu}\log N.$$

Therefore,

$$\begin{aligned}\mu_{UV}\mu_{VU} &\le \frac{\log^3 N}{N} + \frac{\sqrt{\mu}\log^{5/2} N}{\sqrt{N}} \cdot \frac{\sqrt{2m}+\sqrt{2n}}{\sqrt{mn}} + \frac{2\mu\log^2 N}{\sqrt{mn}}\\ &\le \log^2 N \frac{\log N + 4\sqrt{\mu}\sqrt{\log N} + 4\mu}{2\sqrt{mn}} \le \log^2 N \frac{\log N + 4\mu}{\sqrt{mn}}.\end{aligned}$$

For the second term, we have the following upper bound:

$$\begin{aligned} r\sigma_s\left(\frac{\kappa\sqrt{N}}{\sigma_s} + \frac{r\kappa\sqrt{\log N}}{\delta_s} + \frac{r\kappa^2 N}{\delta_s\sigma_s}\right) &= r\left(\kappa\sqrt{N} + r\kappa\sqrt{\log N}\frac{\sigma_s}{\delta_s} + \frac{r\kappa^2 N}{\delta_s}\right)\\ &\le r\left(\kappa\sqrt{N} + r^2\kappa\sqrt{\log N} + \frac{r\kappa\sqrt{N}}{C_0\sqrt{r_{\max}}}\right) \le \frac{\sqrt{2}r^{3/2}K_{A,\xi}}{\sqrt{p}}\left(\sqrt{N} + r^{3/2}\sqrt{\log N}\right).\end{aligned}$$

Multiplying the two terms, we have by Theorem 2.3,

$$\begin{aligned}\|\tilde{A}_s - A_s\|_\infty &\le \log^2 N \cdot \frac{\log N + 4\mu}{\sqrt{mn}} \cdot \frac{\sqrt{2}r^{3/2}K_{A,\xi}}{\sqrt{p}}\left(\sqrt{N} + r^{3/2}\sqrt{\log N}\right)\\ &\le \sqrt{\frac{r^3 K_{A,\xi}^2 \log^6 N}{p}\left(1 + \frac{4\mu^2}{\log^2 N}\right)\left(1 + \frac{r^3\log N}{N}\right)\left(\frac{1}{m} + \frac{1}{n}\right)} \le \varepsilon/4.\end{aligned} \tag{38}$$

where the last inequality comes from the condition (33) if C is large enough.

After the two steps above, we obtain $\|\tilde{A}_s - A\|_\infty \le \varepsilon/2$ with probablity $1 - O(N^{-1})$. Finally, we get, using Fact (34) and the triangle inequality,

$$\|\hat{A}_s - A\|_\infty = \left\|\rho^{-1}\tilde{A}_s - A\right\|_\infty \le \frac{1}{\rho}\|\tilde{A}_s - A\|_\infty + \left|\frac{1}{\rho} - 1\right| \|A\|_\infty \le \frac{\varepsilon/2}{.9} + \frac{K_A}{.9\sqrt{N}} < \varepsilon.$$

The first part is proven, which immediately implies the rounding step recovers A exactly with $\varepsilon = 0.4$. Summing up the exceptional probabilities, we get $O(N^{-1})$. The proof is complete. □

In the next section, we will prove Theorem 2.2. Then we will provide necessary probablistic bounds and prove Theorem 2.3 in Section 4. The technical bounds whose proof do not fit in the main body will be in Section 5.

3 Proof of main results: the deterministic case

To prove Theorems 2.2, we are interested in the differences $\tilde{V}_S\tilde{V}_S^T - V_S V_S^T$ and $\tilde{A}_s - A_s$. As we shall see in the next part, both can be expressed as almost identical power series of the noise matrix E, provided the conditions of the theorem hold. We will establish a common procedure to bound both, consisting of three steps:

1. Expand the matrices above as instances of the same generic power series.
2. Bound each individual term in the generic power series with terms that shrink geometrically with each power of E.
3. Sum all bounding terms as a geometric series to obtain the final bound.

3.1 Step 1: A Taylor-like expansion using contour integration

We first introduce the symmetrization trick. For any $m \times n$ matrix A of rank r, let

$$A_{\text{sym}} := \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}. \tag{39}$$

If A has the SVD: $A = \sigma_1 u_1 v_1^T + \ldots + \sigma_r u_r v_r^T$, then A_{sym} has the eigendecomposition:

$$A_{\text{sym}} = W \Lambda W^T = \sum_{i=1}^{m+n} \lambda_i w_i w_i^T,$$

where for each $i \in [m+n]$:

- If $1 \le i \le r$: $\lambda_i = \sigma_i$ and $w_i = \frac{1}{\sqrt{2}}[u_i,\ v_i]^T$.
- If $r + 1 \le i \le 2r$: $\lambda_i = \sigma_i$ and $w_i = \frac{1}{\sqrt{2}}[u_i,\ -v_i]^T$.
- The remaining eigenvalues are all 0 and the corresponding eigenvectors are an arbitrary basis for the complement space of the span of the nonzero eigenspace.

Note that the singular values of A_{sym} are again $\sigma_1, \dots, \sigma_r$, but each with multiplicity 2, thus the matrices

$$W_s := \begin{bmatrix} w_1, & w_2, & \dots & w_s, & w_{r+1}, & \dots & w_{r+s} \end{bmatrix}, \quad (A_{\text{sym}})_s := \sum_{i=1}^{s} \lambda_i w_i w_i^T + \sum_{i=r+1}^{r+s} \lambda_i w_i w_i^T$$

are respectively, the singular basis of the most significant $2s$ vectors and the best rank-$2s$ approximation of A_{sym}. However, we still use the subscript s instead of $2s$ to emphasize their relation to the quantities U_s, V_s and A_s. For an arbitrary subset $S \subset [r]$, we analogously denote

$$W_S := \begin{bmatrix} w_i, & w_{i+r} \end{bmatrix}_{i \in S}, \quad \text{and} \quad (A_{\text{sym}})_S := \sum_{i \in S} \lambda_i w_i w_i^T + \sum_{i-r \in S} \lambda_i w_i w_i^T.$$

We define $\tilde{\Lambda}$, $\tilde{\lambda}_i$, $\tilde{w}_i$ and $\tilde{W}_s$, $\tilde{W}_S$, $\tilde{A}_s$, $\tilde{A}_S$ similarly for $\tilde{A} = A + E$. The resolvent of A_{sym}, which is a function of a complex variable z, can now be written:

$$(zI - A_{\text{sym}})^{-1} = \sum_{i=1}^{m+n} \frac{w_i w_i^T}{z - \lambda_i} = \sum_{i=1}^{2r} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z},$$

where W is the matrix whose columns are $\{w_i\}_{i=1}^{2 \operatorname{rank} M}$.

The main idea for proving Theorems 2.2 and 2.3 is a Taylor-like expansion of the difference of resolvents

$$(zI - \tilde{A}_{\text{sym}})^{-1} - (zI - A_{\text{sym}})^{-1} = \sum_{\gamma=1}^{\infty} \left[(zI - A_{\text{sym}})^{-1} E_{\text{sym}} \right]^{\gamma} (zI - A_{\text{sym}})^{-1},$$

which has been proven in [26] to hold whenever the right-hand side converges.

Assuming this is true, we can then extract out the differences of the singular vector projections. The above can be rewritten as

$$\sum_{i=1}^{m+n} \frac{\tilde{w}_i \tilde{w}_i^T}{z - \tilde{\lambda}_i} - \sum_{i=1}^{m+n} \frac{w_i w_i^T}{z - \lambda_i} = \sum_{\gamma=1}^{\infty} \left[\Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big) E_{\text{sym}} \right]^{\gamma} \Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big). \tag{40}$$

Let Γ_S denote an arbitrary contour in $\mathbb{C}$ that encircles $\{\pm\sigma_i, \pm\tilde{\sigma}_i\}_{i \in S}$ and none of the other eigenvalues of $\tilde{W}$ and W. Integrating over Γ_S of both sides and dividing by $2\pi i$, we have

$$\begin{aligned} \begin{bmatrix} \tilde{U}_S \tilde{U}_S^T - U_S U_S^T & 0 \\ 0 & \tilde{V}_S \tilde{V}_S^T - V_S V_S^T \end{bmatrix} &= \tilde{W}_S \tilde{W}_S - W_S W_S^T \\ &= \sum_{\gamma=1}^{\infty} \oint_{\Gamma_S} \frac{\mathrm{d}z}{2\pi i} \left[\Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big) E_{\text{sym}} \right]^{\gamma} \Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big). \end{aligned} \tag{41}$$

We quickly note that the following identity can be obtained by multiplying both sides of Eq. (40) with z, dividing by $2\pi i$ and integrating over Γ_S:

$$\begin{aligned} (\tilde{A}_S - A_S)_{\text{sym}} = (\tilde{A}_{\text{sym}})_S - (A_{\text{sym}})_S &= \sum_{i \in S \vee i-r \in S} \left(\frac{z \tilde{w}_i \tilde{w}_i^T}{z - \tilde{\lambda}_i} - \frac{z w_i w_i^T}{z - \lambda_i} \right) \\ &= \sum_{\gamma=1}^{\infty} \oint_{\Gamma_S} \frac{z\mathrm{d}z}{2\pi i} \left[\Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big) E_{\text{sym}} \right]^{\gamma} \Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big). \end{aligned} \tag{42}$$

It is thus beneficial to find a *common strategy* that can bound both these expressions at the same time, It is thus beneficial to consider the following general expression for $\nu \in \mathbb{N}$:

$$\mathcal{T}_\nu := \oint_{\Gamma_S} \frac{z^\nu \mathrm{d}z}{2\pi i} \left[\Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big) E_{\text{sym}} \right]^\gamma \Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big). \tag{43}$$

Our common strategy will establish a bound on this generic form, which implies Eqs. (22) and (23) for $\nu = 0$ and Eq. (24) for $\nu = 1$.

Next, we will expand each power in this series into a big sum and describe the idea to bound each summand individually. To ease the notation, denote

$$P_i := w_i w_i^T, \quad \text{for } i = 1, 2, \ldots, 2r, \quad \text{and} \quad Q := I - WW^T.$$

Note that we have the following identity of terms:

$$\sum_{\lambda_i \neq 0} \frac{P_i}{z - \lambda_i} + \frac{Q}{z} = \sum_{\lambda_i \neq 0} \frac{\lambda_i P_i}{z(z - \lambda_i)} + \frac{I}{z}$$

Plugging the above into Eq. (43), we get

$$\mathcal{T}_\nu = \sum_{\gamma=1}^{\infty} \oint_{\Gamma_S} \frac{z^\nu \mathrm{d}z}{2\pi i} \left[\Big(\sum_{\lambda_i \neq 0} \frac{\lambda_i P_i}{z(z - \lambda_i)} + \frac{I}{z} \Big) E_{\text{sym}} \right]^\gamma \Big(\sum_{\lambda_i \neq 0} \frac{\lambda_i P_i}{z(z - \lambda_i)} + \frac{I}{z} \Big). \tag{44}$$

Fix $\gamma \in \mathbb{N}$, $\gamma \geq 1$ and consider the γ-power term in the series. Expanding the power yields a sum of terms of the form

$$\oint_{\Gamma_S} \frac{z^\nu \mathrm{d}z}{2\pi i} \left(\frac{I}{z} E_{\text{sym}} \right)^{\alpha_0} \frac{\lambda_{i_{11}} P_{i_{11}}}{z(z - \lambda_{i_{11}})} E_{\text{sym}} \frac{\lambda_{i_{12}} P_{i_{12}}}{z(z - \lambda_{i_{12}})} \ldots E_{\text{sym}} \frac{\lambda_{i_{1\beta_1}} P_{i_{1\beta_1}}}{z(z - \lambda_{i_{1\beta_1}})} E_{\text{sym}} \left(\frac{I}{z} E_{\text{sym}} \right)^{\alpha_1}$$
$$\frac{\lambda_{i_{21}} P_{i_{21}}}{z(z - \lambda_{i_{21}})} E_{\text{sym}} \ldots E_{\text{sym}} \frac{\lambda_{i_{2\beta_2}} P_{i_{2\beta_2}}}{z(z - \lambda_{i_{2\beta_2}})} E_{\text{sym}} \left(\frac{I}{z} E_{\text{sym}} \right)^{\alpha_2} \ldots E_{\text{sym}} \frac{\lambda_{i_{h\beta_h}} P_{i_{h\beta_h}}}{z(z - \lambda_{i_{h\beta_h}})} \left(E_{\text{sym}} \frac{I}{z} \right)^{\alpha_h},$$

which can be rewritten as

$$\mathcal{C}_\nu(\mathbf{I}) E_{\text{sym}}^{\alpha_0} \left[\prod_{k=1}^{h-1} \mathcal{M}(\mathbf{i}_k) E_{\text{sym}}^{\alpha_k + 1} \right] \mathcal{M}(\mathbf{i}_h) E_{\text{sym}}^{\alpha_h}, \tag{45}$$

where we denote

$$\mathbf{I} := [\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_h], \qquad \mathbf{i}_k := [i_{k1}, i_{k2}, \ldots, i_{k\beta_k}],$$

and for a non-empty sequence $\mathbf{i} = [i_1, i_2, \ldots, i_\beta]$ we denote the *monomial matrix*

$$\mathcal{M}(\mathbf{i}) := P_{i_1} \prod_{j=2}^{\beta} E_{\text{sym}} P_{i_j}, \tag{46}$$

and the scalar *integral coefficient* for the non-empty sequence $\mathbf{I} = [i_{11}, i_{12}, \ldots, i_{h\beta_h}]$

$$\mathcal{C}_\nu(\mathbf{I}) := \oint_{\Gamma_S} \frac{\mathrm{d}z}{2\pi i} \frac{1}{z^{\gamma+1}} \prod_{k=1}^{h} \prod_{j=1}^{\beta_k} \frac{\lambda_{i_{kj}}}{z - \lambda_{i_{kj}}}. \tag{47}$$

Let $\Pi_h(\gamma)$ be the set of all tuples of $\boldsymbol{\alpha} = [\alpha_k]_{k=0}^h$ and $\boldsymbol{\beta} = [\beta_k]_{k=1}^h$ such that:

$$\begin{aligned} &\bullet \quad \alpha_0, \alpha_h \geq 0, \quad \text{ and } \alpha_k \geq 1 \text{ for } 1 \leq k \leq h-1, \\ &\bullet \quad \beta_k \geq 1 \text{ for } 1 \leq k \leq h, \\ &\bullet \quad \alpha + \beta = \gamma + 1, \quad \text{ where } \alpha := \sum_{k=0}^h \alpha_k, \quad \text{ and } \beta := \sum_{k=1}^h \beta_k. \end{aligned} \tag{48}$$

Note that the conditions above imply $2h - 1 \leq \gamma + 1$, so the maximum value for h is $\lfloor \gamma/2 \rfloor + 1$.

Combining Eqs. (44), (47), and (46), we get the expansion

$$\mathcal{T}_\nu = \sum_{\gamma=1}^{\infty} \mathcal{T}_\nu^{(\gamma)}, \quad \text{ where } \mathcal{T}_\nu^{(\gamma)} = \sum_{h=0}^{\lfloor \gamma/2 \rfloor + 1} \mathcal{T}_\nu^{(\gamma,h)}, \quad \text{ where } \mathcal{T}_\nu^{(\gamma,h)} = \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \Pi_h(\gamma)} \mathcal{T}_\nu(\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{49}$$

where

$$\mathcal{T}_\nu(\boldsymbol{\alpha},\boldsymbol{\beta}) := \sum_{\mathbf{I} \in [2r]^{\beta_1 + \ldots + \beta_h}} \mathcal{C}_\nu(\mathbf{I}) E_{\text{sym}}^{\alpha_0} \left[\prod_{k=1}^{h-1} \mathcal{M}\left(\mathbf{i}_k\right) E_{\text{sym}}^{\alpha_k + 1} \right] \mathcal{M}\left(\mathbf{i}_h\right) E_{\text{sym}}^{\alpha_h}. \tag{50}$$

Let us look at the main theorem again. Consider Eqs. (22) and (24). The most common way to bound the infinity norms of $\tilde{V}_S \tilde{V}_S^T - V_S V_S^T$ and $\tilde{A}_s - A_s$ with high probability is the following:

1. Consider an arbitrary of the matrices above.

2. Bound this entry with probability close to 1.

3. Apply a union bound over all possible entries.

For Eq. (23), since the 2-to-∞ norm of a matrix is simply the norm of the largest row, we can use the aforementioned strategy by replacing the fixed entry with a fixed row in the first step.

Consider an arbitrary jl-entry of $\tilde{V}_S \tilde{V}_S^T - V_S V_S^T$. It corresponds to the $(m+j)(m+l)$-entry of $\tilde{W}_S \tilde{W}_S^T - W_S W_S^T$. From Eqs. (49) and (50), this entry will can be written as

$$e_{m+n,m+j}^T \mathcal{T}_0 e_{m+n,m+k} = \sum_{\gamma=1}^{\infty} \sum_{h=0}^{\lfloor \gamma/2 \rfloor + 1} \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \Pi_h(\gamma)} e_{m+n,m+j}^T \mathcal{T}_0(\boldsymbol{\alpha},\boldsymbol{\beta}) e_{m+n,m+k}. \tag{51}$$

Similarly, the expansions for a single row of $\tilde{V}_S \tilde{V}_S^T - V_S V_S^T$ or a single entry of $\tilde{A}_s - A_s$ also have the form $M^T \mathcal{T}_\nu M'$, where M and M' is either a standard basis vector $e_{m+n,j}$ or I_{m+n}.

It is thus beneficial to establish a bound for generic choices of M and M' in the operator norm, which covers both the Euclidean norm of a vector and the absolute value of a number. In fact, our proof works for any sub-multiplicative norm, which we use the generic $\|\cdot\|$ to denote from this point to the end of this section.

We summarize the objects involved in the bound and its proof below.

Setting 3.1. Let the following objects and properties be given:

- $\boldsymbol{\Lambda} = \{\lambda_i\}_{i \in [2r]}$: a set of real numbers such that $\delta_i := \lambda_i - \lambda_{i+1} > 0$ for each $i \leq r-1$, $\delta_r = \lambda_r > 0$, and $\lambda_i = -\lambda_{i-r}$ for $i \geq r+1$.

- $W = \{w_i\}_{i \in [2r]}$ for $i \in [2r]$: a set of orthonormal vectors in $\mathbb{R}^{m+n}$. We slightly abuse the notation here and use W as an orthogonal matrix when necessary.

- S: an arbitrary subset of $\{\lambda_i\}_{i\in[r]}$.
- γ and h: positive integers such that $2h-1 \leq \gamma+1$.
- $\Pi_h(\gamma)$: the set of pairs of index sequences $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ satisfying Eq. (48).
- $\mathcal{C}_\nu = \mathcal{C}_{\nu,\mathbf{\Lambda},S} : \bigcup_{\beta=0}^{\gamma+1}[2r]^\beta \to \mathbb{R}$: the mapping from an index sequence $\mathbf{I}$ to its integral coefficient defined in Eq. (47).
- E: an arbitrary matrix in $\mathbb{R}^{m\times n}$, and let E_{sym} be defined analogously to Eq. (39).
- $\mathcal{M} = \mathcal{M}_{W,E} : \bigcup_{\beta=0}^{\gamma+1}[2r]^\beta \to \mathbb{R}^{(m+n)\times(m+n)}$: the mapping from an index sequence $\mathbf{i}$ to its monomial matrix defined in Eq. (46).
- M and M': arbitrary matrices with $m+n$ rows.
- $\mathcal{T}_\nu$: the target sum defined in Eq. (49). The sub-terms $\mathcal{T}_\nu^{(\gamma)}$, $\mathcal{T}_\nu^{(\gamma,h)}$ are defined in the same equation, while $\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is defined in Eq. (50).

We aim to upper bound $\|M^T\mathcal{T}_\nu M'\|$ for a sub-multiplicative norm $\|\cdot\|$.

This concludes the first of the three-step common strategy. In the next section, we will establish a bound on $\|M^T\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})M'\|$ for each pair $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, and in the section after, will sum all such bounds over all choices of α, β, h and γ.

3.2 Step 2: Bounding the terms in the infinite series

We begin with a bound for the integral coefficients, in the cases that we are interested in.

Lemma 3.2. *Consider the objects defined in Setting 3.1 and an arbitrary tuple $\mathbf{I} := \{i_k\}_{k\in[\beta]} \in [2r]^\beta$ and denote the following:*

$$\begin{aligned}\beta_S(\mathbf{I}) &:= \left|\{1 \leq k \leq \beta : \; i_k \in S\}\right|, \\ \lambda_S(\mathbf{I}) &:= \min\{|\lambda_{i_k}| : k \in S\}, \\ \Delta_S(\mathbf{I}) &:= \min\{|\lambda_{i_k} - \lambda_{i_l}| : i_k \in S, i_l \notin S\}.\end{aligned}$$

We have,

$$|\mathcal{C}_\nu(\mathbf{I})| \leq \mathrm{L}_\nu(\mathbf{I}) \left(1 + \frac{\Delta_S(\mathbf{I})}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma+\beta_S(\mathbf{I})-2}{\beta_S(\mathbf{I})-1} \frac{1}{\lambda_S(\mathbf{I})^{\gamma+1-\beta}\Delta_S(\mathbf{I})^{\beta-1}}, \tag{52}$$

where $\mathrm{L}_0(\mathbf{I}) = 2$ and $\mathrm{L}_1(\mathbf{I}) = \lambda_S(\mathbf{I})$. Consequently, when either $\nu = 0$ or $\nu = 1$ and $\lambda_S(\mathbf{I})$ if $S = [s]$ for some $s \in [r]$:

$$|\mathcal{C}_\nu(\mathbf{I})| \leq \mathrm{L}_\nu \left(1 + \frac{\Delta_S}{\lambda_S}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma+\beta_S(\mathbf{I})-2}{\beta_S(\mathbf{I})-1} \frac{1}{\lambda_S^{\gamma+1-\beta}\Delta_S^{\beta-1}}, \tag{53}$$

where $\mathrm{L}_0 = 2$ and $\mathrm{L}_1 = \lambda_s$.

Proof. See Section 5.1. □

Note that in order for Eq. (53) to hold for $\nu = 1$, S needs to contain exactly the first s indices for some s. In the steps that follow, we will mainly use Eq. (53), with one exception where the more precise Eq. (52) is needed. It thus makes sense to keep both.

Using the above lemma, we obtain the following bound for $\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})$:

Lemma 3.3. *Consider objects in Setting 3.1 and Lemma 3.2. Assume that either $\nu = 0$ or $\nu = 1$ and $S = [s]$ for some $s \in [r]$. For each $\gamma \geq 1$, $0 \leq h \leq \gamma/2 + 1$, $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Pi_h(\gamma)$, we have*

$$\begin{aligned} \|\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})\| \leq \mathrm{L}_\nu \binom{\gamma + \beta - 2}{\beta - 1} \|W^T E_{\text{sym}} W\|_\infty^{\beta - h} \|E\|^{\alpha - \alpha_0 - \alpha_h + h - 1} \frac{(2|S| + 2\rho|S^c|)^{\beta - 2}}{\lambda_S^{\gamma + 1 - \beta} \Delta_S^{\beta - 1}} \\ \cdot \sum_{i=1}^{2r} \rho^{\mathbf{1}\{i_k \in S^c\}} \left\|w_i^T E_{\text{sym}}^{\alpha_0} M\right\| \cdot \sum_{i=1}^{2r} \rho^{\mathbf{1}\{i_k \in S^c\}} \left\|w_i^T E_{\text{sym}}^{\alpha_h} M'\right\|, \end{aligned} \tag{54}$$

where $\rho := (\lambda_S + \Delta_S)/(2\lambda_S)$. Consequently,

$$\begin{aligned} \|\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})\| \leq \mathrm{L}_\nu \binom{\gamma + \beta - 2}{\beta - 1} \|W^T E_{\text{sym}} W\|_\infty^{\beta - h} \|E\|^{\alpha - \alpha_0 - \alpha_h + h - 1} \\ \cdot \frac{(2r)^{\beta - 2}}{\lambda_S^{\gamma + 1 - \beta} \Delta_S^{\beta - 1}} \cdot \sum_{i=1}^{2r} \left\|w_i^T E_{\text{sym}}^{\alpha_0} M\right\| \cdot \sum_{i=1}^{2r} \left\|w_i^T E_{\text{sym}}^{\alpha_h} M'\right\|. \end{aligned} \tag{55}$$

Since the bound in Eq. (54) is rather cumbersome, we will mostly use Eq. (55). Eq. (54) is still useful for future development in this direction.

Proof. We can simplify the monomial matrix by extracting many scalars from it as below:

$$\mathcal{M}\left(\mathbf{I}\right) = w_{i_1} w_{i_1}^T \prod_{j=2}^{\beta} E_{\text{sym}} w_{i_j} w_{i_j}^T = w_{i_1} \Big(\prod_{j=2}^{\beta} w_{i_{j-1}}^T E_{\text{sym}} w_{i_j}\Big) w_{i_\beta}^T = \Big(\prod_{j=2}^{\beta} X_{i_{j-1} i_j}\Big) w_{i_1} w_{i_\beta}^T,$$

where we temporarily denote $X := W^T E_{\text{sym}} W$. From Eq. (50), we thus have

$$\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{\mathbf{I} \in [2r]^\beta} \mathcal{C}_\nu(\mathbf{I}) D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{I}) \left(M^T E_{\text{sym}}^{\alpha_0} w_{i_1}\right) \left(w_{i_{h\beta_h}}^T E_{\text{sym}}^{\alpha_h} M'\right), \tag{56}$$

where

$$D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{I}) := \prod_{k=1}^{h} \prod_{j=2}^{\beta_k} X_{i_{k(j-1)} i_{kj}} \cdot \prod_{k=1}^{h-1} w_{i_{k\beta_k}}^T E_{\text{sym}}^{\alpha_k + 1} w_{i_{(k+1)1}}. \tag{57}$$

We can obtain an upper bound for $|D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{I})|$ by replacing every term in the first product of Eq. (57) with $\|X\|_\infty$, and replacing every instance of E with $\|E\|$ in the second product. Therefore

$$|D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{I})| \leq \|X\|_\infty^{\beta - h} \|E\|^{\alpha - \alpha_0 - \alpha_h + h - 1}. \tag{58}$$

This bound does not depend on the choice of $\mathbf{I}$, so we have

$$\|\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})\| \leq \|X\|_\infty^{\beta - h} \|E\|^{\alpha - \alpha_0 - \alpha_h + h - 1} \sum_{\mathbf{I} \in [2r]^\beta} |\mathcal{C}_\nu(\mathbf{I})| \left\|M^T E_{\text{sym}}^{\alpha_0} w_{i_1}\right\| \left\|w_{i_{h\beta_h}}^T E_{\text{sym}}^{\alpha_h} M'\right\|.$$

Note that the previous two steps use the sub-multiplicativity of $\|\cdot\|$. Temporarily let T be the sum on the right-hand side. Applying the integral coefficient bound from Lemma 3.2, we have

$$\begin{aligned} |\mathcal{C}_\nu(\mathbf{I})| &\leq \mathrm{L}_\nu \left(1 + \frac{\Delta_S}{\lambda_S}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \frac{1}{\lambda_S^{\gamma + 1 - \beta} \Delta_S^{\beta - 1}} \\ &\leq \mathrm{L}_\nu \binom{\gamma + \beta - 2}{\beta - 1} \left(\frac{\lambda_S + \Delta_S}{2\lambda_S}\right)^{\beta_{S^c}(\mathbf{I})} \frac{1}{\lambda_S^{\gamma + 1 - \beta} \Delta_S^{\beta - 1}}. \end{aligned}$$

Replacing $(\lambda_S + \Delta_S)/(2\lambda_S)$ with ρ for convenience and plugging this bound into T, we get

$$T \leq \mathrm{L}_\nu \binom{\gamma+\beta-2}{\beta-1} \frac{1}{\lambda_S^{\gamma+1-\beta} \Delta_S^{\beta-1}} \sum_{\mathbf{I} \in [2r]^\beta} \left\| M^T E_{\text{sym}}^{\alpha_0} w_{i_1} \right\| \left\| w_{i_{h\beta_h}}^T E_{\text{sym}}^{\alpha_h} M' \right\| \prod_{k=1}^{\beta} \rho^{\mathbf{1}\{i_k \in S^c\}}$$
$$= \mathrm{L}_\nu \binom{\gamma+\beta-2}{\beta-1} \frac{(2|S| + 2\rho|S^c|)^{\beta-2}}{\lambda_S^{\gamma+1-\beta} \Delta_S^{\beta-1}} \sum_{i=1}^{2r} \rho^{\mathbf{1}\{i_k \in S^c\}} \left\| w_i^T E_{\text{sym}}^{\alpha_0} M \right\| \sum_{i=1}^{2r} \rho^{\mathbf{1}\{i_k \in S^c\}} \left\| w_i^T E_{\text{sym}}^{\alpha_h} M' \right\|.$$

The proof of Eq. (54) is complete. Eq. (55) follows from the fact that $\rho \leq 1$. □

To bring the bound above closer to the form required in Theorem 2.2, we will need to find good bounds for the sum $\sum_{i=1}^{2r} \left\| w_i^T E_{\text{sym}}^{\alpha} M \right\|$ for a given α and M, which should involve the terms μ_{UV} and μ_{VU}. We are interested in the cases $M = I_{m+n}$ and $M = e_{m+n,k}$ for some fixed k. In the former, one cannot do much better than the naive bound $2r\|E_{\text{sym}}\|^\alpha \|M\|$, even when E is random. We are interested in the cases $M = I_{m+n}$ and $M = e_{m+n,k}$ for some fixed k. In the former, one cannot do much better than the naive bound $2r\|E_{\text{sym}}\|^\alpha \|M\|$, even when E is random. In the latter, recall the condition (20) in Theorem 2.2. The bound below is a direct result.

Lemma 3.4. *Consider the objects in Setting 3.1 and terms μ_{UV}, μ_{VU} and $\mathcal{K}$ from Theorem 2.2. Then for a matrix $M \in \{e_{m+n,k}\}_{k \in [m+n]} \cup \{I_{m+n}\}$, we have $\sum_{i=1}^{2r} \left\| w_i^T E_{\text{sym}}^{\alpha} M \right\| \leq 2r\mu\mathcal{K}_0^\alpha$, where*

- *$\mu = \frac{1}{\sqrt{2}}\mu_{UV}$ and $\mathcal{K}_0 = \mathcal{K}$ for $M \in \{e_{m+n,k}\}_{k \in [[m+1, m+n]]}$.*
- *$\mu = \frac{1}{\sqrt{2}}\mu_{VU}$ and $\mathcal{K}_0 = \mathcal{K}$ for $M \in \{e_{m+n,k}\}_{k \in [m]}$.*
- *$\mu = 1$ and $\mathcal{K}_0 = \|E\|$ for $M = I_{m+n}$.*

In the upcoming third step, the formulas of μ is not important for the calculations, so we can treat this lemma as a black box and leave its proof for later. This concludes the second step.

3.3 Step 3: Summing up the term-wise bounds in the series

Consider Eq. (49) again. The series we need to bound has the form

$$\mathcal{T} = \sum_{\gamma=1}^{\infty} \mathcal{T}_\nu^{(\gamma)}, \quad \text{where } \mathcal{T}_\nu^{(\gamma)} = \sum_{h=0}^{\lfloor \gamma/2 \rfloor + 1} \mathcal{T}_\nu^{(\gamma,h)}, \quad \text{where } \mathcal{T}_\nu^{(\gamma,h)} = \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \Pi_h(\gamma)} \mathcal{T}\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{59}$$

We will bound each term in these series progressively, starting from $\mathcal{T}_\nu^{(\gamma,h)}$. The end result will be a general bound that implies all all of Eq. (22), (23) and (24). We use the assumption below.

Assumption 3.5. *Consider the objects given by Setting 3.1. Assume that the real numbers L_ν, μ, μ', $\mathcal{K}_0$ and $\mathcal{K}_0'$ satisfy*

$$\forall S \subset [2r], \mathbf{I} \in [2r]^\beta: \qquad |\mathcal{C}_\nu(\mathbf{I})| \leq \mathrm{L}_\nu \left(1 + \frac{\Delta_S}{\lambda_S}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \frac{1}{\lambda_S^{\gamma+1-\beta} \Delta_S^{\beta-1}}, \tag{60}$$

$$\forall \alpha \in [\lceil 10\log(m+n) \rceil]: \qquad \sum_{i=1}^{2r} \|w_i^T E_{\text{sym}}^\alpha M\| \leq 2r\mu\mathcal{K}_0^\alpha \quad \text{and} \quad \sum_{i=1}^{2r} \|w_i^T E_{\text{sym}}^\alpha M'\| \leq 2r\mu'\mathcal{K}_0'^\alpha. \tag{61}$$

Assume the subset $S \in [r]$ and the real numbers R, R_1 and R_2 satisfy

$$\frac{\|E\|}{\lambda_S} \vee \frac{\mathcal{K}_0}{\lambda_S} \vee \frac{\mathcal{K}_0'}{\lambda_S} \vee \frac{2r\|W^T E_{\text{sym}} W\|_\infty}{\Delta_S} \le R_1, \qquad \frac{\sqrt{2r}\|E\|}{\sqrt{\Delta_S \lambda_S}} \le R_2, \qquad R_1 \vee R_2 \le R \le \frac{1}{8}. \tag{62}$$

Additionally, let R_3 be a real number such that

$$\frac{2r}{\Delta_S \sigma_S} \max_{|i-j| \ne 0, r} \left| w_i E_{\text{sym}}^2 w_j \right| \le R_3 \le R_2^2. \tag{63}$$

Lemma 3.6. *Under Setting 3.1 and Assumption 3.5, for $\mathcal{T}_\nu^{(\gamma)}$ defined in Eq. (59), we have*

$$\begin{aligned}\left|\mathcal{T}_\nu^{(\gamma)}\right| &\le r\mathrm{L}_\nu \left(\mu\mu' \mathbf{1}\{\gamma \le 10\log(m+n)\} + \|M\|\|M'\|\mathbf{1}\{\gamma > 10\log(m+n)\}\right) \\ &\quad \cdot \left[9R_1(6R)^{\gamma-1} + \mathbf{1}\{\gamma \text{ even}\}\left(4(R_1\sqrt{27/2})^\gamma + 27R_3(R_2\sqrt{27/4})^{\gamma-2}\right)\right].\end{aligned}$$

Proof. Let us consider the case $\gamma \le 10\log(m+n)$ first. Then we have $\alpha_0 \vee \alpha_h \le 10\log(m+n)$, so the bound (61) holds. Lemma 3.3 gives

$$\|\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})\| \le \mathrm{L}_\nu \binom{\gamma+\beta-2}{\beta-1} \frac{(2r)^\beta \|W^T E_{\text{sym}} W\|_\infty^{\beta-h} \|E\|^{\alpha-\alpha_0-\alpha_h+h-1}}{\lambda_S^{\gamma+1-\beta} \Delta_S^{\beta-1}} \mu\mu' \mathcal{K}_0^{\alpha_0+\alpha_h}.$$

Again, we temporarily define $X := W^T E_{\text{sym}} W$ for convenience. Rearranging the terms, we have

$$\|\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})\| \le 2r\mathrm{L}_\nu \mu\mu' \binom{\gamma+\beta-2}{\beta-1} \left[\frac{\|E\|}{\lambda_S}\right]^{\gamma_1} \left[\frac{\mathcal{K}_0}{\lambda_S}\right]^{\alpha_0} \left[\frac{\mathcal{K}_0'}{\lambda_S}\right]^{\alpha_h} \left[\frac{2r\|X\|_\infty}{\Delta_S}\right]^{\gamma_2} \left[\frac{\sqrt{2r}\|E\|}{\sqrt{\Delta_S \lambda_S}}\right]^{\gamma_3}, \tag{64}$$

where we temporarily define the following:

$$\begin{aligned}&\gamma_1 = \gamma_1(\boldsymbol{\alpha}) := \alpha_1 + \ldots + \alpha_{h-1} - (h-1), \\ &\gamma_2 = \gamma_2(\boldsymbol{\alpha}) := \alpha_0 + \alpha_h, \quad \gamma_2 = \gamma_2(\boldsymbol{\beta}) := \beta - h, \quad \gamma_3 = \gamma_3(h) := 2(h-1).\end{aligned}$$

Since R_1 upper bounds the former four powers and R_2 upper bounds the latter, we get

$$\begin{aligned}\|\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})\| &\le 2r\mathrm{L}_\nu \mu\mu' \binom{\gamma+\beta-2}{\beta-1} R_1^{\gamma_1(\boldsymbol{\alpha})+\alpha_0+\alpha_h+\gamma_2(\boldsymbol{\beta})} R_2^{\gamma_3(h)} \\ &= 2r\mathrm{L}_\nu \mu\mu' \binom{\gamma+\beta-2}{\beta-1} R_1^{\gamma-2h+2} R_2^{2h-2}.\end{aligned}$$

Plugging this bound into Eq. (59), we get

$$\begin{aligned}\left\|\mathcal{T}_\nu^{(\gamma)}\right\| &\le 2r\mathrm{L}_\nu \mu\mu' \sum_{h=1}^{\lfloor \gamma/2 \rfloor + 1} \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \Pi_h(\gamma)} \binom{\gamma+\beta-2}{\beta-1} R_1^{\gamma-2h+2} R_2^{2h-2} \\ &\le 2r\mathrm{L}_\nu \mu\mu' \sum_{h=1}^{\lfloor \gamma/2 \rfloor + 1} \sum_{\beta=h}^{\gamma+2-h} \binom{\gamma+\beta-2}{\beta-1} R_1^{\gamma-2h+2} R_2^{2h-2} \left|\Pi_h(\gamma, \beta)\right|,\end{aligned} \tag{65}$$

where

$$\Pi_h(\gamma, \beta) := \{(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \Pi_h(\gamma) : \beta_1 + \ldots + \beta_h = \beta\}$$

An element of this set is just a tuple $(\alpha_0, \ldots, \alpha_h, \beta_1, \ldots, \beta_h)$ such that

$$\beta_1, \ldots, \beta_h \geq 1, \quad \sum_{i=1}^{h} \beta_i = \beta, \quad \text{and} \quad \alpha_0, \alpha_h \geq 0, \quad \alpha_1, \ldots, \alpha_{h-1}, \quad \sum_{i=0}^{h} \alpha_i = \gamma + 1 - \beta.$$

The number of ways to choose such a tuple is

$$|\{(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \Pi_h(\gamma) : \beta_1 + \ldots + \beta_h = \beta\}| = \binom{\beta - 1}{h - 1}\binom{\gamma + 2 - \beta}{h}.$$

Plugging into Eq. (65), we obtain

$$\begin{aligned} \left|\mathcal{T}_\nu^{(\gamma)}\right| &\leq 2r\mathrm{L}_\nu \mu\mu' \sum_{h=1}^{\lfloor \gamma/2 \rfloor + 1} \sum_{\beta=h}^{\gamma+2-h} \binom{\gamma+\beta-2}{\beta-1}\binom{\beta-1}{h-1}\binom{\gamma+2-\beta}{h} R_1^{\gamma-2h+2} R_2^{2h-2} \\ &= 2r\mathrm{L}_\nu \mu\mu' \sum_{\beta=1}^{\gamma+1} \binom{\gamma+\beta-2}{\beta-1} \sum_{h=1}^{\beta \wedge (\gamma+2-\beta)} \binom{\beta-1}{h-1}\binom{\gamma+2-\beta}{h} R_1^{\gamma-2h+2} R_2^{2h-2}. \end{aligned} \tag{66}$$

Consider two cases for h and γ:

1. $\gamma \geq 2h - 1$. Let $R := R_1 \vee R_2$. The contribution is at most:

$$\begin{aligned} &2r\mathrm{L}_\nu \mu\mu' \sum_{\beta=1}^{\gamma+1} \binom{\gamma+\beta-2}{\beta-1} \sum_{h=1}^{\beta \wedge (\gamma+2-\beta)} \binom{\beta-1}{h-1}\binom{\gamma+2-\beta}{h} R_1 R^{\gamma-1} \\ &\leq 2r\mathrm{L}_\nu \mu\mu' R_1 R^{\gamma-1} \sum_{\beta=1}^{\gamma+1} \binom{\gamma+\beta-2}{\beta-1}\binom{\gamma+1}{\beta} \leq 2r\mathrm{L}_\nu \mu\mu' R_1 R^{\gamma-1} \sum_{\beta=1}^{\gamma+1} \binom{\gamma+1}{\beta} 2^{\gamma+\beta-2} \\ &\leq 2r\mathrm{L}_\nu \mu\mu' R_1 R^{\gamma-1} 2^{\gamma-2} 3^{\gamma+1} = 9r\mathrm{L}_\nu \mu\mu' R_1 (6R)^{\gamma-1}. \end{aligned}$$

2. $\gamma = 2h - 2$. This can only happens if γ is even. Then $\alpha_0 = \alpha_h = 0$ and $\alpha_1 = \ldots = \alpha_{h-1} = \beta_1 = \ldots = \beta_h = 1$. Let $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ denote the corresponding tuple. The previous computations are bad for this case, so we will go back to the definition to bound $\mathcal{T}_\nu(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$. Plugging into Eq. (50) and simplifying, we have

$$\mathcal{T}_\nu(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \sum_{\mathbf{I} \in [2r]^h} \mathcal{C}_\nu(\mathbf{I}) \left(M^T w_{i_1}\right) \left(w_{i_{h\beta_h}}^T M'\right) \prod_{k=1}^{h-1} w_{i_k}^T E_{\text{sym}}^2 w_{i_{k+1}},$$

where each block of consecutive indices in $\mathbf{I}$ now consists of only one index, so we can just denote $\mathbf{I} = (i_1, i_2, \ldots, i_h)$. Temporarily let $Y \in \mathbb{R}^{(m+n)\times(m+n)}$ be a matrix such that $Y_{ij} = w_i^T E_{\text{sym}}^2 w_j$ for $|i - j| \in \{0, r\}$ and 0 otherwise, and let $\lambda_+(\mathbf{I}) := (|\lambda_{i_k}|)_k$ for $\mathbf{I} = (i_k)_k$. We further consider two subcases for $\mathbf{I}$:

2.1. $\lambda_+(\mathbf{I})$ is non-uniform, i.e. there is k so that $|\lambda_{i_k}| \neq |\lambda_{i_{k+1}}|$, meaning $|i_k - i_{k+1}| \notin \{0, r\}$. Then $\left|w_{i_k}^T E_{\text{sym}}^2 w_{i_{k+1}}\right| \leq \|Y\|_\infty$. The rest of the product at the end can be bounded by $\|E\|^2$. The total contribution of this subcase is at most

$$2r\mathrm{L}_\nu \mu\mu' \binom{\gamma+\beta-2}{\beta-1} R_2^{2h-4} \frac{2r\|Y\|_\infty}{\lambda_S \Delta_S} \leq 2r\mathrm{L}_\nu \mu\mu' \binom{3\gamma/2}{\gamma/2} R_3 R_2^{\gamma-2},$$

where we use the same computations leading up to Eq. (66), noting that $h = \beta = \gamma/2+1$.

2.2. $\lambda_+(\mathbf{I})$ is uniform, i.e. $\mathbf{I} = (i_k)_{k=1}^h$ where each $i_k \in \{i, i+r\}$ for some $i \in [r]$. If $i \notin S$, then $\mathcal{C}_\nu(\mathbf{I}) = 0$. Suppose $i \in S$, we can apply Eq. (52) in Lemma 3.2, noting that $\beta_S(\mathbf{I}) = \beta$, $\beta_{S^c}(\mathbf{I}) = 0$, and $\lambda_S(\mathbf{I}) = \Delta_S(\mathbf{I}) = \lambda_i$ in this case, to get

$$|\mathcal{C}_\nu(\mathbf{I})| \le \mathrm{L}_\nu \binom{\gamma+\beta-2}{\beta-1} \frac{1}{\lambda_S(\mathbf{I})^{\gamma+1-\beta}\Delta_S(\mathbf{I})^{\beta-1}} = \mathrm{L}_\nu \binom{\gamma+h-2}{h-1} \frac{1}{\lambda_i^\gamma}.$$

Therefore the total contribution of this subcase is at most

$$\mu\mu' \binom{\gamma+h-2}{h-1} \sum_{i\in S} \sum_{\mathbf{I}\in\{i,i+1\}^h} \frac{|w_i^T E_{\text{sym}}^2 w_i|^{h-1}}{\lambda_i^\gamma} \le s\mu\mu' \binom{3\gamma/2}{\gamma/2} \frac{2^h \|E\|^{2(h-1)}}{\lambda_S^\gamma}$$
$$\le 2r\mathrm{L}_\nu\mu\mu' \binom{3\gamma/2}{\gamma/2} \frac{2^{\gamma/2}\|E\|^\gamma}{\lambda_S^\gamma} \le 2r\mathrm{L}_\nu\mu\mu' \binom{3\gamma/2}{\gamma/2} (R_1\sqrt{2})^\gamma.$$

Therefore, the contribution of the case $h = \gamma/2 + 1$ is at most

$$2r\mathrm{L}_\nu\mu\mu' \binom{3\gamma/2}{\gamma/2} \left[R_3 R_2^{\gamma-2} + (R_1\sqrt{2})^\gamma\right] \le 4r\mathrm{L}_\nu\mu\mu' \left[\frac{27}{4} R_3 \left(R_2\sqrt{27/4}\right)^{\gamma-2} + \left(R_1\sqrt{27/2}\right)^\gamma\right],$$

since $R_3 \ge 2r\|Y\|/(\lambda_S\Delta_S)$.

Summing up the contributions from both cases, we obtain

$$\left\|\mathcal{T}_\nu^{(\gamma)}\right\| \le r\mathrm{L}_\nu\mu\mu' \left[9R_1(6R)^{\gamma-1} + \mathbf{1}\{\gamma \text{ even}\}\left(4(R_1\sqrt{27/2})^\gamma + 27R_3(R_2\sqrt{27/4})^{\gamma-2}\right)\right]. \tag{67}$$

Now let us consider the case $\gamma > 10\log(m+n)$. Instead of using the bound in Eq. (61), we can simply use the naive bounds

$$\sum_{i=1}^{2r} \|w_i^T E_{\text{sym}}^\alpha M\| \le 2r\|E\|^\alpha\|M\| \quad \text{and} \quad \sum_{i=1}^{2r} \|w_i^T E_{\text{sym}}^\alpha M'\| \le 2r\|E\|^\alpha\|M'\|.$$

Plugging in these bounds into the previous computations has the same effect as using $\mu = \|M\|$, $\mu' = \|M'\|$, and $\mathcal{K} = \mathcal{K}' = \|E\|$. The conditions on R_1, R_2 and R_3 remain the same so they still hold. Thus we still have Eq. (67), with the substitutions above. The proof is complete. □

Lemma 3.7. *Under Setting 3.1 and Assumption 3.5, for $\mathcal{T}_\nu$ defined in Eq. (59), we have*

$$\|\mathcal{T}_\nu\| \le 4r\mathrm{L}_\nu\mu\mu'(18R_1 + 27R_3) + r\mathrm{L}_\nu\|M\|\|M'\|(m+n)^{-2.5}.$$

Proof. For convenience, let $N = \lfloor 10\log(m+n) \rfloor$. Applying Lemma 3.6, we have

$$\sum_{\gamma=1}^N \left|\mathcal{T}_\nu^{(\gamma)}\right| \le r\mathrm{L}_\nu\mu\mu' \left[9\sum_{\gamma=1}^\infty R_1(6R)^{\gamma-1} + \sum_{\gamma=1}^\infty \left(4(R_1\sqrt{27/2})^{2\gamma} + 27R_3(R_2\sqrt{27/4})^{2\gamma-2}\right)\right]$$
$$\le r\mathrm{L}_\nu\mu\mu' \left[\frac{9R_1}{1-6R} + \frac{108R_1^2}{2-27R_1^2} + \frac{108R_3}{4-27R_2^2}\right] \le 4r\mathrm{L}_\nu\mu\mu'(18R_1 + 27R_3),$$

and

$$\begin{aligned}\sum_{\gamma=N}^{\infty}\left|\mathcal{T}_\nu^{(\gamma)}\right| &\le r\mathrm{L}_\nu\|M\|\|M'\|\left[9\sum_{\gamma=N}^{\infty}R_1(6R)^{\gamma-1}+\sum_{\gamma=\lceil N/2\rceil}^{\infty}\left(4\left(\tfrac{R_1\sqrt{27}}{\sqrt{2}}\right)^{2\gamma}+27R_3\left(\tfrac{R_2\sqrt{27}}{2}\right)^{2\gamma-2}\right)\right]\\ &\le r\mathrm{L}_\nu\|M\|\|M'\|\left[\frac{9R_1(6R)^{N-1}}{1-6R}+\frac{4(27R_1^2/2)^{\lceil N/2\rceil}}{1-27R_1^2/2}+\frac{27R_3(27R_2^2/4)^{\lceil N/2\rceil-1}}{1-27R_2^2/4}\right]\\ &\le \frac{r\mathrm{L}_\nu\|M\|\|M'\|}{1-6R}\left[2(6R)^N+4\left(\frac{3}{8}\right)^{N/2}+\frac{3}{4}\left(\frac{3}{16}\right)^{N/2-1}\right]\le\frac{r\mathrm{L}_\nu\|M\|\|M'\|}{(m+n)^{2.5}}.\end{aligned}$$

The proof is complete. □

3.4 Conclusion: Proof of Theorem 2.2

We are now ready to use the common three-step strategy above to prove Theorem 2.2.

Proof of Theorem 2.2. Consider the objects defined in Theorem 2.2 and the additional objects in Setting 3.1. Note that $\lambda_i = \sigma_i$ for $i \in [r]$ and $-\sigma_{i-r}$ for $i \in [[r+1, 2r]]$, and $\lambda_S = \sigma_S$ in this context.

Let us prove Eq. (22). Consider arbitrary $j, k \in [n]$. We choose $M = e_{m+n,j+m}$, $M' = e_{m+n,k+m}$ and $\nu = 0$, the expansion in Section 3.1 gives

$$\left(\tilde{V}_S\tilde{V}_S^T - V_SV_S^T\right)_{jk} = \left(\tilde{W}_S\tilde{W}_S^T - W_SW_S^T\right)_{(j+m)(k+m)} = \mathcal{T}_0.$$

We apply Lemma 3.7. Let us choose the parameters to satisfy Assumption 3.5. For $\nu = 0$, Lemma 3.2 guarantees $\mathrm{L}_\nu = 2$ satisfies Eq. (60). For the above choices of M and M', Lemma 3.4 guarantees that the choices $\mu = \mu' = \frac{1}{\sqrt{2}}\mu_{UV}$ and $\mathcal{K}_0 = \mathcal{K}_0' = \mathcal{K}$ satisfy Eq. (61).

For convenience, define the following shorthands:

$$R_1 := \frac{\|E\|}{\sigma_S} \vee \frac{2r\|U^TEV\|_\infty}{\Delta_S}, \qquad R_2 := \frac{\sqrt{2r}\|E\|}{\sqrt{\sigma_S\Delta_S}}, \qquad R_3 := \frac{r\max_{i\ne j}(|u_iEE^Tu_j| + |v_iE^TEv_j|)}{\Delta_S\sigma_S}. \tag{68}$$

Next, we show that the terms R_1, R_2 and R_3 satisfy Eqs. (62) and (63). It suffices to check the guarantees on R_1 and R_3. The former follows from the fact $\|W^TE_{\text{sym}}W\|_\infty \le \|U^TEV\|_\infty$, which follows from $w_iE_{\text{sym}}w_j = (u_i^TEv_j + u_j^TEv_i)/2$. For the latter, fix arbitrary $i, j \in [m+n]$ such that $0 < i - j$. There are 3 cases:

- $j \in [r]$ and $i \in [[r+1, 2r]] \setminus \{j+r\}$. Then $w_i^TE_{\text{sym}}^2w_j = \frac{1}{2}(u_{i-r}^TEE^Tu_j - v_{i-r}^TE^TEv_j)$.
- $j \in [r]$ and $i \in [r] \setminus \{j\}$. Then $w_i^TE_{\text{sym}}^2w_j = \frac{1}{2}(u_i^TEE^Tu_j + v_i^TE^TEv_j)$.
- $j \in [[r+1, 2r]]$ and $i \in [[r+1, 2r]] \setminus \{j\}$. Then $w_i^TE_{\text{sym}}^2w_j = \frac{1}{2}(u_{i-r}^TEE^Tu_{j-r} + v_{i-r}^TE^TEv_{j-r})$.

From the above, it follows that

$$\max_{|i-j|\ne 0,r}|w_i^TE_{\text{sym}}^2w_j| \le \frac{1}{2}\max_{i\ne j}(|u_iEE^Tu_j| + |v_iE^TEv_j|),$$

proving Eq. (63) for this choice of R_3. Plugging in these values into the bound in Lemma 3.7 gives

$$\left|\left(\tilde{V}_S\tilde{V}_S^T - V_SV_S^T\right)_{jk}\right| \le 4r\mu_{UV}^2(18R_1 + 27R_3) + \frac{2r\|M\|\|M'\|}{(m+n)^{2.5}} \le C_0r\mu_{UV}^2(R_1 + R_3) + \frac{1}{m+n},$$

for a constant C_0. The above holds uniformly over all $j, k \in [n]$, so it holds for the infinity norm, proving Eq. (22).

Let us prove Eq. (23). Consider an arbitrary $j \in [n]$. We choose $\nu = 0$, $M = e_{m+n,j+m}$ and $M' = I_{m+n}$. The expansion in Section 3.1 gives

$$\left(\tilde{V}_S\tilde{V}_S^T - V_SV_S^T\right)_{j,\cdot} = \left(\tilde{W}_S\tilde{W}_S^T - W_SW_S^T\right)_{(j+m),\cdot} = \mathcal{T}_0,$$

for $M = e_{m+n,j+m}$, $M' = I_{m+n}$, and $\mathcal{C}$ from Eq. (47). From the previous argument, we can choose L_ν, μ and $\mathcal{K}_0$ the same way, and choose $\mu' = 1$ and $\mathcal{K}_0' = \|E\|$, which trivially satisfy the condition on M'. The same choice of R_1, R_2 and R_3 also satisfy Eqs. (62) and (63). Therefore

$$\left\|\left(\tilde{V}_S\tilde{V}_S^T - V_SV_S^T\right)_{j,\cdot}\right\| \le C_0 r\mu_{UV}(R_1 + R_3) + \frac{1}{m+n},$$

which holds uniformly over $j \in [n]$, proving Eq. (23)..

Let us prove Eq. (24). Consider arbitrary $j \in [m]$ and $k \in [n]$. Choose $M = e_{m+n,j}$, $M' = e_{m+n,k+m}$ and $\nu = 1$. The expansion in Section 3.1 gives

$$\left(\tilde{A}_s\tilde{A}_s^T - A_sA_s^T\right)_{jk} = \left(\tilde{W}_S\tilde{W}_S^T - W_SW_S^T\right)_{j(k+m)} = \mathcal{T}_1.$$

We choose $\mathrm{L}_\nu = \lambda_s = \sigma_s$, $\mu = \frac{1}{\sqrt{2}}\mu_{VU}$, $\mu' = \frac{1}{\sqrt{2}}\mu_{UV}$, $\mathcal{K}_0 = \mathcal{K}$. The choices of R_1, R_2 and R_3 are the same as those in the theorem statement. Similarly to the previous two parts, these choices satisfy all requirements of Assumption 3.5 and Lemma 3.7, so we have

$$\left|\left(\tilde{A}_s\tilde{A}_s^T - A_sA_s^T\right)_{jk}\right| \le C_0 r\sigma_s\mu_{UV}\mu_{VU}(R_1 + R_3) + \frac{1}{m+n},$$

for a constant C_0. This bound holds uniformly over $j \in [m]$ and $k \in [n]$, so it holds in the infinity norm. The proof is complete. □

4 Proof of main results: the random case

In this section, we prove Theorem 2.3. Our job is to replace the terms that depend on E, the random noise matrix, with deterministic terms that upper bound them with high probability, then apply Theorem 2.2. These terms are:

- $\|E\|$. There are tight bounds in the literature. For E following the Model (26), with the assumption $K \le (m+n)^{1/2}\log^{-5}(m+n)$, the moment argument in [28] can be used.
- $\|U^TEV\|_\infty = \max_{i,j}|u_i^TEv_j|$. These terms can be bounded with a simple Chernoff bound.
- $\max_{i\neq j}(|u_iEE^Tu_j| + |v_iE^TEv_j|)$. These terms can be bounded with the moment method. The most saving occurs when E is a stochastic matrix, meaning its row norms and column norms have the same second moment. For our purpose, the naive bound $2\|E\|^2$ suffices.
- μ_{UV}, μ_{VU} and $\mathcal{K}$ for the powers of E. We will use the moment method, with walk-counting, to bound these terms.

We summarize the first three in the lemma below.

Lemma 4.1. *Consider the objects in Setting 2.1. Let $E \in \mathbb{R}^{m\times n}$ be a random matrix satisfying Model (26) with parameters K and κ. Suppose $K \le (m+n)^{1/2}\log^{-3}(m+n)$. Then with probability $1 - O((m+n)^{-2})$, all of the following hold:*

$$\|E\| \le 2\kappa\sqrt{m+n}, \tag{69}$$

$$\max_{i\neq j}(|u_i E E^T u_j| + |v_i E^T E v_j|) \le 2\|E\|^2 \le 8\kappa^2(m+n). \tag{70}$$

$$\|U^T E V\|_\infty = \max_{i,j}|u_i^T E v_j| \le 2\kappa\log^{1/2}(m+n). \tag{71}$$

Proof. Eq. (69) follows from the moment argument in [28]. Eq. (70) follows from Eq. (69). Eq. (71) follows from the Hoeffding inequality [16]. □

The last bound is technically heavy, so we will dedicate a separate lemma for it, whose proof is postponed to Section 5.2.

Lemma 4.2. *Let K and κ be positive real numbers and E be a $m \times n$ random matrix with independent entries following Model (26) with parameters K and κ. Let $E_{\text{sym}} := \left(\begin{smallmatrix} 0 & E \\ E^T & 0\end{smallmatrix}\right)$ be the symmetrization of E. For each $p > 0$, define*

$$D_{U,V,p} := \frac{Kp^3\|U\|_{2,\infty}}{\sqrt{r(m+n)}} + \frac{p^{3/2}}{\sqrt{m+n}} + \frac{p\|V\|_{2,\infty}}{\sqrt{r}}. \tag{72}$$

There are universal constants C and c such that, for any $t, \varepsilon > 0$, if $K \le ct^{-2}\log^{-2}(m+n)\sqrt{m+n}$, then for each fixed $k \in [n]$, with probability $1 - \log^{-\Omega(1)}(m+n)$,

$$\max_{0\le\alpha\le t\log(m+n)} \frac{\|e_{n,k}^T(E^TE)^a E^T U\|}{(1.9\kappa\sqrt{m+n})^{2a+1}} \vee \frac{\|e_{n,k}^T(E^TE)^a V\|}{(1.9\kappa\sqrt{m+n})^{2a}} \le C\sqrt{r}D_{U,V,\log\log(m+n)}. \tag{73}$$

If the stronger bound $K \le ct^{-2}\log^{-5}(m+n)\sqrt{m+n}$ holds, then with probability $1 - O((m+n)^{-2})$,

$$\max_{0\le\alpha\le t\log(m+n)} \frac{\|e_{n,k}^T(E^TE)^a E^T U\|_{2,\infty}}{(1.9\kappa\sqrt{m+n})^{2a+1}} \vee \frac{\|e_{n,k}^T(E^TE)^a V\|_{2,\infty}}{(1.9\kappa\sqrt{m+n})^{2a}} \le C\sqrt{r}D_{U,V,\log(m+n)}. \tag{74}$$

We only use Eq. (74) to prove Theorem 2.3, but for completeness, we still include (73), which has a better bound at the cost of not being uniform over all $k \in [n]$.

Let us prove the main theorem using these lemmas.

Proof of Theorem 2.3. Consider the objects from Setting 2.1. We aim to apply Theorem 2.2.

Firstly, we want to choose R_1, R_2 and R_3 to satisfy (21). Let R_1 and R_2 be given from hypothesis, and $R_3 := R_2^2$. Lemma 4.1 then guarantees (21) with probability $1 - O((m+n)^{-1})$.

Secondly, we want to choose μ_{UV}, μ_{VU} and $\mathcal{K}$ to satisfy (20). There are terms of the same name in Theorem 2.3, so we denote them by μ'_{UV} and μ'_{VU} respectively. Note that $\mu'_{UV} = \mu_{U,V,\log(m+n)}$ from Lemma 4.2. Let C be the constant from that lemma, then $\mu_{UV} := C\mu'_{UV}$ and $\mathcal{K} := 2\kappa\sqrt{m+n}$ satisfy the first half of Eq. (20) with probability $1 - O((m+n)^{-2})$. By symmetry, $\mu_{VU} := C\mu'_{VU}$ and the same $\mathcal{K}$ satisfy the second half, also with probability $1 - O((m+n)^{-2})$.

We can now apply Theorem 2.2. Plugging each of the choices above into their corresponding places in the right-hand sides of Eqs. (22), (23) and (24), we get the desired bounds with probability $1 - O((m+n)^{-1})$. □

5 Proofs of technical lemmas

5.1 Proof of bound for contour integrals of polynomial reciprocals

In this section, we prove Lemma 3.2, which provides the necessary bounds on the integral coefficients to advance the second step of the main proof (Section 3.2). Recall that the integrals we are interested in have the form

$$\mathcal{C}_\nu(\mathbf{I}) := \oint_{\Gamma_S} \frac{z^\nu \mathrm{d}z}{2\pi i} \frac{1}{z^{\gamma+1}} \prod_{k=1}^{\beta} \frac{\lambda_{i_k}}{z-\lambda_{i_k}}, \qquad \text{where } \nu \in \{0,1\} \text{ and } \beta \leq \gamma+1. \tag{75}$$

Let the multiset $\{\lambda_{i_k}\}_{k\in[\beta]} = A \cup B$, where $A := \{a_i\}_{i\in[l]}$ and $B := \{b_j\}_{j\in[k]}$, where each $a_i \in S$ and each $b_j \notin S$, having multiplicities m_i and n_j respectively. We can rewrite the above into

$$\mathcal{C}_\nu(\mathbf{I}) = \prod_{i=1}^{l} a_i^{m_i} \prod_{j=1}^{k} b_j^{n_j} C(n_0; A, \mathbf{m}; B, \mathbf{n}), \tag{76}$$

where

$$C(n_0; A, \mathbf{m}; B, \mathbf{n}) := \oint_{\Gamma_A} \frac{\mathrm{d}z}{2\pi i} \frac{1}{z^{n_0}} \prod_{j=1}^{k} \frac{1}{(z-b_j)^{n_j}} \prod_{i=1}^{l} \frac{1}{(z-a_i)^{m_i}}, \tag{77}$$

where $n_0 = \gamma + 1 - \nu$. The m_i's and n_j's satisfy $\sum_i m_i + \sum_j n_j \leq \gamma + 1$. We can remove the set S and simply denote the contour by Γ_A without affecting its meaning. The next three results will build up the argument to bound these sums and ultimately prove the target lemmas.

Lemma 5.1. *Let $A = \{a_i\}_{i\in[l]}$ and $B = \{b_j\}_{j\in[k]}$ be disjoint set of complex non-zero numbers and $\mathbf{m} = \{m_i\}_{i\in[l]}$ and n_0 and $\mathbf{n} = \{n_j\}_{j\in[k]}$ be nonnegative integers such that $m + n + n_0 \geq 2$, where $m = \sum_i m_i$ and $n := \sum_{i\geq 1} n_i$. Let Γ_A be a contour encircling all numbers in A and none in $B \cup \{0\}$. Let $a, d > 0$ be arbitrary such that:*

$$d \leq a, \qquad a \leq \min_i |a_i|, \qquad d \leq \min_{i,j} |a_i - b_j|. \tag{78}$$

Suppose that $0 \leq m_i' \leq m_i$ for each $i \in [l]$ and that $m' := \sum_{i=1}^{k} m_i' \leq n_0$. Then for $C(n_0; A, \mathbf{m}; B, \mathbf{n})$ defined Eq. (77), we have

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m+n+n_0-2}{m-1} \frac{1}{a^{n_0-m'} d^{m+n-1}} \prod_{i=1}^{l} \frac{1}{|a_i|^{m_i'}} \tag{79}$$

Proof. Firstly, given the sets A and B and the notations and conditions in Lemma 5.1, the weak bound below holds

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m+n+n_0-2}{m-1} \frac{1}{d^{m+n+n_0-1}}. \tag{80}$$

We omit the details of the proof, which is a simple induction argument. We now use Eq. (80) to prove the following:

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m+n+n_0-2}{m-1} \frac{1}{a^{n_0} d^{m+n-1}}. \tag{81}$$

We proceed with induction. Let $P_1(N)$ be the following statement: "For any sets A and B, and the notations and conditions described in Lemma 5.1, such that $m + n + n_0 = N$, Eq. (81) holds."

Since $m + n + n_0 \geq 2$, consider $N = 2$ for the base case. The only case where the integral is non-zero is when $m = 1$ and $n + n_0 = 1$, meaning $A = \{a_1\}$, $m_1 = 1$ and either $B = \varnothing$ and $n_0 = 1$, or $B = \{b_1\}$ and $n_1 = 1$, $n_0 = 0$. The integral yields a_1^{-1} in the former case and $(a_1 - b_1)^{-1}$ in the latter, confirming the inequality in both.

Consider $n \geq 3$ and assume $P_1(n-1)$. If $m = 0$, the integral is again 0. If $n_0 = 0$, Eq. (81) automatically holds by being the same as Eq. (80). Assume $m, n_0 \geq 1$. There must then be some $i \in [l]$ such that $m_i \geq 1$, without loss of generality let 1 be that i. We have

$$C(n_0; A, \mathbf{m}; B, \mathbf{n}) = \frac{1}{a_1}\Big[C(n_0 - 1; A, \mathbf{m}; B, \mathbf{n}) - C(n_0; A, \mathbf{m}^{(1)}; B, \mathbf{n})\Big] \tag{82}$$

where $\mathbf{m}^{(i)}$ is the same as $\mathbf{m}$ except that the i-entry is $m_i - 1$.

Consider the first integral on the right-hand side. Applying $P_1(N-1)$, we get

$$|C(n_0 - 1; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m + n + n_0 - 3}{m - 1} \frac{1}{a^{n_0 - 1} d^{m+n-1}}. \tag{83}$$

Analogously, we have the following bound for the second integral:

$$\left|C(n_0; A, \mathbf{m}^{(1)}; B, \mathbf{n})\right| \leq \binom{m + n + n_0 - 3}{m - 2} \frac{1}{a^{n_0} d^{m+n-2}} \leq \binom{m + n + n_0 - 3}{m - 2} \frac{1}{a^{n_0 - 1} d^{m+n-1}}. \tag{84}$$

Notice that the binomial coefficients in Eqs. (83) and (84) sum to the binomial coefficient in Eq. (81), we get $P_1(N)$, which proves Eq. (81) by induction.

Now we can prove Eq. (79). The logic is almost identical, with Eq. (81) playing the role of Eq. (80) in its own proof, handling an edge case in the inductive step. Let $P_2(n)$ be the statement: "For any sets A and B, and the notations and conditions described in Lemma 5.1, such that $m + n + n_0 = N$, Eq. (79) holds."

The cases $N = 1$ and $N = 2$ are again trivially true. Consider $N \geq 3$ and assume $P_2(N-1)$. Fix any sequence $m_1', m_2', \ldots, m_l'$ satisfying $0 \leq m_i' \leq m_i$ for each $i \in [k]$ and $n_0 \geq m_1' + \ldots + m_k'$. If $m_1' = m_2' = \ldots = m_k' = 0$, we are done by Eq. (81). By symmetry among the indices, assume $m_1' \geq 1$. This also means $n_0 \geq 1$. Consider Eq. (82) again. For the first integral on the right-hand side, applying $P_2(N-1)$ for the parameters $n_0 - 1, n_1, \ldots, n_k$, $m_1, \ldots, m_l$ and $m_1' - 1, m_2', \ldots, m_k'$ yields the bound

$$|C(n_0 - 1; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m + n + n_0 - 3}{m - 1} \frac{1}{a^{n_0 - m'} d^{m+n-1}} \frac{1}{|a_1|^{m_1' - 1}} \prod_{i=2}^{l} \frac{1}{|a_i|^{m_i'}}. \tag{85}$$

Applying $P_2(N-1)$ for the parameters $n_0, n_1, \ldots, n_k$, $m_1 - 1, \ldots, m_l$ and $m_1' - 1, m_2', \ldots, m_k'$, we get the following bound for the second integral on the right-hand side of Eq. (82):

$$\begin{aligned} \left|C(n_0; A, \mathbf{m}^{(1)}; B, \mathbf{n})\right| &\leq \binom{m + n + n_0 - 3}{m - 2} \frac{1}{a^{n_0 - m' + 1} d^{m+n-2}} \frac{1}{|a_1|^{m_1' - 1}} \prod_{i=2}^{l} \frac{1}{|a_i|^{m_i'}} \\ &\leq \binom{m + n + n_0 - 3}{m - 2} \frac{1}{a^{n_0 - m'} d^{m+n-1}} \frac{1}{|a_1|^{m_1' - 1}} \prod_{i=2}^{l} \frac{1}{|a_i|^{m_i'}}. \end{aligned}$$

Summing up the bounds by summing the binomial coefficients, we get exactly $P_2(N)$, so Eq. (79) is proven by induction. □

Lemma 5.2. *Let A, B, $\mathbf{m}$, $\mathbf{n}$, n_0, Γ_A and a, d be the same, with the same conditions as in Lemma 5.1. Suppose that $0 \leq m_i' \leq m_i$ and $0 \leq n_j' \leq n_j$ for each $i, j \geq 1$ and*

$$m' + n' \leq n_0 \quad \textit{for} \quad m' := \sum_i m_i', \quad n' := \sum_j n_j'.$$

Then for $C(n_0; A, \mathbf{m}; B, \mathbf{n})$ defined in Eq. (77), we have

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{n + n_0 - n' + m - 2}{m - 1} \frac{(1 + d/a)^{n'}}{a^{n_0 - m' - n'} d^{m+n-1}} \prod_{i=1}^{l} \frac{1}{|a_i|^{m_i'}} \prod_{j=1}^{k} \frac{1}{|b_j|^{n_j'}}. \tag{86}$$

Proof. We have the expansion

$$\begin{aligned} \frac{1}{z^{n_0}} \prod_{j=1}^{k} \frac{b_j^{n_j'}}{(z - b_j)^{n_j}} \prod_{i=1}^{l} \frac{1}{(z - a_i)^{m_i}} &= \frac{1}{z^{n_0 - n'}} \prod_{j=1}^{k} \frac{1}{(z - b_j)^{n_j - n_j'}} \prod_{j=1}^{k} \left(\frac{1}{z} - \frac{1}{z - b_j} \right)^{n_j'} \prod_{i=1}^{l} \frac{1}{(z - a_i)^{m_i}} \\ = \frac{1}{z^{n_0 - n'}} \prod_{j=1}^{k} \frac{1}{(z - b_j)^{n_j - n_j'}} &\sum_{0 \leq r_j \leq n_j' \forall j} \frac{(-1)^{r_1 + \ldots + r_k}}{z^{n' - r_1 - \ldots - r_k}} \prod_{j=1}^{k} \binom{n_j'}{r_j} \frac{1}{(z - b_j)^{r_j}} \prod_{i=1}^{l} \frac{1}{(z - a_i)^{m_i}} \\ = \sum_{0 \leq r_j \leq n_j' \forall j} \frac{(-1)^{r_1 + \ldots + r_k}}{z^{n_0 - r_1 - \ldots - r_k}} &\prod_{j=1}^{k} \binom{n_j'}{r_j} \frac{1}{(z - b_j)^{r_j + n_j - n_j'}} \prod_{i=1}^{l} \frac{1}{(z - a_i)^{m_i}}. \end{aligned}$$

Integrating both sides over Γ_A, we have

$$C(n_0; A, \mathbf{m}; B, \mathbf{n}) \prod_{j=1}^{k} b_j^{n_j'} = \sum_{0 \leq r_j \leq n_j' \forall j} (-1)^{\sum_j r_j} \binom{n_j'}{r_j} C \left(n_0 - \textstyle\sum_j r_j; A, \mathbf{m}; B, \mathbf{r} + \mathbf{n} - \mathbf{n}' \right),$$

where the j-entry of $\mathbf{r} + \mathbf{n} - \mathbf{n}'$ is simply $r_j + n_j - n_j'$. Applying Lemma 5.1 for each summand on the right-hand side and rearranging the powers, we get

$$\left| C \left(n_0 - \textstyle\sum_j r_j; A, \mathbf{m}; B, \mathbf{r} + \mathbf{n} - \mathbf{n}' \right) \right| \leq \binom{m + n + n_0 - n' - 2}{m - 1} \frac{(a/d)^{\sum_j r_j}}{a^{n_0 - m'} d^{n - n' + m - 1}} \prod_{i=1}^{l} \frac{1}{|a_i|^{m_i'}}.$$

Summing up the bounds, we get

$$\begin{aligned} \left| C(n_0; A, \mathbf{m}; B, \mathbf{n}) \prod_{j=1}^{k} b_j^{n_j'} \right| &\leq \binom{m + n + n_0 - n' - 2}{m - 1} \frac{\prod_{i=1}^{l} |a_i|^{-m_i'}}{a^{n_0 - m'} d^{n - n' + m - 1}} \sum_{0 \leq r_j \leq n_j' \forall j} \prod_{j=1}^{k} \binom{n_j'}{r_j} \frac{a^{r_j}}{d^{r_j}} \\ &= \binom{m + n + n_0 - n' - 2}{m - 1} \frac{\prod_{i=1}^{l} |a_i|^{-m_i'}}{a^{n_0 - m'} d^{n - n' + m - 1}} \left(\frac{a}{d} + 1 \right)^{n'}. \end{aligned}$$

Rearranging the term, we get precisely the desired inequality. □

The lemma above is the main ingredient in the proof of Lemma 3.2.

Proof of Eq. (52) of Lemma 3.2. First rewrite the integral into the forms of (75), then (76) and (77). Let us consider two cases for $\mathcal{C}$:

1. $\nu = 0$, so $n_0 = \gamma + 1$. Let $a = \lambda_S(\mathbf{I})$, $d = \delta_S(\mathbf{I})$, $m = \beta_S(\mathbf{I})$, $n = n' = \beta_{S^c}(\mathbf{I})$, $m'_i = m_i$ and $n'_j = n_j$ for all i, j, then $m' + n' = \beta \leq \gamma + 1 = n_0$, so we can apply Lemma 5.2 to get
$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{n_0 + m - 2}{m - 1} \frac{(1 + d/a)^{n'}}{a^{n_0 - m - n} d^{m+n-1}} \prod_{i=1}^{l} \frac{1}{|a_i|^{m_i}} \prod_{j=1}^{k} \frac{1}{|b_j|^{n_j}},$$
or equivalently,
$$|\mathcal{C}_0(\mathbf{I})| \leq \left(1 + \frac{\Delta_S(\mathbf{I})}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma + \beta_S(\mathbf{I}) - 1}{\beta_S(\mathbf{I}) - 1} \frac{1}{\lambda_S(\mathbf{I})^{\gamma+1-\beta} \Delta_S(\mathbf{I})^{\beta-1}}.$$
Note that since $\beta_S(\mathbf{I}) \leq \beta \leq \gamma + 1$ and $\mathrm{L}_0 = 2$,
$$\binom{\gamma + \beta_S(\mathbf{I}) - 1}{\beta_S(\mathbf{I}) - 1} = \frac{\gamma + \beta_S(\mathbf{I}) - 1}{\gamma} \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \leq \mathrm{L}_0 \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1},$$
so we can replace the former with the latter to the product on the right-hand side to get an upper bound, which is also the desired bound.

2. $\nu = 1$, so $n_0 = \gamma$. Without loss of generality, assume $|a_1| = \lambda_S(\mathbf{I})$, then we are guaranteed $m_1 \geq 1$. Applying Lemma 5.2 for the same parameters as in the previous case, except that $m'_1 = m_1 - 1$, we get
$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq |a_1| \binom{n_0 + m - 2}{m - 1} \frac{(1 + d/a)^{n'}}{a^{n_0 - m + 1 - n} d^{m+n-1}} \prod_{i=1}^{l} \frac{1}{|a_i|^{m_i}} \prod_{j=1}^{k} \frac{1}{|b_j|^{n_j}},$$
which translates to
$$|\mathcal{C}_1(\mathbf{I})| \leq \lambda_S(\mathbf{I}) \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \left(1 + \frac{\Delta_S(\mathbf{I})}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})} \frac{1}{\lambda_S(\mathbf{I})^{\gamma+1-\beta} \Delta_S(\mathbf{I})^{\beta-1}},$$
which is the desired bound since $\mathrm{L}_1 = \lambda_S(\mathbf{I})$.

Let us now prove Eq. (53). We can assume $\beta_S(\mathbf{I}) \geq 1$, since the integral is 0 otherwise, making the inequality trivial. It suffices to show that we can substitute $\lambda_S(\mathbf{I})$ with λ_S and $\Delta_S(\mathbf{I})$ with Δ_S in Eq. (52) to make the right-hand side larger. Let us again split into the cases as above.

1. $\nu = 0$. We have
$$\frac{\left(1 + \frac{\Delta_S(\mathbf{I})}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})}}{\lambda_S(\mathbf{I})^{\gamma+1-\beta} \Delta_S(\mathbf{I})^{\beta-1}} = \frac{\left(\frac{1}{\Delta_S(\mathbf{I})} + \frac{1}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})}}{\lambda_S(\mathbf{I})^{\gamma+1-\beta} \Delta_S(\mathbf{I})^{\beta_S(\mathbf{I})-1}}.$$
From this new form, it is evident that the right-hand side will increase if we make the aforementioned substitutions, since $\lambda_S \leq \lambda_S(\mathbf{I})$ and $\Delta_S \leq \Delta_S(\mathbf{I})$.

2. $\nu = 1$ and additionally, $S = [s]$ for some $s \in [r]$. If $\beta \leq \gamma$, the rewriting in the previous case works in the same way. Suppose $\beta = \gamma + 1$, we now write
$$\frac{\lambda_S(\mathbf{I}) \left(1 + \frac{\Delta_S(\mathbf{I})}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})}}{\lambda_S(\mathbf{I})^{\gamma+1-\beta} \Delta_S(\mathbf{I})^{\beta-1}} = \frac{1}{\lambda_S(\mathbf{I})^{\gamma-1}} \left(\frac{\lambda_S(\mathbf{I})}{\Delta_S(\mathbf{I})}\right)^{\beta_S(\mathbf{I})-1} \left(1 + \frac{\lambda_S(\mathbf{I})}{\Delta_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})}.$$
Since $\lambda_S(\mathbf{I}) \leq \lambda_S = \lambda_s$, it suffices to show $\lambda_S(\mathbf{I})/\Delta_S(\mathbf{I}) \leq \lambda_s/\delta_s$ to make the substitution work as in the previous case. Choose $t \in [s]$ where $\lambda_t = \lambda_S(\mathbf{I})$, then $\Delta_S(\mathbf{I}) \geq \lambda_t - \lambda_{s+1}$, thus
$$\frac{\lambda_S(\mathbf{I})}{\Delta_S(\mathbf{I})} \leq \frac{\lambda_t}{\lambda_t - \lambda_{s+1}} \leq \frac{\lambda_s}{\lambda_s - \lambda_{s+1}} = \frac{\lambda_s}{\delta_s}.$$

Thus both Eqs. (52) and (53) hold. The proof is complete. □

5.2 Proof of semi-isotropic bounds for powers of random matrices

In this section, we prove Lemma 4.2, which gives semi-itrosopic bounds for powers of E_{sym} in the second step of the main proof strategy.

The form of the bound naturally implies that we should handle the even and odd powers separately. We split the two cases into the following lemmas.

Lemma 5.3. *Let $m, r \in \mathbb{N}$ and $U \in \mathbb{R}^{m \times r}$ be a matrix whose columns $u_1, u_2, \ldots, u_r$ are unit vectors. Let E be a $m \times n$ random matrix following Model (26) with parameters K and $\kappa = 1$, meaning E has independent entries and*

$$\mathbf{E}\left[E_{ij}\right] = 0, \quad \mathbf{E}\left[\|E\|_{ij}^2\right] \le 1, \quad \mathbf{E}\left[\|E\|_{ij}^p\right] \le K^{p-2} \quad \textit{for all } \ p.$$

For any $a \in \mathbb{N}$, $k \in [n]$, for any $D > 0$, for any $p \in \mathbb{N}$ such that

$$m + n \ge 2^8 K^2 p^6 (2a+1)^4,$$

we have, with probability at least $1 - (2^5/D)^{2p}$,

$$\left\|e_{n,k}^T (E^T E)^a E^T U\right\| \le D r^{1/2} p^{3/2} \sqrt{2a+1} \left(16 p^{3/2} (2a+1)^{3/2} K \frac{\|U\|_{2,\infty}}{\sqrt{r}} + 1\right) [2(m+n)]^a.$$

Lemma 5.4. *Let E be a $m \times n$ random matrix following the model in Lemma 5.3. For any matrix $V \in \mathbb{R}^{m \times l}$ with unit columns $v_1, v_2, \ldots, v_l$, any $a \in \mathbb{N}$, $k \in [n]$, any $D > 0$, and any $p \in \mathbb{N}$ such that*

$$m + n \ge 2^8 K^2 p^6 (2a)^4,$$

we have, with probability at least $1 - (2^4/D)^{2p}$,

$$\left\|e_{n,k}^T (E^T E)^a V\right\| \le D p \|V\|_{2,\infty} [2(m+n)]^a.$$

Let us prove the main objective of this section, Lemma 4.2, before delving into the proof of the technical lemmas.

Proof of Lemma 4.2. Consider Eq. (73) and assume $K \le \log^{-2-\varepsilon}(m+n)\sqrt{m+n}$. Fix $k \in [n]$. It suffices to prove the following two bounds uniformly over all $a \in [\lfloor t \log(m+n) \rfloor]$:

$$\left\|e_{n,k}^T (E^T E)^a E^T U\right\| \le C\sqrt{r} D_{U,V,\log\log(m+n)} (1.9\kappa\sqrt{m+n})^{2a+1} \tag{87}$$

$$\left\|e_{n,k}^T (E^T E)^a V\right\| \le C\sqrt{r} D_{U,V,\log\log(m+n)} (1.9\kappa\sqrt{m+n})^{2a}. \tag{88}$$

Fix $a \in [\lfloor t \log(m+n) \rfloor]$. Let $p = \log\log(m+n)$. We can assume p is an integer for simplicity without any loss. This choice ensures

$$K^2 p^6 (2a)^4 < K^2 p^6 (2a+1)^4 \le \frac{(m+n) t^4 \log^4(m+n) \log^6 \log(m+n)}{\log^{4+2\varepsilon}(m+n)} \ll m+n,$$

so we can apply both Lemmas 5.3 and 5.4.

Let us prove Eq. (87) for a. Applying Lemma 5.3 for the random matrix E/κ and $D = 2^{13}$ gives, with probability $1 - \log^{-4.04}(m+n)$,

$$\begin{aligned}\frac{\|e_{n,k}^T (E^T E)^a E^T U\|}{(1.9\kappa\sqrt{m+n})^{2a+1}} &\le \frac{D r^{1/2} p^{3/2} \sqrt{2a+1}}{1.9\sqrt{m+n}} \left(16 p^{3/2} (2a+1)^{3/2} K \frac{\|U\|_{2,\infty}}{\sqrt{r}} + 1\right) \left(\frac{2}{3.61}\right)^a \\ &\le \frac{D r^{1/2} p^{3/2}}{\sqrt{m+n}} \left(16 p^{3/2} K \frac{\|U\|_{2,\infty}}{\sqrt{r}} + 1\right) \le 2^{17}\sqrt{r} \left(\frac{K p^3 \|U\|_{2,\infty}}{\sqrt{r(m+n)}} + \frac{p^{3/2}}{\sqrt{m+n}}\right),\end{aligned}$$

where the second inequality is due to $\alpha \leq (\sqrt{2}/1.9)^{\alpha}$. A union bound over all $a \in [\lfloor t\log(m+n)\rfloor]$ makes the bound uniform, with probability at least $1-\log^{-3}(m+n)$. The term inside parentheses in the last expression is less than $D_{U,V,\log\log(m+n)}$, so Eq. (87) follows.

Let us prove Eq. (88). Applying Lemma 5.3 for the random matrix E/κ and $D = 2^{10}$ gives, with probability $1-\log^{-8}(m+n)$,

$$\frac{\|e_{n,k}^T(E^TE)^aV\|}{(1.9\kappa\sqrt{m+n})^{2a+1}} \leq Dp\|V\|_{2,\infty}\left(\frac{2}{3.61}\right)^a \leq 2^{10}p\|V\|_{2,\infty} \leq 2^{10}\sqrt{r}D_{U,V,p},$$

proving Eq. (88) after a union bound, similar to the previous case. Combining the two cases, Eq. (73) is proven.

Let us now prove Eq. (74). Since the 2-to-∞ norm is the the largest norm among the rows, it suffices to prove Eq. (73) holds uniformly over all $k \in [n]$ for $p = \log(m+n)$. Substituting this new choice of p into the previous argument, for a fixed k, we have Eq. (73), but with probability at least $1-(m+n)^{-4.04}$. Applying another union bound over $k \leq [n]$ gives Eq. (74) with probability at least $1-(m+n)^{-3}$. The proof is complete. □

Now let us handle the technical lemmas 5.3 and 5.4. The odd case (Lemma 5.3) is more difficult, so we will handle it first to demonstrate our technique. The argument for the even case (Lemma 5.4) is just a simpler version of the same technique.

5.2.1 Case 1: odd powers

Proof. Without loss of generality, let $k = 1$. Let us fix $p \in \mathbb{N}$ and bound the $(2p)^{th}$ moment of the expression of concern. We have

$$\mathbf{E}\left[\left\|e_{n,1}^T(E^TE)^aE^TU\right\|^{2p}\right] = \mathbf{E}\left[\left(\sum_{l=1}^{r}(e_{n,1}^T(E^TE)^aE^Tu_l)^2\right)^p\right] = \sum_{l_1,\ldots,l_p\in[r]}\mathbf{E}\left[\prod_{h=1}^{p}(e_{n,1}^T(E^TE)^aE^Tu_{l_h})^2\right]. \tag{89}$$

Temporarily let $\mathcal{W}$ be the set of walks $W = (j_0i_0j_1i_1\ldots i_a)$ of length $2a+1$ on the complete bipartite graph $K_{m,n}$ such that $j_0 = 1$. Here the two parts of K are $I = \{1', 2', \ldots, m'\}$ and $J = \{1, 2, \ldots, n\}$, where the prime symbol serves to distinguish two vertices on different parts with the same number. Let $E_W = E_{i_0j_0}E_{i_0j_1}\ldots E_{i_{a-1}j_a}E_{i_aj_a}$. We can rewrite the final expression in the above as

$$\sum_{l_1,l_2,\ldots,l_p\in[r]}\ \sum_{W_{11},W_{12},W_{21},\ldots,W_{p2}\in\mathcal{W}}\mathbf{E}\left[\prod_{h=1}^{p}E_{W_{h1}}E_{W_{h2}}u_{l_hi_{(h1)a}}u_{l_hi_{(h2)a}}\right],$$

where we denote $W_{hd} = (j_{(hd)0}, i_{(hd)0}, \ldots, i_{(hd)a})$. We can swap the two summation in the above to get

$$\sum_{W_{11},W_{12},W_{21},\ldots,W_{p2}\in\mathcal{W}}\mathbf{E}\left[\prod_{h=1}^{p}E_{W_{h1}}E_{W_{h2}}\right]\sum_{l_1,l_2,\ldots,l_p\in[r]}\prod_{h=1}^{p}u_{l_hi_{(h1)a}}u_{l_hi_{(h2)a}}.$$

The second sum can be recollected in the form of a product, so we can rewrite the above as

$$\sum_{W_{11},W_{12},W_{21},\ldots,W_{p2}\in\mathcal{W}}\mathbf{E}\left[\prod_{h=1}^{p}E_{W_{h1}}E_{W_{h2}}\right]\prod_{h=1}^{p}U_{\cdot,i_{(h1)a}}^TU_{\cdot,i_{(h2)a}}$$

Define the following notation:

1. $\mathcal{P}$ is the set of all *star*, i.e. tuples of walks $P = (P_1, \dots, P_{2p})$ on the complete bipartite graph $K_{m,n}$, such that each walk $P_r \in \mathcal{W}$ and each edge appears at least twice.

   Rename each tuple $(W_{h1}, W_{h2})_{h=1}^{p}$ as a star P with $W_{hd} = P_{2h-2+d}$.

   For each P, let $V(P)$ and $E(P)$ respectively be the set of vertices and edges involved in P.

   Define the partition $V(P) = V_I(P) \cup V_J(P)$, where $V_I(P) := V(P) \cap I$ and $V_J(P) := V(P) \cap J$.

2. $E_P := E_{P_1} E_{P_2} \dots E_{P_{2p}}$.

3. $P^{\text{end}} := (i_{1a}, i_{2a}, \dots, i_{(2p)a})$, which we call the *boundary* of P. Then $u_Q := \prod_{r=1}^{2p} u_{q_r}$ for any tuple $Q = (q_1, \dots, q_r)$.

4. $\mathcal{S}$ is the subset of "shapes" in $\mathcal{P}$. A shape is a tuple of walks $S = (S_1, \dots, S_{2p})$ such that all S_r start with 1 and for all $r \in [2p]$ and $s \in [0, a]$, if i_{rs} appears for the first time in $\{i_{r's'} : r' \le r, s' \le s\}$, then it is stricly larger than all indices before it, and similarly for j_{rs}. We say a star $P \in \mathcal{P}$ has shape $S \in \mathcal{S}$ if there is a bijection from $V(P)$ to $[|V(P)|]$ that transforms P into S. The notations $V(S)$, $V_I(S)$, $V_J(S)$, $E(S)$ are defined analogously. Observe that the shape of P is unique, and $\mathcal{S}$ forms a set of equivalent classes on $\mathcal{P}$.

5. Denote by $\mathcal{P}(S)$ the class associated with the shape S, namely the set of all stars P having shape S.

We can rewrite the previous sum as:

$$\sum_{P \in \mathcal{P}} \mathbf{E}\left[E_P\right] \prod_{h=1}^{p} U^T_{\cdot, i_{(2h-1)a}} U_{\cdot, i_{(2h)a}}$$

Using triangle inequality and the sub-multiplicity of the operator norm, we get the following upper bound for the above:

$$\sum_{P \in \mathcal{P}} |\mathbf{E}\left[E_P\right]| \prod_{h=1}^{p} \|U_{\cdot, i_{(2h-1)a}}\| \|U_{\cdot, i_{(2h)a}}\| = r^p \sum_{P \in \mathcal{P}} u_{P^{\text{end}}} |\mathbf{E}\left[E_P\right]| = r^p \sum_{S \in \mathcal{S}} \sum_{P \in \mathcal{P}(S)} u_{P^{\text{end}}} |\mathbf{E}\left[E_P\right]|, \quad (90)$$

where the vector u is given by $u_i = r^{-1/2} \|U_{\cdot, i}\|$ for $i \in [m]$. Observe that

$$\|u\| = 1 \quad \text{and} \quad \|u\|_\infty = r^{-1/2} \|U\|_{2,\infty}.$$

Fix $P \in \mathcal{P}$. Let us bound $\mathbf{E}\left[E_P\right]$. For each $(i,j) \in E(P)$, let $\mu_P(i,j)$ be the number of times (i,j) is traversed in P. We have

$$|\mathbf{E}\left[E_P\right]| = \prod_{(i,j) \in E(P)} \mathbf{E}\left[|E_{ij}|^{\mu_P(i,j)}\right] \le \prod_{(i,j) \in E(P)} K^{\mu_P(i,j) - 2} = K^{2p(2a+1) - 2|E(P)|}.$$

Since the entries u_i are related by the fact their squares sum to 1, it will be better to bound their symmetric sums rather than just a product $u_{P^{\text{end}}}$. Fix a shape S, we have

$$\sum_{P \in \mathcal{P}(S)} |u_{P^{\text{end}}}| = \sum_{f : V(S) \hookrightarrow [m]} \prod_{k=1}^{|V(S^{\text{end}})|} |u_{f(k)}|^{\mu_{S^{\text{end}}}(k)} \le m^{|V_I(S)| - |V(S^{\text{end}})|} n^{|V_J(S)| - 1} \prod_{k=1}^{|V(S^{\text{end}})|} \sum_{i=1}^{m} |u_i|^{\mu_{S^{\text{end}}}(k)}$$

$$= m^{|V_I(S)| - |V(S^{\text{end}})|} n^{|V_J(S)| - 1} \prod_{k=1}^{|V(S^{\text{end}})|} \|u\|_{\mu_{S^{\text{end}}}(k)}^{\mu_{S^{\text{end}}}(k)},$$

where we slightly abuse notation by letting $\mu_Q(k)$ be the number of time k appears in Q.

Consider $\|u\|_l^l$ for an arbitrary $l \in \mathbb{N}$. When $l = 1$, $\|u\|_l^l \leq \sqrt{m}$ by Cauchy-Schwarz. When $l \geq 2$, we have $\|u\|_l^l \leq \|u\|_\infty^{l-2}\|u\|_2^2 = \|u\|_\infty^{l-2}$. Thus

$$\sum_{P\in\mathcal{P}(S)} |u_{P^{\text{end}}}| \leq \prod_{k=1}^{|V(S)|} \|u\|_{\mu_{S^{\text{end}}}(k)}^{\mu_{S^{\text{end}}}(k)} \leq \prod_{k\in V_2(S)} \|u\|_\infty^{\mu_{S^{\text{end}}}(k)-2}(\sqrt{m})^{|V_1(S^{\text{end}})|} = \|u\|_\infty^{2p-\nu(S)} m^{|V_1(S^{\text{end}})|/2},$$

where, we define $V_1(Q)$ as the set of vertices appearing in Q exactly once and $V_2(Q)$ as the set of vertices appearing at least twice, and to shorten the notation, we let $\nu(S) := |V_1(S^{\text{end}})|+2|V_2(S^{\text{end}})|$. Combining the bounds, we get the upper bound below for (90):

$$\begin{aligned}&K^{2p(2a+1)} \sum_{S\in\mathcal{S}} K^{-2|E(S)|} m^{|V_I(S)|-|V(S^{\text{end}})|} n^{|V_J(S)|-1} \|u\|_\infty^{2p-\nu(S)} m^{|V_1(S^{\text{end}})|/2} \\ &= K^{2p(2a+1)+2} \sum_{S\in\mathcal{S}} K^{-2|V(S)|} m^{|V_I(S)|-\nu(S)/2} n^{|V_J(S)|-1} \|u\|_\infty^{2p-\nu(S)}.\end{aligned}$$

Suppose we fix $|V_1(S^{\text{end}})| = x$, $|V_2(S^{\text{end}})| = y$, $|V_I(S)| = z$, $|V_J(S)| = t$. Let $\mathcal{S}(x, y, z, t)$ be the subset of shapes having these quantities. To further shorten the notation, let $K_1 := K^{2p(2a+1)}\|u\|_\infty^{2p}$. Then we can rewrite the above as:

$$K_1 \sum_{x,y,z,t\in\mathcal{A}} K^{-2(z+t)} m^{z-x/2-y} n^{t-1} \|u\|_\infty^{-x-2y} |\mathcal{S}(x, y, z, t)|, \tag{91}$$

where $\mathcal{A}$ is defined, somewhat abstractly, as the set of all tuples (x, y, z, t) such that $\mathcal{S}(x, y, z, t) \neq \varnothing$. We first derive some basic conditions for such tuples. Trivially, one has the following initial bounds:

$$0 \leq x, y, \qquad 1 \leq x + y \leq z, \qquad x + 2y \leq 2p, \qquad 0 \leq z, t, \qquad z + t \leq p(2a + 1) + 1,$$

where the last bound is due to $z + t = |V(S)| \leq |E(S)| + 1 \leq p(2a + 1) + 1$, since each edge is repeated at least twice. However, it is not strong enough, since we want the highest power of m and n combined to be at most $2ap$, so we need to eliminate a quantity of p.

Claim 5.5. *When each edge is repeated at least twice, we have $z - x/2 - y + t - 1 \leq 2ap$.*

Proof of Claim 5.5. Let $S = (S_1, \ldots, S_{2p})$, where $S_r = j_{r0}i_{r0}j_{r1}i_{r1} \ldots j_{ra}i_{ra}$. We have $j_{r0} = 1$ for all r. It is tempting to think (falsely) that when each edge is repeated at least twice, each vertex appears at least twice too. If this were to be the case, then each vertex in the set

$$A(S) := \{i_{rs} : 1 \leq r \leq 2p, 0 \leq s \leq a - 1\} \cup \{j_{rs} : 1 \leq r \leq 2p, 1 \leq s \leq a\} \cup V_1(S^{\text{end}})$$

appears at least twice. The sum of their repetitions is $4ap + x$, so the size of this set is at most $2ap + x/2$. Since this set covers every vertex, with the possible exceptions of $1 \in I$ and $V_2(S^{\text{end}})$, its size is at least $z - y + t - 1$, proving the claim. In general, there will be vertices appearing only once in S. However, we can still use the simple idea above. Temporarily let $A_1(S)$ be the set of vertices appearing once in S and $f(S)$ be the sum of all edges' repetitions in S. Let $S^{(0)} := S$. Suppose for $k \geq 0$, $S^{(k)}$ is known and satisfies $|A(S^{(k)})| = |A(S)| - k$, $f(S^{(k)}) = 4pa + x - 2k$ and each edge appears at least twice in $S^{(k)}$. If $A_1(S^{(k)}) = \varnothing$, then by the previous argument, we have

$$2(z - y + t - 1 - k) \leq 4pa + x - 2k \implies z - x/2 - y + t - 1 \leq 2pa,$$

proving the claim. If there is some vertex in $A_1(S^{(k)})$, assume it is some i_{rs}, then we must have $s \leq a - 1$ and $j_{rs} = j_{r(s+1)}$, otherwise the edge $j_{rs}i_{rs}$ appears only once. Create $S^{(k+1)}$ from

$S^{(k)}$ by removing i_{rs} and identifying j_{rs} and $j_{r(s+1)}$, we have $|A(S^{(k+1)})| = |A(S)| - (k+1)$ and $|f(S^{(k)}) = 4pa + x - 2(k+1)$. Further, since i_{rs} is unique, $j_{rs}i_{rs} \equiv i_{rs}j_{r(s+1)}$ are the only 2 occurences of this edge in $S^{(k)}$, thus the edges remaining in $S^{(k+1)}$ also appears at least twice. Now we only have $|A_1(S^{(k+1)})| \leq |A_1(S^{(k)})|$, with possible equality, since j_{rs} can be come unique after the removal, but since there is only a finite number of edges to remove, eventually we have $A_1(S^{(k)}) = \varnothing$, completing the proof of the claim. □

Claim 5.5 shows that we can define the set $\mathcal{A}$ of *eligible sizes* as follows:

$$\mathcal{A} = \left\{(x,y,z,t) \in \mathbb{N}_{\geq 0}^4 : \quad 1 \leq t; \quad 1 \leq x+y \leq z; \quad x+2y \leq 2p; \quad z - x/2 - y + t - 1 \leq 2ap\right\}. \tag{92}$$

Now it remains to bound $|\mathcal{S}(x,y,z,t)|$.

Claim 5.6. *Given a tuple $(x,y,z,t) \in \mathcal{A}$, where $\mathcal{A}$ is defined in Eq. (92), we have*

$$|\mathcal{S}(x,y,z,t)| \leq \frac{2^{l+1}(2p(a+1))!(2pa)!(l+1)^{2p(2a+1)-2l}}{(2p(2a+1)-2l)!l!z!(t-1)!}(16p(a+1)-8l-2)^{4p(a+1)-2l-1}.$$

Proof. We use the following coding scheme for each shape $S \in \mathcal{S}(x,y,z,t)$: Given such an S, we can progressively build a codeword $W(S)$ and an associated tree $T(S)$ accoding to the following scheme:

1. Start with $V_J = \{1\}$ and $V_I = \varnothing$, $W = []$ and T being the tree with one vertex, 1.
2. For $r = 1, 2, \ldots, 2p$:
    (a) Relabel S_r as $1k_1k_2\ldots k_{2a}$.
    (b) For $s = 1, 2, \ldots, 2a$:
    - If $k_s \notin V(T)$ then add k_s to T and draw an edge connecting k_{s-1} and k_s, then mark that edge with a $(+)$ in T, and append $(+)$ to W. We call its instance in S_r a *plus edge.*
    - If $k_s \in V(T)$ and the edge $k_{s-1}k_s \in E(T)$ and is marked with $(+)$: unmark it in T, and append $(-)$ to W. We call its instance in S_r a *minus edge.*
    - If $k_s \in V(T)$ but either $k_{s-1}k_s \notin E(T)$ or is unmarked, we call its instance in S_r a *neutral edge*, and append the symbol k_s to W.

This scheme only creates a *preliminary codeword* W, which does not yet uniquely determine the original S. To be able to trace back S, we need the scheme in [28] to add more details to the preliminary codewords. For completeness, we will describe this scheme later, but let us first bound the number of preliminary codewords.

Claim 5.7. *Let $\mathcal{PC}(x,y,z,t)$ denote the set of preliminary codewords generable from shapes in $\mathcal{S}(x,y,z,t)$. Then for $l := z + t - 1$ we have*

$$|\mathcal{PC}(x,y,z,t)| \leq \frac{2^l(2p(a+1))!(2pa)!(l+1)^{2p(2a+1)-2l}}{(2p(2a+1)-2l)!l!z!(t-1)!}.$$

Note that the bound above does not depend on x and y. In fact, for fixed z and t, the right-hand side is actually an upper bound for the sum of $|\mathcal{S}(x,y,z,t)|$ over all pairs (x,y) such that (x,y,z,t) is eligible. We believe there is plenty of room to improve this bound in the future.

Proof. To begin, note that there are precisely z and $t-1$ plus edges whose right endpoint is respectively in I and J. Suppose we know u and v, the number of minus edges whose right endpoint is in I and J, respectively. Then

- The number of ways to place plus edges is at most $\binom{2p(a+1)}{z}\binom{2pa}{t-1}$.
- The number of ways to place minus edges, given the position of plus edges, is at most $\binom{2p(a+1)-z}{u}\binom{2pa-t+1}{v}$.
- The number of ways to choose the endpoint for each neutral edge is at most $z^{2p(a+1)-z-u}t^{2pa-t+1-v}$.

Combining the bounds above, we have

$$|\mathcal{S}(x,y,z,t)| \leq \binom{2p(a+1)}{z}\binom{2pa}{t-1}\sum_{u+v=z+t-1}\binom{2p(a+1)-z}{u}\binom{2pa-t+1}{v}z^{f(z,u)}t^{g(t,v)}, \quad (93)$$

where $f(z,u) = 2p(a+1)-z-u$ and $g(u,v) = 2pa-t+1-v$. Let us simplify this bound. The sum on the right-hand side has the form

$$\sum_{i+j=k}\binom{N}{i}\binom{M}{j}z^i t^j,$$

where $k = 2(p(2a+1)-(z+t-1))$, $N = 2p(a+1)-z$, $M = 2pa-t+1$. We have

$$\sum_{i+j=k}\binom{N}{i}\binom{M}{j}z^i t^j = \sum_{i+j=k}\frac{N!M!}{k!(N-i)!(M-j)!}\binom{k}{i}z^i t^j \leq \sum_{i+j=k}\frac{N!M!}{k!}\frac{(z+t)^k}{(N-i)(M-j)!}$$
$$\leq \frac{N!M!(z+t)^k}{k!(M+N-k)!}\sum_{i+j=k}\binom{M+N-k}{N-i} \leq \frac{2^{M+N-k}N!M!(z+t)^k}{k!(M+N-k)!}.$$

Replacing M, N and k with their definitions, we get

$$\sum_{u+v=z+t-1}\binom{2p(a+1)-z}{u}\binom{2pa-t+1}{v}z^{f(z,u)}t^{g(t,v)}$$
$$\leq \frac{2^{z+t-1}(2p(a+1)-z)!(2pa-t+1)!(z+t)^{2p(2a+1)-2(z+t-1)}}{(2p(2a+1)-2(z+t-1))!(z+t-1)!},$$

replacing $z+t-1$ with l, we prove the claim. □

Back to the proof of Claim 5.6, to uniquely determine the shape S, the general idea is the following. We first generated the preliminary codeword W from S, then attempt to decode it. If we encounter a plus or neutral edge, we immediately know the next vertex. If we see a minus edge that follows from a plus edge (u,v), we know that the next vertex is again u. Similarly, if there are chunks of the form $(++\ldots+--\ldots-)$ with the same number of each sign, the vertices are uniquely determined from the first vertex. Therefore, we can create a condensed codeword W^* repeatedly removing consecutive pairs of $(+-)$ until none remains. For example, the sections $(-+-+-)$ and $(-++--)$ both become $(-)$. Observe that the condensed codeword is always unique regardless of the order of removal, and has the form

$$W^* = [(+\ldots+) \text{ or } (-\ldots-)] \text{ (neutral) } [(+\ldots+) \text{ or } (-\ldots-)] \ldots \text{ (neutral) } [(+\ldots+) \text{ or } (-\ldots-)],$$

where we allow blocks of pure pluses and minuses to be empty. The minus blocks that remain in W^* are the only ones where we cannot decipher.

Recall that during decoding, we also reconstruct the tree $T(S)$, and the partial result remains a tree at any step. If we encounter a block of minuses in W^* beginning with the vertex i, knowing the right endpoint j of the last minus edge is enough to determine the rest of the vertices, which is just the unique path between i and j in the current tree. We call the last minus edge of such a block an *important edge*. There are two cases for an important edge.

1. If i and all vertices between i and j (excluding j) are only adjacent to at most two plus edges in the current tree (exactly for the interior vertices), we call this important edge *simple* and just mark the it with a direction (left or right, in addition to the existing minus). For example, $(- - \ldots -)$ becomes $(- - \ldots (-dir))$ where dir is the direction.
2. If the edge is non-simple, we just mark it with the vertex j, so $(- - \ldots -)$ becomes $(- - \ldots (-j))$.

It has been shown in [28] that the fully codeword $\overline{W}$ resulting from W by marking important edges uniquely determines S, and that when the shape of S *is that of a single walk*, the cost of these markings is at most a multiplicative factor of $2(4N+8)^N$, where N is the number of neutral edges in the preliminary W. To adapt this bound to our case, we treat the star shape S as a single walk, with a neutral edge marked by 1 after every $2a+1$ edges. There are $2p-1$ additional neutral edges from this perspective, making $N = 4p(a+1) - 2l - 1$ in total. Combining this with the bound on the number of preliminary codewords (Claim 5.7) yields

$$|\mathcal{S}(x,y,z,t)| \le \frac{2^{l+1}(2p(a+1))!(2pa)!(l+1)^{2p(2a+1)-2l}}{(2p(2a+1)-2l)!l!z!(t-1)!}(16p(a+1)-8l-2)^{4p(a+1)-2l-1},$$

where $l = z+t-1$. Claim 5.6 is proven. □

Back to the proof of Lemma 5.3. Temporarily let

$$G_l := 2p(2a+1) - 2l \quad \text{and} \quad F_l := \frac{2^{l+1}(l+1)^{G_l}}{G_l!l!}(4G_l + 8p - 2)^{G_l+2p-1}.$$

Note that $(2p(a+1))!(2pa)!F_l$ is precisely the upper bound on $|\mathcal{S}(x,y,z,t)|$ in Claim 5.6. Also let

$$K_2 = K_1(2p(a+1))!(2pa)! = K^{2p(2a+1)}(2p(a+1))!(2pa)!\|u\|_\infty^{2p}.$$

Replacing the appropriate terms in the bound in Claim 5.6 with these short forms, we get another series of upper bounds for the last double sum in Eq. (90):

$$\begin{aligned}
&K_2 \sum_{x,y} \|u\|_\infty^{-x-2y} \sum_{l=x+y}^{\lfloor 2pa+x/2+y \rfloor} K^{-2(l+1)} F_l \sum_{z+t=l+1} \frac{m^{z-x/2-y} n^{t-1}}{z!(t-1)!} \\
&\le K_2 \sum_{x,y} \|u\|_\infty^{-x-2y} \sum_{l=x+y}^{\lfloor 2pa+x/2+y \rfloor} \frac{K^{-2(l+1)} F_l}{(l - \lfloor \frac{x}{2} \rfloor - y)!} \sum_{z+t=l+1} \binom{l - \lfloor \frac{x}{2} \rfloor - y}{z - \lfloor \frac{x}{2} \rfloor - y} m^{z - \lfloor \frac{x}{2} \rfloor - y} n^{t-1} \\
&\le K_2 \sum_{x,y} \|u\|_\infty^{-x-2y} \sum_{l=x+y}^{\lfloor 2pa+x/2+y \rfloor} \frac{K^{-2(l+1)} F_l}{(l - \lfloor \frac{x}{2} \rfloor - y)!} (m+n)^{l - \lfloor \frac{x}{2} \rfloor - y}.
\end{aligned}$$

Temporarily let C_l be the term corresponding to l in the sum above. For $l \geq x+y+1$, we have

$$\frac{C_l}{C_{l-1}} = \frac{2(m+n)(G_l+1)(G_l+2)}{K^2 l^3 (4G_l+8p-2)^2 (l-\lfloor \frac{x}{2} \rfloor - y)} \left(1+\frac{1}{l}\right)^{G_l} \left(1-\frac{4}{2G_l+4p+3}\right)^{G_l+2p+1}.$$

The last power is approximately $e^{-2} \approx 0.135$, and for $p \geq 7$ a routine numerical check shows that it is at least $1/8$. The second to last power is at least 1. The fraction be bounded as below.

$$\frac{2(m+n)(G_l+1)(G_l+2)}{K^2 l^3 (4G_l+8p-2)^2 (l-\lfloor \frac{x}{2} \rfloor - y)} \geq \frac{2(m+n)\cdot 1 \cdot 2}{K^2 l^4 (8p-2)^2} \geq \frac{m+n}{16K^2 l^4 p^2} \geq \frac{m+n}{16K^2 p^6 (2a+1)^4}.$$

Therefore, under the assumption that $m+n \geq 256K^2p^6(2a+1)^4$, we have $C_l \geq 2C_{l-1}$ for all $l \geq 1$, so $\sum_l C_l \leq 2C_{l^*}$, where $l^* = \lfloor 2pa + x/2 + y \rfloor$, the maximum in the range. We have

$$\begin{aligned} 2C_{l^*} \leq 2(m+n)^{2pa} & \frac{(2K^{-2})^{2pa+\lfloor \frac{x}{2} \rfloor + y + 1}(2pa+\lfloor \frac{x}{2} \rfloor + y + 1)^{2(p-\lfloor \frac{x}{2} \rfloor - y)}}{(2(p-\lfloor \frac{x}{2} \rfloor - y))! \cdot (2pa+\lfloor \frac{x}{2} \rfloor + y)! \cdot (2pa)!} \\ & \cdot \left(16p - 8\left\lfloor \frac{x}{2} \right\rfloor - 8y - 2\right)^{4p-2\lfloor \frac{x}{2} \rfloor - 2y - 1}. \end{aligned}$$

Temporarily let $d = p - (\lfloor \frac{x}{2} \rfloor + y)$ and $N = p(2a+1)$, we have

$$2C_{l^*} \leq 2(m+n)^{2pa} \frac{(2K^{-2})^{N-d+1}(N-d+1)^{2d}(8p+8d-2)^{2p+2d-1}}{(2pa)! \cdot (2d)! \cdot (N-d)!}.$$

For each d, there are at most $2(p-d)$ pairs (x,y) such that $d = p - (\lfloor \frac{x}{2} \rfloor + y)$, so overall we have the following series of upper bounds for the last double sum in Eq. (90):

$$\begin{aligned} & K_2(m+n)^{2pa} \sum_{d=0}^{p-1} 4(p-d)\|u\|_\infty^{-2(p-d)} \cdot \frac{(2K^{-2})^{N-d+1}(N-d+1)^{2d}(8p+8d-2)^{2p+2d-1}}{(2pa)! \cdot (2d)! \cdot (N-d)!} \\ & \leq K_3(m+n)^{2pa} \sum_{d=0}^{p-1} \|u\|_\infty^{2d} \cdot \frac{2^{-d}K^{2d}(N-d+1)^{2d}(8p+8d-2)^{2p+2d-1}}{(2d)! \cdot (N-d)!}, \end{aligned} \tag{94}$$

where

$$K_3 = 4p\frac{K_2\|u\|_\infty^{-2p}(2K^{-2})^{N+1}}{(2pa)!} = 2^{p(2a+1)+3}pK^{-2}(2p(a+1))!.$$

Let us bound the sum at the end of Eq. (94). Temporarily let A_d be the term corresponding to d and $x := 2^{-1/2}K\|u\|_\infty$. We have

$$A_d = \frac{x^{2d}(N-d+1)^{2d}}{(2d)!(N-d)!}(8p+8d-2)^{2p+2d-1} \leq \frac{x^{2d}N^{3d}}{(2d)!N!}\frac{(16p)^{2p+2d}}{8p}.$$

Therefore

$$\begin{aligned} \sum_{d=0}^{p-1} A_d \leq \frac{(16p)^{2p}}{8pN!}\sum_{d=0}^{p-1}\frac{(16pN^{3/2}x)^{2d}}{(2d)!} &\leq \frac{(16p)^{2p}}{8pN!}\sum_{d=0}^{p-1}\binom{2p}{2d}(16pN^{3/2}x)^{2d}\frac{e^{2d}}{(2p)^{2d}} \\ &= \frac{(16p)^{2p}}{8pN!}(8eN^{3/2}x+1)^{2p} \leq \frac{(16p)^{2p}}{8pN!}(16N^{3/2}K\|u\|_\infty+1)^{2p}. \end{aligned}$$

Plugging this into Eq. (94), we get another upper bound for (90):

$$K_4(16N^{3/2}K\|u\|_\infty+1)^{2p}(m+n)^{2ap},$$

where

$$K_4 := K_3\frac{(16p)^{2p}}{8pN!} = 2^{p(2a+1)+3}pK^{-2}(2p(a+1))!\frac{(16p)^{2p}}{8p(2ap+p)!} \le \frac{2^{2ap}2^{10p}p^{3p}(a+1)^p}{8K^2}.$$

To sum up, we have

$$\begin{aligned}\mathbf{E}\left[\left\|e_{n,1}^T(E^TE)^aE^TU\right\|^{2p}\right] &\le r^p\sum_{S\in\mathcal{S}}\sum_{P\in\mathcal{P}(S)}u_{P^{\text{end}}}\left|\mathbf{E}\left[E_P\right]\right| \\ &\le \frac{r^p2^{2ap}2^{10p}p^{3p}(a+1)^p}{8K^2}(16N^{3/2}K\|u\|_\infty+1)^{2p}(m+n)^{2ap} \\ &\le \left(2^5r^{1/2}p^{3/2}\sqrt{2a+1}(2^4p^{3/2}(2a+1)^{3/2}K\|u\|_\infty+1)\cdot[2(m+n)]^a\right)^{2p}.\end{aligned}$$

Let $D>0$ be arbitrary. By Markov's inequality, for any p such that $m+n\ge 2^8K^2p^6(2a+1)^4$, the moment bound above applies, so we have

$$\left\|e_{n,1}^T(E^TE)^aE^TU\right\| \le Dr^{1/2}p^{3/2}\sqrt{2a+1}(16p^{3/2}(2a+1)^{3/2}K\|u\|_\infty+1)[2(m+n)]^a$$

with probability at least $1-(2^5/D)^{2p}$. Replacing $\|u\|_\infty$ with $\frac{1}{\sqrt{r}}\|U\|_{2,\infty}$, we complete the proof. □

5.2.2 Case 2: even powers

Proof. Without loss of generality, assume $k=1$. We can reuse the first part and the notations from the proof of Lemma 5.3 to get the bound

$$\mathbf{E}\left[\left\|e_{n,1}^T(E^TE)^aV\right\|^{2p}\right] \le r^p\sum_{S\in\mathcal{S}}\sum_{P\in\mathcal{P}(S)}v_{P^{\text{end}}}\left|\mathbf{E}\left[E_P\right]\right|,$$

where $v_i = r^{-1/2}\|V_{\cdot,i}\|$. Again,

$$\|v\|=1 \quad\text{and}\quad \|v\|_\infty = r^{-1/2}\|V\|_{2,\infty},$$

and $\mathcal{S}$ is the set of shapes such that every edge appears at least twice, $\mathcal{P}(S)$ is the set of stars having shape S, and

$$E_P=\prod_{ij\in E(P)}E_{ij}^{m_P(ij)}, \quad\text{and}\quad v_Q=\prod_{j\in V(Q)}v_j^{m_Q(j)}.$$

Note that a shape for a star now consists of walks of length $2a$:

$$S=(S_1,S_2,\dots,S_{2p}) \quad\text{where } S_r=j_{r0}i_{r0}j_{r1}i_{r1}\dots j_{ra}.$$

We have, for any shape S and $P\in\mathcal{P}(S)$,

$$\mathbf{E}\left[E_P\right]\le K^{4pa-2|E(S)|}\le K^{2pa-2|V(S)|+2}, \quad |v_{P^{\text{end}}}|\le\|v\|_\infty^{2p}, \quad\text{and } |\mathcal{P}(S)|\le m^{|V_I(S)|}n^{|V_J(S)|-1},$$

where the power of n in the last inequality is due to 1 having been fixed in $V_J(S)$. Therefore

$$\sum_{S\in\mathcal{S}}\sum_{P\in\mathcal{P}(S)}v_{P^{\text{end}}}\left|\mathbf{E}\left[E_P\right]\right|\le K_1\sum_{S\in\mathcal{S}}K^{-2|V(S)|}m^{|V_I(S)|}n^{|V_J(S)|-1}, \quad\text{where } K_1:=K^{4pa+2}\|v\|_\infty^{2p}.$$

Let $\mathcal{S}(z,t)$ be the set of shapes S such that $|V_I(S)| = z$ and $|V_J(S)| = t$. Let $\mathcal{A}$ be the set of eligible indices:

$$\mathcal{A} := \Big\{(z,t) \in \mathbb{N}^2 : 0 \le z,\ 1 \le t, \ \text{ and } z + t \le 2pa + 1\Big\}.$$

Using the previous argument in the proof of Lemma 5.3 for counting shapes, we have for $(z,t) \in \mathcal{A}$:

$$|\mathcal{S}(z,t)| \le \frac{[(2pa)!]^2 F_l}{z! \cdot (t-1)!} m^z n^{t-1}, \quad \text{where } l := z + t - 1 \in [2pa],$$

where

$$G_l := 4ap - 2l \ \text{ and } \ F_l := \frac{2^{l+1}(l+1)^{G_l}}{G_l! l!}(4G_l + 8p - 2)^{G_l + 2p - 1}.$$

We have

$$\sum_{S \in \mathcal{S}} \sum_{P \in \mathcal{P}(S)} v_{P^{\text{end}}} |\mathbf{E}[E_P]| \le K_1 \sum_{l=0}^{2ap} K^{-2(l+1)}[(2ap)!]^2 F_l \sum_{z+t=l+1} \frac{m^z n^{t-1}}{z! \cdot (t-1)!}$$
$$= K_2 \sum_{l=0}^{2ap} \frac{K^{-2l}F_l}{l!} \sum_{z+t=l+1} \binom{l}{z} m^z n^{t-1} = K_2 \sum_{l=0}^{2ap} \frac{K^{-2l}F_l}{l!}(m+n)^l,$$

where $K_2 := K_1[(2pa)!]^2 K^{-2} = K^{4ap}[(2pa)!]^2 \|v\|_\infty^{2p}$. Let C_l be the term corresponding to l in the last sum above. An anlogous calculation from the proof of Lemma 5.3 shows that under the assumption that $m + n \ge 256K^2p^6(2a)^4$, $C_l \ge 2C_{l-1}$ for each l, so $\sum_{l=0}^{2pa} C_l \le 2C_{2pa}$, where

$$C_{2pa} = \frac{K^{-4ap}2^{2ap+1}(8p-2)^{2p-1}}{[(2ap)!]^2}(m+n)^{2ap}.$$

Therefore

$$\mathbf{E}\left[\left\|e_{n,1}^T (E^T E)^a V\right\|^{2p}\right] \le r^p \sum_{S \in \mathcal{S}} \sum_{P \in \mathcal{P}(S)} v_{P^{\text{end}}} |\mathbf{E}[E_P]|$$
$$\le 2r^p K_2 \frac{K^{-4ap}2^{2ap+1}(8p-2)^{2p-1}}{[(2ap)!]^2}(m+n)^{2ap} = 4\left(2^3 p r^{1/2}\|v\|_\infty [2(m+n)]^a\right)^{2p}.$$

Pick $D > 0$, by Markov's inequality, we have

$$\mathbf{P}\left(\left\|e_{n,1}^T (E^T E)^a V\right\| \ge D p r^{1/2} \|v\|_\infty [2(m+n)]^a\right) \le \left(\frac{16p}{D}\right)^{2p}.$$

Replacing $\|v\|_\infty$ with $r^{-1/2}\|V\|_{2,\infty}$, we complete the proof. □

References

[1] Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4), Jul 2016.

[2] Abhinav Bhardwaj and Van Vu. Matrix perturbation: Davis-kahan in the infinity norm, 2023.

[3] Jian-Feng Cai, Emmanuel Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20:1956–1982, 03 2010.

[4] Emmanuel Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98:925 – 936, 07 2010.

[5] Emmanuel Candes and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56:2053 – 2080, 06 2010.

[6] Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 9:717–772, 11 2008.

[7] Emmanuel Candès and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23, 11 2006.

[8] Herman Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *The Annals of Mathematical Statistics*, 23(4):493 – 507, 1952.

[9] Alexander Chistov and Dima Grigoriev. *Complexity of quantifier elimination in the theory of algebraically closed fields*, volume 176, pages 17–31. 04 2006.

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

[11] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

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

[13] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, page 11–20, New York, NY, USA, 2014. Association for Computing Machinery.

[14] Moritz Hardt. Understanding alternating minimization for matrix completion. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 651–660, 12 2014.

[15] Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 638–678, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[16] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[17] Prateek Jain, Praneeth Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Symposium on the Theory of Computing*, 2012.

[18] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research - JMLR*, 11, 06 2009.

[19] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56:2980 – 2998, 07 2010.

[20] Xiao Peng Li, Lei Huang, H.C. So, and Bo Zhao. A survey on matrix completion: Perspective of signal processing, 01 2019.

[21] Sean O'Rourke, Van Vu, and Ke Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, 2013.

[22] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12, 10 2009.

[23] Jos F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(1-4):625–653, 1999.

[24] Kim-Chuan Toh, Michael Todd, and R Z. Sdpt3—a matlab software package for semidefinite programming, version 2.1. *Optimization Methods & Software - OPTIM METHOD SOFTW*, 11, 10 1999.

[25] Linh Tran and Van Vu. The "power of few" phenomenon: The sparse case. *Random Structures and Algorithms*, page to appear, 2023.

[26] Phuc Tran and Van Vu. Matrices with random perturbation: Davis-kahan theorem, 2023.

[27] Vinita Vasudevan and M. Ramakrishna. A hierarchical singular value decomposition algorithm for low rank matrices. *ArXiv*, abs/1710.02812, 2017.

[28] Van H. Vu. Spectral norm of random matrices. *Combinatorica*, 27:721–736, 2005.

[29] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12(1):99–111, mar 1972.