Fast exact recovery of noisy matrix from few entries: the infinity norm approach

BaoLinh Tran[1], Van Vu[1]

[1]Department of Mathematics, Yale University

Abstract

The matrix recovery (completion) problem, a central problem in data science and theoretical computer science, is to recover a matrix A from a relatively small set of entries.

While such a task is impossible in general, it has been shown that one can recover A exactly in polynomial time, with high probability, from a random subset of entries, under three 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. These assumptions are necessary and we refer to them as the basic assumptions.

There are many different algorithms for the task, including convex optimization by Candes et al. [10, 9, 30], alternating projection by Hart et al. [20, 21], and low rank approximation with gradient descent by Keshavan et al. [25, 24].

In applications, it is more realistic to assume that data is noisy. In this case, it has been shown that these approaches provide a approximate recovery with small root mean square error, again with high probability, but it is hard to an approximate recovery to an exact one.

Recently, results by Abbe et al. [1] and Bhardwaj et al. [6] concerning approximation in the infinity norm showed that we can achieve exact recovery even in the noisy case, given that the ground matrix has bounded precision. Beyond the three basic assumptions, they required either the condition number of A is small [1] or the gap the between consecutive singular values is large [6].

In this paper, we remove these extra spectral assumptions. As a result, we obtain a simple algorithm for exact recovery in the noisy case, under only three basic assumptions. This is the first such algorithm. T The algorithm computes a low rank approximation of the input (properly scalled) at a certain threshold, and then round off the entries to given the precision level. Thus, its main step is a truncated SVD, which runs very fast in both theory and practice.

The mathematics behind our analysis is totally different from all previous approaches. Using the contour integral method from operator theory combined with combinatorial ideas, we show that under some mild conditions, the best rank k approximations of two matrices A and $A+E$ (where E represents noise) are close in the infinity norm. This method may be of independent interest.

1 Introduction

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 task is to recover A, given A_Ω. This is the *matrix recovery* (or *matrix completion*) problem, a central problem in data science which has been received lots of attention in recent years. In this paper, we focus on *exact recovery*, where we want to recover all entries of A exactly.

It is standard to assume that the set Ω is *random*, and researchers propose two models: (a) Ω is sampled uniformly among subsets with the same size, or (b) that Ω has independently chosen entries, each with the same probability p, called the *sampling density*, which can be known or hidden. It is simple to replace the former model by the latter, using a simple conditioning trick. One samples the entries independently, and condition on the event that the sample size equals a given number.

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 each $s \in [r]$, define $A_s = \sum_{i=1}^s \sigma_i u_i v_i^T$ as the best rank-s approximation of A. Define B_s analogously for any matrix 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 $\|M\|_{2,\infty} := \sup\{\|Mu\|_\infty : \|u\|_2 = 1\}$, which is the largest row norm of M. Define $\mu(V)$ similarly. When U and V are 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.

  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 the absolute value of the largest entry. We stick to the definition above, which is consistent with the parameter μ_0 in many popular papers [10, 9, 30, 25, 24, 8].

- The following parameter has also been used:

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

  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. It is simple to see that

$$\mu_1 \le \mu_0 \sqrt{r}. \tag{4}$$

- We use C to denote a positive universal constant, whose value changes from place to place. The asymptotics notation are used under the assumption that $N \to \infty$.

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

- *Low-rank:* One assumes that $r := \text{rank}\, A$ is much smaller than $\min\{m, n\}$. This assumption is crucial as it reduces the degree of freedom of A significantly, making the problem solvable, at least from an information theory stand point. Many papers assume r is bounded ($r = O(1)$), while $m, n \to \infty$.

- *Incoherence:* This assumption ensures that the rows and columns of A are sufficiently "spread out", so the information does not concentrate in a small set of entries, which could be easily overlooked by random sampling. In technical terms, one requires μ_0, and (sometimes) also μ_1 to be small.

- *Sufficient sampling size/density:* Due to a coupon collector effect, 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 [9] gives the lower bound $|\Omega| CrN \log N$ for a sufficient large constant C This is equivalent to $p \geq Cr(m^{-1} + n^{-1}) \log N$ for the independent sampling model.

  For more discussion about the necesity of these assumptions, we refer to [10, 9, 15]. We will refer to these assumptions as the *basic assumptions*. These have been assumed in all results we discuss in this paper.

1.3 Exact recovery in practice: finite precision

Let us make an important comment on the notion of *exact recovery*. Clearly, if an entry of A is irrational, then it is impossible for our computers to write it down exactly, let alone computing it. Thus, exact recovery only makes sense when the entries of A have finite precision, which is the case in all real-life applications. To this end, we says that a matrix A have a finite precision ε_0, if its entries are integer multiples of a parameter $\varepsilon_0 > 0$. For instance, if all entries have two decimal places, then $\varepsilon = .01$.

In many pratical problems, such as compelting/rating a recommendation system, the parameter ε_0 is actually quite large. For instance, in the most influential problem in the field, the Netflix Challenge [4], the entries of A are ratings of movies, which are half integers from 1 to 5, so $\varepsilon_0 = 1/2$. The algorithm we propose and analyze in this paper will exploit this fact to our advantage. (One should not confuse this notion with the machine precision or machine epsilon, which is very small.)

1.4 A brief summary of existing methods

There is a huge literature on matrix completion. In this section, we try to summarize some of the main methods.

- *Nuclear norm minimization:* This method is based on convexifying the intuitive but NP-hard approach of minimizing the rank given the observations. This method is guaranteed to achieve exact recovery under perhaps the most general assumptions. However, the time complexity includes high powers of N and the calculation may be sensitive to noise [9].

- *Alternating projections:* This is based on another intuitive but NP-hard approach of fixing the rank, then minimizing the RMSE with the observations. The basic version of the algorithm switches between optimizing the column and row spaces, given the other, in alternating steps. Existing variants of alternating projections run well in practice.

- *Low-rank approximation:* The general idea here is to view the sample matrix A_Ω as a rescaled and unbiased random perturbation of A. This way, it is natural to first approximate A by taking a low rank approximation of $p^{-1}A_\Omega$ (where p is the density). Next, one can use an extra cleaning step to make the recovery exact. The first step (truncated SVD in this case) here has only one operation, and runs fast in practice. Our algorithm in Section 1.6 belongs to this category.

1.4.1 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 would like to solve the following optimization problem:

$$\text{minimize} \quad \operatorname{rank} X \quad \text{subject to} \quad X_\Omega = A_\Omega. \tag{5}$$

Unfortunately, this problem is NP-hard, and all existing algorithms take doubly exponential time in terms of the dimensions of A [14]. To overcome this problem, Candes and Recht [10], motivated by an idea from the *sparse signal recovery* problem in the field of *compressed sensing* [11, 17], proposed to replace the rank with the nuclear norm of X, leading to

$$\text{minimize} \quad \|X\|_* \quad \text{subject to} \quad X_\Omega = A_\Omega. \tag{6}$$

The paper [10] was shortly followed by Candes and Tao [9], with both improvements and trade-offs, and ultimately by Recht [30], who improved both previous results, proving that A is the unique solution to (6), given the sampling size bound

$$|\Omega| \geq C \max\{\mu_0, \mu_1^2\} r N \log^2 N, \tag{7}$$

for the coherence parameters μ_0 and μ_1 defined previously.

If one replaces μ_1 with $\mu_0\sqrt{r}$ (see (4), the RHS becomes $C\mu_0^2 r^2 N \log^2 N$. This attains the optimal power of N while missing slightly from the optimal powers of r and $\log N$.

The key advantage of replacing the rank in Problem (5) with the nuclear norm is that Problem (6) is a convex program, which can be further translated into a semidefinite program [10, 9], solvable in polynomial time by a number of algorithms. However, convex optimization program usually runs slowly in practice. The survey [27] mentioned the interior point-based methods SDPT3 [32] and SeDuMi [31], which can take up to $O(|\Omega|^2 N^2)$ floating point operations (FLOPs) assuming (7), even if one takes advantage of the sparsity of A_Ω. Indeed, as Ω is at least $NlogN$ (by coupon collector), the number of operations is $\Omega(N^4)$, which is too large even for moderate N. An *iterative singular value thresholding* method aiming to solve a regularized version of nuclear norm minimization, trading exactness for performance, has been proposed [7]. It is thus desirable to develop faster algorithms for the problem, and in what follows we discuss two other methods, which achieve this goal, under extra assumptions.

1.4.2 Modified alternating projections

The intuition behind this approach is to fix the rank, then attempt to match the observations at much as possible. If we *know* the rank r of A precisely, then it is natural to look at the following optimization problem

$$\text{minimize} \quad \|(A - XY^T)_\Omega\|_F^2 \quad \text{over } X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{r\times n}. \tag{8}$$

This, unfortunately, like (5), is NP-hard [15]. There have been many studies proposing variants of *alternating projections*, all of which involve the following basic idea: suppose one already obtains an approximator $X^{(l)}$ of X at iteration l, then $Y^{(l)}$ and $X^{(l+1)}$ are defined by

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

The survey [15] pointed out that these methods tend to outperform nuclear norm minimization in practice. Oh the other hand, there are few rigorous guarantees for recovery. The convergence and final output of the basic algorithm above also depends highly on the choice of $X^{(0)}$ [15].

Jain, Netrapalli and Sanghavi (2012) [23] developed one of the first 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 [23] recovers A within an Frobenius norm error ε in $O(|\Omega|r^2\log(1/\varepsilon))$ time with high probability. Since the Frobenius norm is larger than the infinity norm, this gives us an exact recovery if we set $\varepsilon = \varepsilon_0/3$, where ε_0 is the precision level of A; see subsection 1.3.

Compared to the previous approach, there are two new factors here. First one needs to know the rank of A precisely. Second, there is a strong dependence on the condition number (which means the result is only effective if the least singular value of A is comparble to the largest).

The condition number factor was reduced to quadratic by Hardt [20] and again by Hardt and Wooters [21] to logarithmic, at the cost of an increase in the powers of r, μ_0 and $\log N$.

Remark 1.1. In practice, the common situation is that we do not know the rank r exactly, but have some estimates (for instance, r is between known values $r_{\min}$ and $r_{\max}$). It has been suggested (see, for instance, [26]) that one tries all integers in this range as the potential value of r, which only increase the running time by a factor $r_{\max} - r_{\min}$, which is acceptable from the comlexity view point. The trouble here is that it is not clear that among these cases, which output should we choose.

1.4.3 Low rank approximation with Gradient descent

As discussed earlier, if one assumes the independent sampling model with probability p, then the rescaled sample matrix $p^{-1}A_\Omega$ can be viewed as a *random perturbation* of A. Since $\mathbf{E}[A_\Omega/p] = A$, this perturbation is unbiased, and the matrix $E := p^{-1}A_\Omega - A$ is a random matrix with mean zero.

Assuming that the rank r is known, Keshavan, Montanari and Oh [25] first use the best rank-r approximation of $p^{-1}A_\Omega$ to obtain an approximation of A, then add a cleaning step, using optimization via gradient descent, to achieve exact recovery. Here is the description of their 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:

$$\text{minimize} \quad \left\|A_\Omega - (XSY^T)_\Omega\right\|_F^2 \quad \text{for} \quad X\in\mathbb{R}^{m\times r}, Y\in\mathbb{R}^{n\times r}, S\in\mathbb{R}^{r\times r}, \tag{9}$$

using a gradient descent variant [25], 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 optimal solution. Output $X_* S_* Y_*^T$.

The last cleaning step resembles the optimization problem in alternating projections methods, but they used gradient descent instead. The authors [25] showed that the algorithm returns an output arbitrarily close to A, given *enough* iterations in the cleaning step, provided the following sampling size condition:

$$|\Omega| C \max \left\{ \mu_0 \sqrt{mn} \left(\frac{\sigma_1}{\sigma_r} \right)^2 r \log N, \quad \max\{\mu_0, \mu_1\}^2 r \min\{m, n\} \left(\frac{\sigma_1}{\sigma_r} \right)^6 \right\}. \tag{10}$$

The powers of r and $\log N$ are optimal by the coupon-collector limit, answering a question from [9]. On the other hand, the bound depends heavily on the condition number $\kappa := \sigma_1/\sigma_r$. Furthermore, similar to the situation in the previous subsection, one needs to know the rank r in advance, which is rarely the case in practice.

In a later paper [26], Keshavan and Oh showed that one can compute r (with high probability) if the condition number satisfies $\kappa = O(1)$; see also Remark 1.1. Thus, it seems that the critical extra assumption for this algorithm (apart from the three basic assumptions) to work is that the singular values of A are of the same order of magnitude ($\kappa = O(1)$). This assumption is strong, and we do not know how often it holds in practice. In many data sets, it has been notice that the leading singular value decay fast; see, for instance . Furthermore, it seems hard to decide from the input A_Ω that A has this property or not; so we do not know whether it is a good input for the algorithm.

One can try to compute the singular values of $p^{-1} A_\Omega$ as a prediction to those of A, but this prediction is only realiable for the those singular values of A which are already large, so the problem is sort of circular.

From the complexity point of view, the first part (low rank approximation) of the algorithm is very fast, as it used truncated SVD only once. On the other hand, the authors of [25] did not include a full convergence rate analysis of their gradient descent part, only briefly mentioning that quadratic convergence is possible.

1.4.4 Low rank approximation with rounding off

In this approach, one exploits the fact that A has finite precision; see subsection 1.3. It is clear that if each entry of A is an integer multiple of ε_0, then to achieve an exact recovery, it suffices to compute each entry with error less than $\varepsilon_0/2$, and then round it off. In other words, it is sufficient to obtain an approximation of A in the inifnity norm. It has been shown, under different extra assumptions, that low rank approximation fullfils this purpose.

The first infinity norm result was obtained by Abbe, Fan, Wang, and Zhong [1]. They showed that the rank-r approximation of $p^{-1} A_\Omega$ is close to A in the infinity norm [1, Theorem 3.4]. Technically, they proved that if $p \geq 6N^{-1} \log N$, then

$$\|p^{-1}(A_\Omega)_r - A\|_\infty \leq C \mu_0^2 \kappa^4 \|A\|_\infty \sqrt{\frac{\log N}{pN}},$$

for some universal constant C, provided $\sigma_r \geq C\kappa \|A\|_\infty \sqrt{\frac{N \log N}{p}}$, where $\kappa = \sigma_1/\sigma_r$ is the condition number.

If we turn this result into an algorithm (by simply rounding off the approximation), then we face the same problem as in the previous subsection, namely we needs to know the rank r and the condition number κ has to be small.

Getting rid of the condition number. Very recently, Bhardwaj and Vu [6] proposed and analysed the following algorithm, where they do not need to know the rank of A and get rid of the condition number, at the cost of a new assumption on the gaps between consecutive singular values. For simplicity, we state their result for matrices with integer entries ($\varepsilon = 1$). One can reduce the case of general case to this by scalling.

Algorithm 1.2 (Approximate-and-Round (**AR**))**.**

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 $\tilde{\sigma}_i \geq \frac{N}{8r\mu}$, where $\mu := N \max\{\|U\|_\infty^2, \|V\|_\infty^2\}$ is known.
3. Let $\hat{A} := \sum_{i=1}^{\tilde{s}} \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^T$.
4. Round off every entry of $\hat{A}$ to the nearest integer.

They showed that with probability $1 - o(1)$, before the rounding step, $\|\hat{A} - A\|_\infty < 1/2$, guaranteeing an exact recovery of A, under the following assumptions:

- *Low-rank:* $r = O(1)$.
- *Incoherence:* $\mu = O(1)$.
- *Sampling density:* $p \geq N^{-1} \log^{4.03} N$.
- *Bounded entries:* $\|A\|_\infty \leq K_A$ for a known constant K_A.
- *Gaps between singular values:* $\min_{i\in[s]}(\sigma_i - \sigma_{i+1}) \geq Cp^{-1} \log N$.

Aside from the first three traditional assumptions, the new assumption that the entries are bounded is standard for real-life datasets. In the step of finding the threshold, it seems that one needs to know both r and μ, but a closer look at the analysis reveals that it is possible to relax to knowing only their upper bounds. (We will do exactly this later in our algorithm, which is a variant of **AR**.)

The main improvement of **AR** over the previous spectral approaches is the removal of the dependence on the condition number. This removal was based on an entirely different mathematical analysis, which shows that the leading singular vectors of A and $p^{-1}A_\Omega$ are close in the infinity norm.

This removal, however, comes at the cost of the (new) gap assumption. While the required bound for the gaps is mild (much better than what one requires for the application of Davis-Kahan theorem; see [6] for more discussion), we do not know how often matrices in practice satisfy it, and again it is also hard to decide if the input matrix satisfies this requirement.

From the mathematical view point, it is interesting to notice that this assumption goes into the opposite direction of the small condition number assumption. Indeed, if the condition number is large, then the least singular value σ_r is considerably smaller than the largest one σ_1, which suggests, at the intuitve level at least, that the gaps between the consecutive singular values are large. So, mathematically, we have two valid theorems with two constrasting extra assumptions (beyond the three basic assumptions). The most logical explanation here should be that neither assumption is in fact needed. This conjecture, in the (more difficult) noisy setting presented in the next section, is the motivation of our study.

1.5 Recovery with Noise

Candes and Plan, in their influential survey [8], pointed out that data is often noisy, and a more realistic model of the recovery problem is to consider $A' = A + Z$, for A being the low rank ground truth matrix and Z the noise. We observe a sparse matrix A'_Ω, where each entry of A' is observed with probability p and set to 0 otherwise. In other words, we have access to a small random set of noisy entries. Notice that in this case, the truth matrix A is still low rank, but the noisy matrix A', the only thing we can observe, can have full rank. In what follow, we denote our input by $A_{\Omega,Z}$, emphasizing the presence of the noise.

Recovery from noisy observation is clearly a harder problem, and most papers concerning noisy recovery aim for recovery in the normalized Frobenius norm (root mean square error; RMSE), rather than exact recovery.

Continuing the nuclear norm minimization approach, Candes and Plan [8] adapted to the noisy situation by relaxing the constraint on the observations, leading to the following problem:

$$\text{minimize} \quad \|X\|_* \quad \text{subject to} \quad \|X_\Omega - A_{\Omega,Z}\|_F \leq \delta, \tag{11}$$

where δ is a known upper bound on $\|Z_\Omega\|_F$. The authors showed that, under the same sample size condition in [30], with probability $1 - o(1)$, the optimal solution $\hat{A}$ satisfies

$$\frac{1}{\sqrt{mn}}\|\hat{A} - A\|_F \leq C\|Z_\Omega\|_F\sqrt{\frac{\min\{m,n\}}{|\Omega|}}. \tag{12}$$

If one would like the RMSE to be at most ε, then one needs to require

$$|\Omega| \geq C\frac{\|Z_\Omega\|_F^2 \min\{m,n\}}{\varepsilon^2}, \tag{13}$$

which grows quadratically with $1/\varepsilon$.

For exact recovery, one needs to turn the approximation in the Frobenius norm into an approximation in the infinity norm; see subsection 1.3. This is a major mathematical challange, and in general, there is no efficient way to do this. The trivial bound that $\|M\|_\infty \geq \|M\|_F$ is too generous. If we use this and then use (12) to bound the RHS, then the corresponding bound on $|\Omega|$ in (13) becomes larger than mn, which is meaningless. This is the common situation with all Frobenius norm bounds discussed in this section.

Concerning the alternating method, a corollary of [21, Theorem 1] shows that we can obtain an approximation $\hat{A}$ of rank r, where

$$\|\hat{A} - A\| \leq (2 + o(1))\|Z\| + \varepsilon\sigma_1, \tag{14}$$

given that

$$p = \tilde{\Omega}\left(\frac{1}{n}\Big(1 + \frac{\|Z\|_F}{\varepsilon\sigma_1}\Big)\right)^2.$$

The bound here is in the spectral norm, and one can translate into Frobenius norm by the fact that $\|M\|_F \leq \sqrt{\operatorname{rank} M}\|M\|$. Again, it is not clear of how to obtain exact recovery from here.

Concerning the spectral approach, Keshavan, Montanari and Oh [24] also extended their result from [26] to the noisy case, using the same algorithm. They proved that with the same sample size condition as (10), the output satisfies w.h.p.

$$\|\hat{A} - A\|_F \leq C\left(\frac{\sigma_1}{\sigma_r}\right)^2 \frac{r^{1/2}mn}{|\Omega|}\|Z_\Omega\|_{op}. \tag{15}$$

If one would like to have $\frac{1}{\sqrt{mn}}\|\hat{A} - A\|_F \leq \varepsilon$, this translates to the following sample size condition:

$$|\Omega| \geq C\frac{\sigma^2 rN}{\varepsilon^2}\left(\frac{\sigma_1}{\sigma_r}\right)^2, \tag{16}$$

where the dependence on ε is again quadratic.

So far, the only approach which adapts well to the noisy situation is the infinity norm approach. As a matter of fact, the infinity norm bounds presented in the section 2.4 hold in both noiseless and noisy case (with some modification). The reason is that even in the noiseless case, one already views the (rescalled) input matrix $p^{-1}A_\Omega$ as the sum of A and a random matrix E. Thus, adding a new noise matrix Z just changes E to $E + Z$. This changes some parameters in the analysis, but all general mathematical arguments are still valid.

The result by Abbe et al. [1, Theorem 3.4] yields the same approximation as in the noiseless case, given

$$p \geq \frac{1}{\varepsilon^2}C^2\mu_0^4\kappa^8(\|A\|_\infty + \sigma_Z)^2 N^{-1}\log N, \tag{17}$$

where σ_Z is the standard deviation of each entry of Z. If we set $\varepsilon < \varepsilon_0/3$ (see subsection 1.3), then again rounding off would give as an exact recovery. Similarly, algorithm **AR** works in the noisy case; see [6] for the exact statement.

Summary. To summarize, in the noisy case, the infinity norm approach is currently the only one that yields exact recovery. The latest results in this directions, [1] and [6], however, requires the extra assumptions that the condition number is small and the gaps are large, respectively. As disucssed at the end of Section 2, these conditions constrast each other, and we conjecture that both of them could be removed. This leads to the main question of this paper:

Question 1. *Can we use the infinity norm approach to obtain exact recovery in the noisy case with only the three basic assumptions (low rank, incoherence, density) ?*

1.6 New results: an affirmative answer to Question 1

The main goal of this paper is to give an affirmative answer to Question 1, in a sufficiently general setting. We will show that a variant of Algorithm **AR** will do the job. The technical core is a new mathematical method to prove infinity norm estimates, which is entirely different from all previous techniques, and is of independent interest.

We would like to point out that this affirmative answer is not only a unifying result for the noisy case, but also an improvment for the noiseless case. In the noiseless, the only approach which does not need extra assumptions is the nuclear norm minimization (subsection 2.1). However, as discussed, the running time for this approach is not great, and various attemps have been made to improve the running time, leading to spectral algorithms such as those by Keshavan et al. Our algorithm is basically a truncated SVD, which is effective in both theory and practice. Compared to the spectral approach by Keshavan et al, it does not need the second, cleaning, phase. Finally, the depenpence of the density on the relevant parameters is comparable to all previous works; see Remarks 1.6 and 1.7.

Setting 1.3 (Matrix completion with noise)**.** Consider the truth matrix A, the observed set Ω, and noise matrix Z. We assume

1. *Known bound on entries:* We assume $\|A\|_\infty \le K_A$ for some known parameter K_A. This is the case for most real-life applications, as entries have physical meaning. For instance, in the Netflix Challenge $K_A = 5$.

2. *Known bound on rank:* We *do not assume* the knowledge of the rank r, but assume that we know some upper bound $r_{\max}$.

3. *Independent, bounded, centered noise:* Z has independent entries satisfying $\mathbf{E}[Z_{ij}] = 0$ and $\mathbf{E}\left[|Z_{ij}|^l\right] \le K_Z^l$ for all $l \in \mathbb{N}$ and $i \in [m]$, $j \in [n]$. We assume the knowledge of the upper bound K_Z. We do not require the entries to have the same distribution or even the same variance.

We allow the parameters r, $r_{\max}$, K_A, K_Z to depend on m and n. Under the setting above, we propose the following algorithm to recover A:

Algorithm 1.4 (AR2). Input: the $m \times n$ matrix $A_{\Omega,Z}$ and the discritization unit ε_0 of A's entries.

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

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

3. *Low-rank approximation:* Compute the truncated SVD $\hat{A}_{r_{\max}} = \sum_{i=1}^{r_{\max}} \hat{\sigma}_i \hat{u}_i \hat{v}_i^T$.

   Take the largest index $s \le r_{\max} - 1$ such that $\hat{\sigma}_s - \hat{\sigma}_{s+1} \ge 20(K_A + K_Z)\sqrt{\frac{r_{\max}(m+n)}{\hat{p}}}$.

   If no such s exists, take $s = r_{\max}$. Let $\hat{A}_s := \sum_{i \le s} \hat{\sigma}_i \hat{u}_i \hat{v}_i^T$,

4. *Rounding off:* Round each entry of $\hat{A}_s$ to the nearest multiple of ε_0. Return $\hat{A}_s$.

Compared to Approximate-and-Round (**AR**), a minor difference is that we use an estimate $\hat{p}$ of p, which is very accurate with high probability. We next use a different cutoff point for the truncated SVD step that does not require knowing the parameter μ_0.

From the complexity view point, both algorithms are very fast. It is basically just truncated SVD, taking only $O(|\Omega| r) = O(pmnr)$ FLOPs. Our main theorem below gives sufficient conditions for exact recovery (for the case of discrete entries).

Theorem 1.5. *There is a universal constant $C > 0$ such that the following holds. Suppose $r_{\max} \le \log^2 N$. Under the model 1.3, assume the following:*

- Large signal: $\sigma_1 \ge 100 r K \sqrt{\frac{r_{\max} N}{p}}$, *for* $K := K_A + K_Z$.

- Sampling density:

$$p \ge C\left(\frac{1}{m} + \frac{1}{n}\right) \max\left\{\log^4 N,\ \frac{r^3 K^2}{\varepsilon_0^2}\left(1 + \frac{\mu_0^2}{\log^2 N}\right)\right\} \log^6 N. \tag{18}$$

*Then with probability $1 - O(N^{-1})$, the first three steps of **AR2** recovers every entry of A within an absolute error $\varepsilon_0/3$. Consequently, if all entries are multiples integer of ε_0, the rounding-off step recovers A exactly.*

Notice that we have removed the gap condition from [6], and thus obtained an exact recovery algorithm, using only the three basic assumptions: low rank, incoherence, density.

Well, almost! The reader, of course, has noticed that we have a new spectral assumption that the leading singular value σ_1 of A has to be sufficiently large (large signal).

First, let us point out that an assumption on the magnitude of σ_1 is necessary. It is a well known phenomenon in engineering that if the intensity of the noise dominates the signal of the data, then the noisy data is often totally corrupted. From the mathematical view point, it is a well known fact, called the BBP threshold phenomenon [2], that if $\|Z\| \geq c\|A\| = c\sigma_1$, for a specific constant c, then the sum $A + Z$ behaves like a random matrix [2, 18, 29, 12, 5, 19]. For instance, the leading singular vectors of $A + Z$ look totally random and have nothing to do with the leading singular vectors of A. This shows that there is no chance that one can recover A from the (even fully observed) noisy matrix $A + Z$. In the results discussed earlier, this large signal assumption (if not explictly stated) is implicit in the setting of the model (see, for instance, the bound in the next paragraph).

Second, we emphasize that the bound required for σ_1 is very mild. In most cases, it is automatically satisfied by the simple fact that $\sigma_1^2 \geq r^{-1}\|A\|_F^2$. To see this, let us consider the base case when $n = \Theta(m)$, $\frac{1}{n}\|A\|_F = \Theta(1)$ (a normalization, for convenience), $r = O(1)$. In this case,

$$\sigma_1 \geq r^{-1}\|A\|_F = \tilde{\Omega}(n). \tag{19}$$

If we assume $K_A, K_Z = O(1)$, then our requirement on σ_1 in the theorem becomes

$$\sigma_1 = \Theta(\sqrt{n/p}),$$

which is guaranteed automatically (with room to spare) by (19), since we need $p = \Omega(\log n/n)$. As a matter of fact, the requirement $\sigma_1 = \Theta(\sqrt{n/p})$ is consistend with the BBP phenomenon discussed above. Indeed, if we consider the rescalled input

$$p^{-1}A_{\Omega,Z} = p^{-1}A_\Omega + p^{-1}Z,$$

then the $\sqrt{n/p}$ is the order of magnitude of the spectral norm of noisy part $\|\frac{1}{p}Z_\Omega\|$, and the spectral norm of $p^{-1}A_\Omega$ is approximately $\|A\|$, with high probability. (It is well known that one can approximate the spectral norm by random sampling.)

Remark 1.6 (Density bound)**.** The condition (18) looks complicated, but in the common case where A has constant rank, constantly bounded entries, and uniformly random singular vectors, we have $\mu_0 = O(\log N)$, $K_A, K_Z = O(1)$ and $r_{\max} = O(1)$, it reduces to

$$p \geq C \max\left\{\log^4 N,\ \varepsilon_0^{-2}\right\}\left(m^{-1} + n^{-1}\right)\log^6 N, \tag{20}$$

which is equivalent to $|\Omega| \geq CN \log^6 N \max\{\log^4 N, \varepsilon_0^{-2}\}$ in the uniform sampling model. The power of $\log N$ can be further reduced but the details are tedious, and the improvement is not really important from the pratical view point.

Even when reduced to the noiseless case, this bound is comparable, up to a polylogarithmic factor, to all previous results, while having the key advantage that it does not contain a power of the condition number.

Remark 1.7 (Quadratic growth in $1/\varepsilon_0$)**.** Notice that $|\Omega|$ is $O(N \log^{10} N)$ until $\varepsilon_0 < \log^{-2} N$, then it grows quadratically with $1/\varepsilon_0$. Interestingly, this quandratic dependence also appears in all results discussed in Section 1.5, for both RMSE and exact recoveries.

Remark 1.8 (Bound on $r_{\max}$)**.** The condition $r_{\max} \le \log^2 N$ in Theorem 1.5 can be avoided, at the cost of a more complicated sampling density bound. The full form of our bound is

$$p \ge C\left(\frac{1}{m} + \frac{1}{n}\right) \max\left\{\log^{10} N,\ \frac{r^4 r_{\max}\mu_0^2 K^2}{\varepsilon_0^2}, \frac{r^3 K^2}{\varepsilon_0^2}\left(1 + \frac{\mu_0^2}{\log^2 N}\right)\left(1 + \frac{r^3 \log N}{N}\right)\log^6 N\right\}. \tag{21}$$

The proof of the more general version of Theorem 1.5 with Eq. (21) replacing Eq. (18) will be in Appendix A. We do not know of a natural setting where one expects $r_{\max} > \log^2 N$. Nevertheless this shows that our technique does not require any extra condition besides the large signal assumption.

In the next section, we will prove Theorem 1.5, asserting the correctness of our algorithm, **AR2**. We will reframe the problem from a matrix perturbation perspective, then introduce the main technical results, which are extensions of the famous Davis-Kahan-Wedin theorem, and use them to prove Theorem 1.5. The rest of the paper is dedicated to proving the technical theorems.

2 Proof of guarantees for recovery

2.1 Proof sketch

The proof of Theorem 1.5 aims to bound the difference $\hat{A}_s - A$ in the infinity norm. Instead of comparing $\hat{A}_s$ with A directly, we use a series of intermediate comparisons, outlined below.

1. Let $\rho := \hat{p}/p$. We bound $\|A - \rho^{-1}A\|_\infty$. As shown by a Chernoff bound, ρ is very close to 1, making this error small. It remains to bound

   $$\|\rho^{-1}A - \hat{A}_s\|_\infty = \rho^{-1}\|A - (p^{-1}A_{\Omega,Z})_s\|_\infty.$$

   Let $\tilde{A} := p^{-1}A_{\Omega,Z}$, this is equivalent to bounding $\|A - \tilde{A}_s\|_\infty$, since $\rho = \Theta(1)$.

2. We bound $\|A - A_s\|_\infty$. Some light calculations give $\|A - A_s\|_\infty \le \sigma_{s+1}\|U\|_{2,\infty}\|V\|_{2,\infty}$. Using the fact that $\hat{\sigma}_i - \hat{\sigma}_{i+1}$ is small for all $i > s$, we can deduce that $\hat{\sigma}_{s+1}$ is small, making σ_{s+1} small too. Coupled with the incoherence property, this error will be small.

3. We bound $\|A_s - \tilde{A}_s\|_\infty$. Most of the heavy lifting is done here. We discuss it in detail below.

Observe that $\mathbf{E}[A_\Omega] = pA$ and $\mathbf{E}[Z_\Omega] = 0$, we have

$$\mathbf{E}\big[\tilde{A}\big] = \mathbf{E}\left[p^{-1}A_{\Omega,Z}\right] = \mathbf{E}\left[p^{-1}(A_\Omega + Z_\Omega)\right] = \mathbf{E}\left[p^{-1}A_\Omega\right] + \mathbf{E}\left[p^{-1}Z_\Omega\right] = A.$$

Let $E := \tilde{A} - A$. The above shows that E is a random matrix with mean 0. From a *matrix perturbation theory* point of view, $\tilde{A}$ is an *unbiased* perturbation of A. Establishing a bound on $(A+E)_s - A_s$ in the infinity norm is one of the major goals of perturbation theory, and is the main technical contribution of our paper.

Since both A_Ω and Z have independent entries, so does E. The entries of E satisfy

$$E_{ij} = \begin{cases} p^{-1}(A_{ij} + Z_{ij}) - A_{ij} = A_{ij}(p^{-1} - 1) + p^{-1}Z_{ij} & \text{with probability } p, \\ -A_{ij} & \text{with probability } 1 - p. \end{cases} \tag{22}$$

Consider the moments of E_{ij}. For each $l \geq 2$, e have

$$\begin{aligned}\mathbf{E}\left[|E_{ij}|^l\right] &= \frac{\mathbf{E}\left[|A_{ij}(1-p)+Z_{ij}|^l\right]}{p^{l-1}} + (1-p)|A_{ij}|^l \leq \frac{(K_A(1-p)+K_Z)^l}{p^{l-1}} + (1-p)K_A^l \\ &\leq \frac{1}{p^{l-1}}\left(\sum_{k=0}^{l-1}\binom{l}{k}K_A^k(1-p)^kK_Z^{l-k} + K_A^l(1-p)^l + p^{l-1}(1-p)K_A^l\right) \\ &\leq \frac{1}{p^{l-1}}\left(\sum_{k=0}^{l-1}\binom{l}{k}K_A^kK_Z^{l-k} + K_A^l\right) = \frac{(K_A+K_Z)^l}{p^{l-1}}.\end{aligned} \tag{23}$$

We have the following result for the third step:

Theorem 2.1. *Consider a fixed matrix $A \in \mathbb{R}^{m\times n}$. and a random matrix $E \in \mathbb{R}^{m\times n}$ with independent entries satisfying $\mathbf{E}[E_{ij}] = 0$ and $\mathbf{E}\left[|E_{ij}|^l\right] \leq p^{1-l}K^l$ for some $K > 0$ and $0 < p < 1$. Let $\tilde{A} = A + E$. Let $s \in [r]$ be an index satisfying*

$$\delta_s := \sigma_s - \sigma_{s+1} \geq 40rK\sqrt{N/p},$$

There are constant C and C' such that, if $p \geq C(m^{-1}+n^{-1})\log N$ where $N = m+n$, then

$$\|\tilde{A}_s - A_s\|_\infty \leq C'\frac{(\log N + \mu_0)\log^2 N}{\sqrt{mn}} \cdot r\sigma_s\left(\frac{K}{\sigma_s}\sqrt{\frac{N}{p}} + \frac{rK\sqrt{\log N}}{\delta_s\sqrt{p}} + \frac{r^2\mu_0 K\log N}{p\delta_s\sqrt{mn}}\right). \tag{24}$$

Remark 2.2. This theorem essentially gives a bound on the perturbation of the best rank-s approximation, given that the perturbation on A is random with sufficiently bounded moments and the singular values up to rank s are sufficiently separated from the rest. Consider the case $m = \Theta(n)$, making both $\Theta(N)$. Suppose either of the following two popular scenarios: A is completely incoherent, namely $\mu_0 = O(1)$; or A has independent Gaussian $N(0,1)$ entries, making $\mu_0 = O(\log N)$. The theorem is just the random version of the following:

$$\text{If } \delta_s \geq C_1 r\|E\|, \quad \text{then } \|\tilde{A}_s - A_s\|_\infty \leq \frac{C'\log^3 N}{N}r\sigma_s\left(\frac{\|E\|}{\sigma_s} + \frac{r\|U^TEV\|_\infty}{\delta_s}\right). \tag{25}$$

To see this, note that in the above setting, the first factor in the bound is simply $O\left(\frac{\log^3 N}{N}\right)$. As shown near the end (Section 5), with probability $1 - O(N^{-1})$,

$$\|E\| = O\left(K\sqrt{\frac{N}{p}}\right), \quad \|U^TEV\|_\infty = O\left(\frac{r\mu_0 K\log N}{p\sqrt{mn}} + \frac{K\sqrt{\log N}}{\sqrt{p}}\right).$$

Plugging the above into Eq. (25) recovers Eq. (24). Eq. (25) is important because it reveals the structure of the problem. Theorem 2.1 extends the classic *Davis-Kahan-Wedin theorem* for the perturbation of the best rank-s approximation in terms of the perturbation matrix E and the singular values of A. In Section 3, we will reveal the full form of our Davis-Kahan-Wedin extension, which is the main technical invention of this paper, compare it with previous results, and prove Theorem 2.1 with it.

Let us complete the proof of Theorem 1.5 now, asserting the correctness of **AR2**.

2.2 Proof of the matrix completion theorem

Proof of Theorem 1.5. For convenience, let $K := K_{A,Z} = K_A + K_Z$. Recall that in the discussion, we defined the following terms:

$$\rho := \frac{\hat{p}}{p}, \quad \tilde{A} = \rho \hat{A} = p^{-1} A_{\Omega,Z}, \quad E = \tilde{A} - A,$$

and E is a random matrix with mean 0, independent entries satisfying $\mathbf{E}\left[|E_{ij}|^l\right] = p^{1-l} K^l$ by Eq. (23). Recall that we have the SVD $\hat{A} = \sum_i \hat{\sigma}_i \hat{u}_i \hat{v}_i^T$. Similarly, denote the SVD of $\tilde{A}$ by

$$\tilde{A} = \sum_{i}^{\min\{m,n\}} \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^T.$$

We have the relation $\tilde{u}_i = \hat{u}_i$, $\tilde{v}_i = \hat{v}_i$ and $\tilde{\sigma}_i = \rho^{-1} \hat{\sigma}_i$ for each i.

From the sampling density assumption, a standard application of concentration bounds [22, 13] guarantees that, with probability $1 - O(N^{-2})$,

$$0.9 \le 1 - \frac{1}{\sqrt{N}} \le 1 - \frac{\log N}{\sqrt{pmn}} \le \rho \le 1 + \frac{\log N}{\sqrt{pmn}} \le 1 + \frac{1}{\sqrt{N}} \le 1.1. \tag{26}$$

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

$$\|E\| \le 2K\sqrt{N/p}, \tag{27}$$

with probability $1 - O(N^{-1})$. See [3, 35], [33, Lemma A.7] or [3] for detailed proofs. Therefore we can assume both Eqs. (26) and (27) at the cost of an $O(N^{-1})$ exceptional probability. The sampling density condition (Eq. (18) is equivalent to the conjunction of two conditions:

$$p \ge C\left(\frac{1}{m} + \frac{1}{n}\right) \log^{10} N, \tag{28}$$

$$p \ge \frac{Cr^3K^2}{\varepsilon^2}\left(1 + \frac{\mu_0^2}{\log^2 N}\right)\left(\frac{1}{m} + \frac{1}{n}\right) \log^6 N. \tag{29}$$

Before entering the three steps outlined in the proof sketch, we show that the SVD step is guaranteed to choose a valid $s \in [r]$ such that $\hat{\delta}_s \ge 20K\sqrt{r_{\max} N/\hat{p}}$. Choose an index $l \in [r]$ such that $\delta_l \ge \sigma_1/r$, which exists since $\sum_{l \in [r]} \delta_l = \sigma_1$. We have, by Weyl's inequality,

$$\tilde{\delta}_l \ge \delta_l - \|E\| \ge \frac{\sigma_1}{r} - 2K\sqrt{\frac{N}{p}} \ge (100 r_{\max}^{1/2} - 4)K\sqrt{\frac{N}{p}} \ge 90K\sqrt{\frac{r_{\max}N}{p}}.$$

Therefore

$$\hat{\delta}_l \ge \rho^{-1}\tilde{\delta}_l \ge 90\rho^{-1}K\sqrt{\frac{r_{\max}N}{p}} \ge 80\rho^{-1/2}K\sqrt{\frac{r_{\max}N}{p}} = 80K\sqrt{\frac{r_{\max}N}{\hat{p}}},$$

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

$$\tilde{\delta}_{r+1} \le \tilde{\sigma}_{r+1} \le \sigma_{r+1} + \|E\| = \|E\| \le 2K\sqrt{N/p}.$$

Therefore,

$$\hat{\delta}_{r+1} = \rho^{-1}\tilde{\delta}_{r+1} \le 2\rho^{-1}K\sqrt{\frac{N}{p}} \le 3\rho^{-1/2}K\sqrt{\frac{N}{p}} = 3K\sqrt{\frac{N}{\hat{p}}} < 20K\sqrt{\frac{r_{\max}N}{\hat{p}}}.$$

We want to show that the first three steps of **AR2** recover A up to an absolute error ε, namely $\|\hat{A}_s - A\|_\infty \le \varepsilon$. We will now follow the three steps in the sketch.

1. *Bounding* $\|A - \rho^{-1}A\|_\infty$. We have

$$\|A - \rho^{-1}A\|_\infty = \left|\rho^{-1} - 1\right| \|A\|_\infty \le \frac{K_A}{.9\sqrt{N}} < \varepsilon/4.$$

2. *Bounding* $\|A_s - A\|_\infty$. Firstly, we bound σ_{s+1}. We have, since $\tilde{\delta}_s$ is the last singular value gap of $\tilde{A}$ such that $\tilde{\delta}_s \ge 20\rho^{-1/2}K\sqrt{\frac{r_{\max}N}{p}}$, then

$$\sigma_{s+1} \le \tilde{\sigma}_{s+1} + \|E\| \le 20\rho^{-1/2}rK\sqrt{\frac{r_{\max}N}{p}} + 2K\sqrt{\frac{N}{p}} \le 22rK\sqrt{\frac{r_{\max}N}{p}}. \tag{30}$$

For each fixed indices j, k, we have

$$\begin{aligned}|(A_s - A)_{jk}| &= \left|U_{j,\cdot}^T \Sigma_{[s+1,r]} V_{k,\cdot}\right| \le \sigma_{s+1}\|U\|_{2,\infty}\|V\|_{2,\infty} \le 22rK\sqrt{\frac{r_{\max}N}{p}}\frac{r\mu_0}{\sqrt{mn}} \\ &= \sqrt{\frac{22^2 r^4 r_{\max}\mu_0^2 K^2}{p}\left(\frac{1}{m} + \frac{1}{n}\right)} \le \sqrt{\frac{22^2 r r_{\max}}{C\log^4 N}} \le \varepsilon/4,\end{aligned}$$

where the last inequality comes from Eq. (29) and the assumption that $r_{\max} \le \log^2 N$, if C is large enough. Since this holds for all pairs (j,k), we have $\|A_s - A\|_\infty \le \varepsilon/4$.

3. *Bounding* $\|\tilde{A}_s - A_s\|_\infty$, *with probability* $1 - O(N^{-1})$. The condition (28) guarantees that we can apply Theorem 2.1. We will get, by Eq. (24)

$$\|\tilde{A}_s - A_s\|_\infty = C'\frac{(\mu_0 + \log N)\log^2 N}{\sqrt{mn}} \cdot rK\left(\sqrt{\frac{N}{p}} + \frac{\log N}{p}\frac{\sigma_s}{\delta_s}\right). \tag{31}$$

Under the assumption $r_{\max} \le \log^2 N$, we can further simplify. We have

$$\delta_s \ge \tilde{\delta}_s - 2\|E\| \ge 20\rho^{-1/2}\sqrt{\frac{r_{\max}N}{p}} - 4K\sqrt{\frac{N}{p}} \ge 18\sqrt{\frac{r_{\max}N}{p}}.$$

Therefore, by Eq. (30), $\sigma_{s+1} < 2\delta_s$, so

$$\frac{\sigma_s}{\delta_s} = 1 + \frac{\sigma_{s+1}}{\delta_s} < 1 + 2r \le 3r.$$

Back to Eq. (31), consider the second factor of the right-hand side, we have

$$\frac{\log N}{p}\frac{\sigma_s}{\delta_s} \le \frac{3r\log N}{p} \le 3r\sqrt{\frac{N}{p}}\frac{\log N}{\sqrt{pN}} \le \frac{3r}{\log^4 N} \cdot \sqrt{\frac{N}{p}} < \sqrt{\frac{N}{p}}.$$

Therefore Eq. (31) becomes

$$\|\tilde{A}_s - A_s\|_\infty \le 2C'\frac{(\mu_0 + \log N)rK\sqrt{N}\log^2 N}{\sqrt{pmn}}.$$

We want this term to be at most $\varepsilon/4$. We have

$$\begin{aligned}2C'\frac{(\mu_0 + \log N)rK\sqrt{N}\log^2 N}{\sqrt{pmn}} \le \frac{\varepsilon}{4} &\Leftrightarrow p \ge (2C')^2 r^2 K^2(\mu_0 + \log N)^2 \cdot \frac{N}{mn}\log^4 N \\ &\Leftrightarrow p \ge (2C')^2 r^2 K^2\left(1 + \frac{\mu_0}{\log N}\right)^2\left(\frac{1}{m} + \frac{1}{n}\right)\log^6 N.\end{aligned}$$

This is satisfied by the condition (29), when C is large enough. The third step is complete.

Now we combined the three steps. We have, by triangle inequality,

$$\|\hat{A}_s - A\|_\infty = \|\rho^{-1}\tilde{A}_s - A\|_\infty \le \|A - \rho^{-1}A\|_\infty + \rho^{-1}\left(\|\tilde{A}_s - A_s\|_\infty + \|A_s - A\|_\infty\right)$$
$$\le \frac{\varepsilon}{4} + 1.2\left(\frac{\varepsilon}{4} + \frac{\varepsilon}{4}\right) < .9\varepsilon.$$

The total exceptional probability is $O(N^{-1})$. The proof is complete. □

3 Extension of the Davis-Kahan-Wedin theorem for the ∞-norm

Now that our matrix completion algorithm (**AR2**) has been verified, we will focus on proving Theorem 2.1. From this point onwards, let us put aside the matrix completion context and simply consider matrix perturbation model $\tilde{A} = A + E$.

Setting 3.1 (Matrix perturbation)**.** Consider a fixed $m \times n$ matrix A with SVD

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

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:

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

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 $\tilde{A}$.

3.1 The Davis-Kahan-Wedin theorem for the spectral norm

One of the most well-known results in perturbation theory is the **Davis-Kahan** $\sin\Theta$ **theorem** [16], 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 [36], states that:

$$\|\tilde{U}_s\tilde{U}_s^T - U_sU_s^T\| \vee \|\tilde{V}_s\tilde{V}_s^T - V_sV_s^T\| \le \frac{C\|E\|}{\delta_s} \quad \text{for a universal constant } C. \tag{32}$$

There are three challenges if one wants to apply this theorem in order to attack our problem:

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 [8], Eq. (32) works for all perturbation matrices E. Per the discussion in [34], the worst case (equality) only happens when there are special interactions between E and A. A series of papers by Vu and coauthors [28, 34] 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 [28] proved the following:

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

with high probability, effectively turning the *noise-to-gap* on the right-hand side of Eq. (32) into the *noise-to-signal ratio*, which can be much smaller than the former in many cases.

P. Tran and Vu then [34] 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

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

then

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

replacing the term $\|E\|^2$ in [28] with the smaller $y := \frac{1}{2}\max_{i \neq j}(|u_i^T E E^T u_j| + |v_i^T E^T E v_j|)$. This quantity is at most $\|E\|^2$. However, it can be significantly smaller in many cases, notably when E is *regular* [34], meaning there is a common $\overline{\sigma}$ such that:

$$\overline{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m} \mathbf{Var}\left[E_{ij}\right] = \frac{1}{n}\sum_{j=1}^{n} \mathbf{Var}\left[E_{ij}\right] \quad \text{for all } i \in [m],\ j \in [n]. \tag{33}$$

The notion of random regular matrices cover most models of random matrices used in practice, such as Wigner matrices or genearalized Wigner matrices.

Their method is easily adaptable to prove similar bounds for the deviation of other spectral entities, with respect to difference matrix norms. The main idea is to write the difference of the two entities as a contour integral, 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. See Section 3 of [34] for a detailed discussion.

3.2 A new Davis-Kahan bound for the infinity norm

In this paper, we adapt and further develop the method in [34] 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 3.2. *Consider the objects in Setting 3.1. Suppose some positive numbers τ_1, τ_2, and $\mathcal{H}$ satisfy $\|E\| \le \mathcal{H}$ and for all $0 \le a \le 10\log(m+n)$, we have:*

$$\begin{aligned} &\max\left\{\mathcal{H}^{-2a}\left\|(EE^T)^a U\right\|_{2,\infty},\ \mathcal{H}^{-(2a+1)}\left\|(EE^T)^a EV\right\|_{2,\infty}\right\} \le \tau_1\sqrt{r}, \\ &\max\left\{\mathcal{H}^{-2a}\left\|(E^TE)^a V\right\|_{2,\infty},\ \mathcal{H}^{-(2a+1)}\left\|(E^TE)^a E^TU\right\|_{2,\infty}\right\} \le \tau_2\sqrt{r}. \end{aligned} \tag{34}$$

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

$$\frac{\mathcal{H}}{\sigma_S} \vee \frac{2r\|U^TEV\|_\infty}{\Delta_S} \vee \frac{\sqrt{2r}\|E\|}{\sqrt{\Delta_S\lambda_S}} \le \frac{1}{8}, \tag{35}$$

Then there is a universal constant C such that

$$\left\|\tilde{V}_S\tilde{V}_S^T - V_SV_S^T\right\|_\infty \le C\tau_1^2 r\left(\frac{\mathcal{H}}{\sigma_S} + \frac{2r\|U^TEV\|_\infty}{\Delta_S} + \frac{2ry}{\Delta_S\sigma_S}\right) + \frac{1}{m+n}, \tag{36}$$

$$\left\|\tilde{V}_S\tilde{V}_S^T - V_SV_S^T\right\|_{2,\infty} \le C\tau_1 r\left(\frac{\mathcal{H}}{\sigma_S} + \frac{2r\|U^TEV\|_\infty}{\Delta_S} + \frac{2ry}{\Delta_S\sigma_S}\right) + \frac{1}{m+n}, \tag{37}$$

where

$$y := \frac{1}{2}\max_{i\neq j}(|u_i^TEE^Tu_j| + |v_i^TE^TEv_j|).$$

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

$$\|\tilde{A}_s - A_s\|_\infty \le C\tau_1\tau_2 r\sigma_s\left(\frac{\mathcal{H}}{\sigma_s} + \frac{2r\|U^TEV\|_\infty}{\delta_s} + \frac{2ry}{\delta_s\sigma_s}\right) + \frac{1}{m+n}. \tag{38}$$

Analogous bounds for U and $\tilde{U}$ hold, with U and V swapped. 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. Suppose the entries of E are independent with mean 0, variance at most ς^2 and almost sure upper bound M.

The term $\mathcal{H}$ plays the role of $\|E\|$, but defining it as an upper bound offers more flexibility for Eq. (34). Various works on random matrix norms [3, 35] give the estimate $\mathcal{H} = O(\varsigma\sqrt{m+n})$. The terms τ_1 and τ_2 are the *localization coefficients*. They play the roles of the coherence parameters in the matrix completion setting. A trivial choice is $\tau_1 = \tau_2 = 1/\sqrt{r}$, since we have

$$\|(EE^T)^aU\|_{2,\infty} \le \|E\|^{2a}, \quad \|(E^TE)^aEU\|_{2,\infty} \le \|E\|^{2a+1},$$

and analogously for V. However, they can be vastly improved when $\|U\|_{2,\infty}$ and $\|V\|_{2,\infty}$ are small. Our analysis later will give the estimates

$$\tau_1 = O\left(\log N\sqrt{\frac{\mu(U)}{m}} + \frac{\log^{3/2}N}{\sqrt{N}} + \frac{M\log^3 N}{\sqrt{N}}\cdot\sqrt{\frac{\mu(V)}{n}}\right),$$

and symmetrically for τ_2, where $\mu(U)$ and $\mu(V)$ are the individual incoherence parameters defined in Eq. (2). Assume a setting where $m = \Theta(n) = \Theta(N)$ and $K \le \sqrt{N}\log^{-2}N$. Suppose both $\mu(U = \Theta(\mu(V)) = \Theta(\mu_0)$, then the above is simply

$$\tau_1 = \Theta(\tau_2) = O\left(\log N\sqrt{\frac{\mu_0 + \log N}{N}}\right).$$

Observe that replacing $\mathcal{H}$ with $\|E\|$ in the three bounds recovers the term

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

in the Davis-Kahan bound on $\|\tilde{V}_s \tilde{V}_s^T - V_s V_s^T\|$ in [34]. In the ball part, our bounds are essentially:

$$\begin{aligned}
\left\|\tilde{V}_S \tilde{V}_S^T - V_S V_S^T\right\|_\infty &\leq C\frac{\mu(V)}{n}\left\|\tilde{V}_S \tilde{V}_S^T - V_S V_S^T\right\|, \\
\left\|\tilde{V}_S \tilde{V}_S^T - V_S V_S^T\right\|_{2,\infty} &\leq C\sqrt{\frac{\mu(V)}{n}}\left\|\tilde{V}_S \tilde{V}_S^T - V_S V_S^T\right\|, \\
\left\|\tilde{A}_s - A_s\right\|_\infty &\leq C\sqrt{\frac{\mu(U)\mu(V)}{mn}}\left\|\tilde{V}_s \tilde{V}_s^T - V_s V_s^T\right\|,
\end{aligned} \tag{39}$$

with some additional polylogarithmic factors.

The term $\|U^T EV\|_\infty$ is the maximum among sums of independent random variables, whose sizes are concentrated around $\varsigma(\sqrt{\log N} + M\|U\|_\infty\|V\|_\infty \log N)$ by the Bernstein bound [22, 13].

The term y has been analyzed in [34] and mentioned above. We use the trivial upper bound $\|E\|^2 = O(\varsigma^2(m+n))$, which is enough to prove Theorem 1.5.

The full "random" version of Theorem 3.2 looks like:

Theorem 3.3. *Consider the objects in Setting 3.1. Let $\varepsilon \in (0,1)$ be arbitrary Suppose E is a random $m \times n$ matrix with independent entries satisfying:*

$$\mathbf{E}\left[E_{ij}\right] = 0, \quad \mathbf{E}\left[|E_{ij}|^2\right] \leq \varsigma^2, \quad \mathbf{E}\left[|E_{ij}|^l\right] \leq M^{l-2}\varsigma^l \quad \textit{for all } \ p \in \mathbb{N}_{\geq 2}, \tag{40}$$

where M and ς are parameters. Let $N = m+n$. Define

$$\tau_1 := \frac{M\|U\|_{2,\infty} \log^3 N}{\sqrt{rN}} + \frac{\|V\|_{2,\infty} \log N}{\sqrt{r}} + \frac{\log^{3/2} N}{\sqrt{N}},$$

and define τ_2 symmetrically by swapping U and V. For an arbitrary subset $S \subset [r]$, suppose

$$\frac{\varsigma\sqrt{m+n}}{\sigma_S} \ \vee \ \frac{r\varsigma(\sqrt{\log N} + M\|U\|_\infty\|V\|_\infty \log N)}{\Delta_S} \ \vee \ \frac{\varsigma\sqrt{rN}}{\sqrt{\Delta_S \sigma_S}} \leq \frac{1}{16}. \tag{41}$$

Let

$$R_S := \frac{\varsigma\sqrt{N}}{\sigma_S} + \frac{r\varsigma(\sqrt{\log N} + M\|U\|_\infty\|V\|_\infty \log N)}{\Delta_S} + \frac{2r\varsigma^2 N}{\Delta_S \sigma_S}.$$

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

$$\left\|\tilde{V}_S \tilde{V}_S^T - V_S V_S^T\right\|_\infty \leq C\tau_1^2 r R_S + \frac{1}{N}, \tag{42}$$

$$\left\|\tilde{V}_S \tilde{V}_S^T - V_S V_S^T\right\|_{2,\infty} \leq C\tau_1 r R_S + \frac{1}{N}. \tag{43}$$

Analogous bounds for U and $\tilde{U}$ hold, with τ_2 replacing τ_1.

When $S = [s]$ for some $s \in [r]$, we slightly abuse the notation to let

$$R_s := R_{[s]} = \frac{\varsigma\sqrt{N}}{\sigma_s} + \frac{r\varsigma(\sqrt{\log N} + M\|U\|_\infty\|V\|_\infty \log N)}{\delta_s} + \frac{2r\varsigma^2 N}{\delta_s \sigma_s}.$$

Then with probability $1 - O(N^{-1})$,

$$\|\tilde{A}_s - A_s\|_\infty \le C\tau_1\tau_2 r\sigma_s R_s + \frac{1}{N}. \tag{44}$$

Furthermore, for each $\varepsilon > 0$, *if the term* $\frac{2r\varsigma^2 N}{\Delta_S \sigma_S}$ *in* R_S *is replaced with*

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

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

Going back to the matrix completion setting, we can use this theorem to prove Theorem 2.1.

Proof of Theorem 2.1. Let $\varsigma = K/\sqrt{p}$ and $M = 1/\sqrt{p}$. Then for C sufficiently large, $p \ge C(m^{-1} + n^{-1})\log^{10} N$ implies $M \le c\sqrt{N}\log^{-5} N$, meaning we can apply Theorem 3.3, specifically Eq. (44) for this choice of ς and M if the condition (41) holds. We check it for $S = [s]$. Given that $\sigma_s \ge \delta_s \ge 40K\sqrt{rN/p}$, we have

$$\frac{\varsigma\sqrt{N}}{\sigma_S} = \frac{K}{\sigma_S}\sqrt{\frac{rN}{p}} \le \frac{1}{40\sqrt{r}} < \frac{1}{16}, \qquad \frac{\varsigma\sqrt{rN}}{\sqrt{\delta_s\sigma_s}} \le \frac{K\sqrt{rN}}{\sqrt{p}\cdot 40rK\sqrt{rN/p}} \le \frac{1}{40} < \frac{1}{16},$$

and, using the fact $\mu_0 \le N$ and the assumption $r \le \log^2 N$,

$$\begin{aligned}
\frac{r\varsigma(\sqrt{\log N} + M\|U\|_\infty\|V\|_\infty \log N}{\delta_S} &\le \frac{rK\sqrt{\log N}}{\delta_s\sqrt{p}} + \frac{r^2 K\mu_0 \log N}{\delta_s p\sqrt{mn}} \\
\le \frac{\sqrt{r}\log N}{40\sqrt{N}} + \frac{r^{3/2}\mu_0\log N}{\sqrt{pmnN}} &\le \frac{\sqrt{r}\log N}{40\sqrt{N}} + \frac{r^{3/2}\mu_0\log N}{\sqrt{C}N\log^5 N} \le \frac{1}{\log N} < \frac{1}{16}.
\end{aligned}$$

It remains to transform the right-hand side of Eq. (44) to the right-hand side of Eq. (24). We have

$$\tau_1 \le \frac{\sqrt{\mu_0}\log^3 N}{\sqrt{pmN}} + \frac{\log^{3/2} N}{\sqrt{N}} + \frac{\sqrt{\mu_0}\log N}{\sqrt{n}} \le \frac{\log^{3/2} N}{\sqrt{N}} + \frac{\sqrt{2\mu_0}\log N}{\sqrt{n}}.$$

Combining with the symmetric bound for τ_2, we get

$$\tau_1\tau_2 \le \frac{\log^3 N}{N} + \frac{\sqrt{\mu_0}\log^{5/2} N}{\sqrt{N}} \cdot \frac{\sqrt{2m} + \sqrt{2n}}{\sqrt{mn}} + \frac{2\mu_0\log^2 N}{\sqrt{mn}} \le 4\log^2 N \frac{\log N + \mu_0}{\sqrt{mn}},$$

which is the first factor on the right-hand side of Eq. (24).

Consider the term R_s. From the above, we have

$$R_s \le \frac{K}{\sigma_S}\sqrt{\frac{rN}{p}} + \frac{rK\sqrt{\log N}}{\delta_s\sqrt{p}} + \frac{r^2K\mu_0\log N}{\delta_s p\sqrt{mn}} + \frac{K^2 rN}{p\delta_s\sigma_s}.$$

Since $\delta_s \ge 40K\sqrt{rN/p}$, the fourth term is at most $1/40$ of te first term. Removing it recovers exactly the second factor on the right-hand side of Eq. (24). The proof is complete. ☐

In the next section, we will prove Theorem 3.2. Then we will provide necessary probablistic bounds and prove Theorem 3.3 in Section 5. The technical bounds whose proof do not fit in the main body will be in Section B.

4 Proof of main results: the deterministic case

To prove Theorems 3.2, we are interested in the differences $\tilde{V}_S\tilde{V}_S^T - V_SV_S^T$ 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.

4.1 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

$$A_{\text{sym}} := \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}. \tag{45}$$

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 $\sigma_1, \ldots, \sigma_r$, but each with multiplicity 2, thus the matrices

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

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 $\{w_i\}_{i=1}^{2\,\mathrm{rank}\,M}$.

The main idea for proving Theorems 3.2 and 3.3 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 [34] 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

$$\sum_{i=1}^{m+n} \frac{\tilde{w}_i \tilde{w}_i^T}{z - \tilde{\lambda}_i} - \sum_{i=1}^{m+n} \frac{w_i w_i^T}{z - \lambda_i} = \sum_{\gamma=1}^{\infty} \left[\Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big) E_{\text{sym}} \right]^{\gamma} \Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big). \tag{46}$$

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

$$\begin{aligned} &\begin{bmatrix} \tilde{U}_S \tilde{U}_S^T - U_S U_S^T & 0 \\ 0 & \tilde{V}_S \tilde{V}_S^T - V_S V_S^T \end{bmatrix} = \tilde{W}_S \tilde{W}_S - W_S W_S^T \\ &= \sum_{\gamma=1}^{\infty} \oint_{\Gamma_S} \frac{\mathrm{d}z}{2\pi i} \left[\Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big) E_{\text{sym}} \right]^{\gamma} \Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big). \end{aligned} \tag{47}$$

We quickly note that the following identity can be obtained by multiplying both sides of Eq. (46) with z, dividing by $2\pi i$ and integrating over Γ_S:

$$\begin{aligned} &(\tilde{A}_S - A_S)_{\text{sym}} = (\tilde{A}_{\text{sym}})_S - (A_{\text{sym}})_S = \sum_{i \in S \vee i - r \in S} \left(\frac{z \tilde{w}_i \tilde{w}_i^T}{z - \tilde{\lambda}_i} - \frac{z w_i w_i^T}{z - \lambda_i} \right) \\ &= \sum_{\gamma=1}^{\infty} \oint_{\Gamma_S} \frac{z\mathrm{d}z}{2\pi i} \left[\Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big) E_{\text{sym}} \right]^{\gamma} \Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big). \end{aligned} \tag{48}$$

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 $\nu \in \mathbb{N}$:

$$\mathcal{T}_\nu := \oint_{\Gamma_S} \frac{z^{\nu}\mathrm{d}z}{2\pi i} \left[\Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big) E_{\text{sym}} \right]^{\gamma} \Big(\sum_{\lambda_i \neq 0} \frac{w_i w_i^T}{z - \lambda_i} + \frac{I - WW^T}{z} \Big). \tag{49}$$

Our common strategy will establish a bound on this generic form, which implies Eqs. (36) and (37) for $\nu = 0$ and Eq. (38) 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. (49), we get

$$\mathcal{T}_\nu = \sum_{\gamma=1}^{\infty} \oint_{\Gamma_S} \frac{z^\nu \mathrm{d}z}{2\pi i} \left[\Big(\sum_{\lambda_i \neq 0} \frac{\lambda_i P_i}{z(z-\lambda_i)} + \frac{I}{z} \Big) E_{\text{sym}} \right]^\gamma \Big(\sum_{\lambda_i \neq 0} \frac{\lambda_i P_i}{z(z-\lambda_i)} + \frac{I}{z} \Big). \tag{50}$$

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

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

which can be rewritten as

$$\mathcal{C}_\nu(\mathbf{I}) E_{\text{sym}}^{\alpha_0} \left[\prod_{k=1}^{h-1} \mathcal{M}(\mathbf{i}_k) E_{\text{sym}}^{\alpha_k+1} \right] \mathcal{M}(\mathbf{i}_h) E_{\text{sym}}^{\alpha_h}, \tag{51}$$

where we denote

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

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

$$\mathcal{M}(\mathbf{i}) := P_{i_1} \prod_{j=2}^{\beta} E_{\text{sym}} P_{i_j}, \tag{52}$$

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

$$\mathcal{C}_\nu(\mathbf{I}) := \oint_{\Gamma_S} \frac{\mathrm{d}z}{2\pi i} \frac{1}{z^{\gamma+1}} \prod_{k=1}^{h} \prod_{j=1}^{\beta_k} \frac{\lambda_{i_{kj}}}{z-\lambda_{i_{kj}}}. \tag{53}$$

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:

(54)
- $\alpha_0, \alpha_h \geq 0$, and $\alpha_k \geq 1$ for $1 \leq k \leq h-1$,
- $\beta_k \geq 1$ for $1 \leq k \leq h$,
- $\alpha + \beta = \gamma + 1$, where $\alpha := \sum_{k=0}^{h} \alpha_k$, and $\beta := \sum_{k=1}^{h} \beta_k$.

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. (50), (53), and (52), we get the expansion

$$\mathcal{T}_\nu = \sum_{\gamma=1}^{\infty} \mathcal{T}_\nu^{(\gamma)}, \quad \text{where } \mathcal{T}_\nu^{(\gamma)} = \sum_{h=0}^{\lfloor \gamma/2 \rfloor + 1} \mathcal{T}_\nu^{(\gamma,h)}, \quad \text{where } \mathcal{T}_\nu^{(\gamma,h)} = \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \Pi_h(\gamma)} \mathcal{T}_\nu(\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{55}$$

where

$$\mathcal{T}_\nu(\boldsymbol{\alpha},\boldsymbol{\beta}) := \sum_{\mathbf{I} \in [2r]^{\beta_1 + \dots + \beta_h}} \mathcal{C}_\nu(\mathbf{I}) E_{\text{sym}}^{\alpha_0} \left[\prod_{k=1}^{h-1} \mathcal{M}(\mathbf{i}_k) E_{\text{sym}}^{\alpha_k+1} \right] \mathcal{M}(\mathbf{i}_h) E_{\text{sym}}^{\alpha_h}. \tag{56}$$

Let us look at the main theorem again. Consider Eqs. (36) and (38). 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. (37), 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. (55) and (56), this entry will can be written as

$$e_{m+n,m+j}^T \mathcal{T}_0 e_{m+n,m+k} = \sum_{\gamma=1}^{\infty} \sum_{h=0}^{\lfloor\gamma/2\rfloor+1} \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Pi_h(\gamma)} e_{m+n,m+j}^T \mathcal{T}_0(\boldsymbol{\alpha},\boldsymbol{\beta}) e_{m+n,m+k}. \tag{57}$$

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 4.1. 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 $i \le r-1$, $\delta_r = \lambda_r > 0$, and $\lambda_i = -\lambda_{i-r}$ for $i \ge r+1$.
- $W = \{w_i\}_{i\in[2r]}$ 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 $2h-1 \le \gamma+1$.
- $\Pi_h(\gamma)$: the set of pairs of index sequences $(\boldsymbol{\alpha},\boldsymbol{\beta})$ satisfying Eq. (54).
- $\mathcal{C}_\nu = \mathcal{C}_{\nu,\boldsymbol{\Lambda},S} : \bigcup_{\beta=0}^{\gamma+1}[2r]^\beta \to \mathbb{R}$: the mapping from an index sequence $\mathbf{I}$ to its integral coefficient defined in Eq. (53).
- E: an arbitrary matrix in $\mathbb{R}^{m\times n}$, and let E_{sym} be defined analogously to Eq. (45).
- $\mathcal{M} = \mathcal{M}_{W,E} : \bigcup_{\beta=0}^{\gamma+1}[2r]^\beta \to \mathbb{R}^{(m+n)\times(m+n)}$: the mapping from an index sequence $\mathbf{i}$ to its monomial matrix defined in Eq. (52).
- M and M': arbitrary matrices with $m+n$ rows.
- $\mathcal{T}_\nu$: the target sum defined in Eq. (55). 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. (56).

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.

4.2 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 4.2. *Consider the objects defined in Setting 4.1 and an arbitrary tuple* $\mathbf{I} := \{i_k\}_{k\in[\beta]} \in [2r]^\beta$ *and denote the following:*

$$\beta_S(\mathbf{I}) := |\{1 \le k \le \beta : \ i_k \in S\}|,$$
$$\lambda_S(\mathbf{I}) := \min\{|\lambda_{i_k}| : k \in S\},$$
$$\Delta_S(\mathbf{I}) := \min\{|\lambda_{i_k} - \lambda_{i_l}| : i_k \in S, i_l \notin S\}.$$

We have,

$$|\mathcal{C}_\nu(\mathbf{I})| \le \mathrm{L}_\nu(\mathbf{I}) \left(1 + \frac{\Delta_S(\mathbf{I})}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \frac{1}{\lambda_S(\mathbf{I})^{\gamma+1-\beta} \Delta_S(\mathbf{I})^{\beta-1}}, \tag{58}$$

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* $s \in [r]$*:*

$$|\mathcal{C}_\nu(\mathbf{I})| \le \mathrm{L}_\nu \left(1 + \frac{\Delta_S}{\lambda_S}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \frac{1}{\lambda_S^{\gamma+1-\beta} \Delta_S^{\beta-1}}, \tag{59}$$

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

Proof. See Section B.1. □

Note that in order for Eq. (59) 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. (59), with one exception where the more precise Eq. (58) 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 4.3. *Consider objects in Setting 4.1 and Lemma 4.2. Assume that either* $\nu = 0$ *or* $\nu = 1$ *and* $S = [s]$ *for some* $s \in [r]$*. For each* $\gamma \ge 1$*,* $0 \le h \le \gamma/2 + 1$*,* $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \Pi_h(\gamma)$*, we have*

$$\begin{aligned}\|\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})\| \le \mathrm{L}_\nu &\binom{\gamma + \beta - 2}{\beta - 1} \|W^T E_{\text{sym}} W\|_\infty^{\beta-h} \|E\|^{\alpha - \alpha_0 - \alpha_h + h - 1} \frac{(2|S| + 2\rho|S^c|)^{\beta-2}}{\lambda_S^{\gamma+1-\beta} \Delta_S^{\beta-1}} \\ &\cdot \sum_{i=1}^{2r} \rho^{\mathbf{1}\{i_k \in S^c\}} \left\|w_i^T E_{\text{sym}}^{\alpha_0} M\right\| \cdot \sum_{i=1}^{2r} \rho^{\mathbf{1}\{i_k \in S^c\}} \left\|w_i^T E_{\text{sym}}^{\alpha_h} M'\right\|,\end{aligned} \tag{60}$$

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

$$\begin{aligned}\|\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta})\| \le \mathrm{L}_\nu &\binom{\gamma + \beta - 2}{\beta - 1} \|W^T E_{\text{sym}} W\|_\infty^{\beta-h} \|E\|^{\alpha - \alpha_0 - \alpha_h + h - 1} \\ &\cdot \frac{(2r)^{\beta-2}}{\lambda_S^{\gamma+1-\beta} \Delta_S^{\beta-1}} \cdot \sum_{i=1}^{2r} \left\|w_i^T E_{\text{sym}}^{\alpha_0} M\right\| \cdot \sum_{i=1}^{2r} \left\|w_i^T E_{\text{sym}}^{\alpha_h} M'\right\|.\end{aligned} \tag{61}$$

Since the bound in Eq. (60) is rather cumbersome, we will mostly use Eq. (61). Eq. (60) is still useful for future development in this direction.

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

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

where we temporarily denote $X := W^T E_{\text{sym}} W$. From Eq. (56), we thus have

$$\mathcal{T}_\nu(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{\mathbf{I} \in [2r]^\beta} \mathcal{C}_\nu(\mathbf{I}) D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{I}) \left(M^T E_{\text{sym}}^{\alpha_0} w_{i_1} \right) \left(w_{i_{h\beta_h}}^T E_{\text{sym}}^{\alpha_h} M' \right), \tag{62}$$

where

$$D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{I}) := \prod_{k=1}^{h} \prod_{j=2}^{\beta_k} X_{i_{k(j-1)} i_{kj}} \cdot \prod_{k=1}^{h-1} w_{i_{k\beta_k}}^T E_{\text{sym}}^{\alpha_k + 1} w_{i_{(k+1)1}}. \tag{63}$$

We can obtain an upper bound for $|D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{I})|$ by replacing every term in the first product of Eq. (63) with $\|X\|_\infty$, and replacing every instance of E with $\|E\|$ in the second product. Therefore

$$|D(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{I})| \le \|X\|_\infty^{\beta - h} \|E\|^{\alpha - \alpha_0 - \alpha_h + h - 1}. \tag{64}$$

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

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

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 4.2, we have

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

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

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

The proof of Eq. (60) is complete. Eq. (61) follows from the fact that $\rho \le 1$. □

To bring the bound above closer to the form required in Theorem 3.2, 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 τ_1 and τ_2. 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 (34) in Theorem 3.2. The bound below is a direct result.

Lemma 4.4. *Consider the objects in Setting 4.1 and terms τ_1, τ_2 and $\mathcal{H}$ from Theorem 3.2. Then for a matrix $M \in \{e_{m+n,k}\}_{k\in[m+n]} \cup \{I_{m+n}\}$, we have $\sum_{i=1}^{2r} \left\|w_i^T E_{\text{sym}}^{\alpha} M\right\| \le 2r\tau\mathcal{H}_0^{\alpha}$, where*

- $\tau = \frac{1}{\sqrt{2}}\tau_1$ *and* $\mathcal{H}_0 = \mathcal{H}$ *for* $M \in \{e_{m+n,k}\}_{k\in[[m+1,m+n]]}$.
- $\tau = \frac{1}{\sqrt{2}}\tau_2$ *and* $\mathcal{H}_0 = \mathcal{H}$ *for* $M \in \{e_{m+n,k}\}_{k\in[m]}$.
- $\tau = 1$ *and* $\mathcal{H}_0 = \|E\|$ *for* $M = I_{m+n}$.

In the upcoming third step, the formulas of τ 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.

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

Consider Eq. (55) again. The series we need to bound has the form

$$\mathcal{T} = \sum_{\gamma=1}^{\infty} \mathcal{T}_{\nu}^{(\gamma)}, \quad \text{where } \mathcal{T}_{\nu}^{(\gamma)} = \sum_{h=0}^{\lfloor\gamma/2\rfloor+1} \mathcal{T}_{\nu}^{(\gamma,h)}, \quad \text{where } \mathcal{T}_{\nu}^{(\gamma,h)} = \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Pi_h(\gamma)} \mathcal{T}\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{65}$$

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. (36), (37) and (38). We use the assumption below.

Assumption 4.5. *Consider the objects given by Setting 4.1. Assume that the real numbers L_{ν}, τ, τ', $\mathcal{H}_0$ and $\mathcal{H}_0'$ satisfy*

$$\forall S \subset [2r], \mathbf{I} \in [2r]^{\beta}: \qquad |\mathcal{C}_{\nu}(\mathbf{I})| \le \mathrm{L}_{\nu}\left(1 + \frac{\Delta_S}{\lambda_S}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \frac{1}{\lambda_S^{\gamma+1-\beta}\Delta_S^{\beta-1}}, \tag{66}$$

$$\forall \alpha \in [\lceil 10\log(m+n)\rceil]: \qquad \sum_{i=1}^{2r} \|w_i^T E_{\text{sym}}^{\alpha} M\| \le 2r\tau\mathcal{H}_0^{\alpha} \quad \textit{and} \quad \sum_{i=1}^{2r} \|w_i^T E_{\text{sym}}^{\alpha} M'\| \le 2r\tau'\mathcal{H}_0'^{\alpha}. \tag{67}$$

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

$$\frac{\|E\|}{\lambda_S} \vee \frac{\mathcal{H}_0}{\lambda_S} \vee \frac{\mathcal{H}_0'}{\lambda_S} \vee \frac{2r\|W^T E_{\text{sym}} W\|_{\infty}}{\Delta_S} \le R_1, \qquad \frac{\sqrt{2r}\|E\|}{\sqrt{\Delta_S \lambda_S}} \le R_2, \qquad R_1 \vee R_2 \le R \le \frac{1}{8}. \tag{68}$$

Additionally, let R_3 be a real number such that

$$\frac{2r}{\Delta_S \sigma_S} \max_{|i-j|\ne 0,r} \left|w_i E_{\text{sym}}^2 w_j\right| \le R_3 \le R_2^2. \tag{69}$$

Lemma 4.6. *Under Setting 4.1 and Assumption 4.5, for $\mathcal{T}_{\nu}^{(\gamma)}$ defined in Eq. (65), we have*

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

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 (67) holds. Lemma 4.3 gives

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

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

$$\|\mathcal{T}_\nu(\boldsymbol{\alpha},\boldsymbol{\beta})\| \le 2r\mathrm{L}_\nu \tau\tau' \binom{\gamma+\beta-2}{\beta-1} \left[\frac{\|E\|}{\lambda_S}\right]^{\gamma_1} \left[\frac{\mathcal{H}_0}{\lambda_S}\right]^{\alpha_0} \left[\frac{\mathcal{H}_0'}{\lambda_S}\right]^{\alpha_h} \left[\frac{2r\|X\|_\infty}{\Delta_S}\right]^{\gamma_2} \left[\frac{\sqrt{2r}\|E\|}{\sqrt{\Delta_S \lambda_S}}\right]^{\gamma_3}, \tag{70}$$

where we temporarily define the following:

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

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

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

Plugging this bound into Eq. (65), we get

$$\begin{aligned}\left\|\mathcal{T}_\nu^{(\gamma)}\right\| &\le 2r\mathrm{L}_\nu\tau\tau' \sum_{h=1}^{\lfloor\gamma/2\rfloor+1} \sum_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\Pi_h(\gamma)} \binom{\gamma+\beta-2}{\beta-1} R_1^{\gamma-2h+2} R_2^{2h-2} \\ &\le 2r\mathrm{L}_\nu\tau\tau' \sum_{h=1}^{\lfloor\gamma/2\rfloor+1} \sum_{\beta=h}^{\gamma+2-h} \binom{\gamma+\beta-2}{\beta-1} R_1^{\gamma-2h+2} R_2^{2h-2} \left|\Pi_h(\gamma,\beta)\right|,\end{aligned} \tag{71}$$

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

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

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. (71), we obtain

$$\begin{aligned}\left|\mathcal{T}_\nu^{(\gamma)}\right| &\le 2r\mathrm{L}_\nu\tau\tau' \sum_{h=1}^{\lfloor\gamma/2\rfloor+1} \sum_{\beta=h}^{\gamma+2-h} \binom{\gamma+\beta-2}{\beta-1}\binom{\beta-1}{h-1}\binom{\gamma+2-\beta}{h} R_1^{\gamma-2h+2} R_2^{2h-2} \\ &= 2r\mathrm{L}_\nu\tau\tau' \sum_{\beta=1}^{\gamma+1} \binom{\gamma+\beta-2}{\beta-1} \sum_{h=1}^{\beta\wedge(\gamma+2-\beta)} \binom{\beta-1}{h-1}\binom{\gamma+2-\beta}{h} R_1^{\gamma-2h+2} R_2^{2h-2}.\end{aligned} \tag{72}$$

Consider two cases for h and γ:

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

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

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. (56) 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

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

where we use the same computations leading up to Eq. (72), 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. (58) in Lemma 4.2, 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

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

Therefore the total contribution of this subcase is at most

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

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

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

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

Summing up the contributions from both cases, we obtain

$$\left\|\mathcal{T}_\nu^{(\gamma)}\right\| \le r\mathrm{L}_\nu \tau\tau' \left[9R_1(6R)^{\gamma-1} + \mathbf{1}\{\gamma \text{ even}\}\left(4(R_1\sqrt{27/2})^{\gamma} + 27R_3(R_2\sqrt{27/4})^{\gamma-2}\right)\right]. \tag{73}$$

Now let us consider the case $\gamma > 10\log(m+n)$. Instead of using the bound in Eq. (67), 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 $\tau = \|M\|$, $\tau' = \|M'\|$, and $\mathcal{H} = \mathcal{H}' = \|E\|$. The conditions on R_1, R_2 and R_3 remain the same so they still hold. Thus we still have Eq. (73), with the substitutions above. The proof is complete. □

Lemma 4.7. *Under Setting 4.1 and Assumption 4.5, for $\mathcal{T}_\nu$ defined in Eq. (65), we have*

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

Proof. For convenience, let $N = \lfloor 10\log(m+n)\rfloor$. Applying Lemma 4.6, we have

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

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. □

4.4 Conclusion: Proof of Theorem 3.2

We are now ready to use the common three-step strategy above to prove Theorem 3.2.

Proof of Theorem 3.2. Consider the objects defined in Theorem 3.2 and the additional objects in Setting 4.1. 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. (36). 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 4.1 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 4.7. Let us choose the parameters to satisfy Assumption 4.5. For $\nu = 0$, Lemma 4.2 guarantees $\mathrm{L}_\nu = 2$ satisfies Eq. (66). For the above choices of M and M', Lemma 4.4 guarantees that the choices $\tau = \tau' = \frac{1}{\sqrt{2}}\tau_1$ and $\mathcal{H}_0 = \mathcal{H}_0' = \mathcal{H}$ satisfy Eq. (67).

For convenience, define the following shorthands:

$$R_1 := \frac{\|E\|}{\sigma_S} \vee \frac{2r\|U^T E V\|_\infty}{\Delta_S}, \quad R_2 := \frac{\sqrt{2r}\|E\|}{\sqrt{\sigma_S \Delta_S}}, \quad R_3 := \frac{r \max_{i\neq j}(|u_i E E^T u_j| + |v_i E^T E v_j|)}{\Delta_S \sigma_S}. \tag{74}$$

Then the assumption (35) is just $R_1 \vee R_2 \leq 1/8$, and the term R_S is simply $R_1 + R_3$.

Next, we show that the terms R_1, R_2 and R_3 satisfy Eqs. (68) and (69). If this holds then we can apply Lemma 4.7. It suffices to check the conditions on R_1 and R_3. The former follows from the fact $\|W^T E_{\text{sym}} W\|_\infty \leq \|U^T E V\|_\infty$, which follows from $w_i E_{\text{sym}} w_j = (u_i^T E v_j + u_j^T E v_i)/2$. 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 $w_i^T E_{\text{sym}}^2 w_j = \frac{1}{2}(u_{i-r}^T E E^T u_j - v_{i-r}^T E^T E v_j)$.
- $j \in [r]$ and $i \in [r] \setminus \{j\}$. Then $w_i^T E_{\text{sym}}^2 w_j = \frac{1}{2}(u_i^T E E^T u_j + v_i^T E^T E v_j)$.
- $j \in [[r+1, 2r]]$ and $i \in [[r+1, 2r]] \setminus \{j\}$. Then $w_i^T E_{\text{sym}}^2 w_j = \frac{1}{2}(u_{i-r}^T E E^T u_{j-r} + v_{i-r}^T E^T E v_{j-r})$.

From the above, it follows that

$$\max_{|i-j|\neq 0, r} |w_i^T E_{\text{sym}}^2 w_j| \leq \frac{1}{2} \max_{i\neq j}(|u_i E E^T u_j| + |v_i E^T E v_j|),$$

proving Eq. (69) for this choice of R_3. Plugging in these values into the bound in Lemma 4.7 gives

$$\left| \left(\tilde{V}_S \tilde{V}_S^T - V_S V_S^T \right)_{jk} \right| \leq 4r\tau_1^2 (18R_1 + 27R_3) + \frac{2r\|M\|\|M'\|}{(m+n)^{2.5}} \leq C_0 r \tau_1^2 (R_1 + R_3) + \frac{1}{m+n},$$

for a constant C_0. The above holds uniformly over all $j, k \in [n]$, so it holds for the infinity norm, proving Eq. (36).

Let us prove Eq. (37). 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 4.1 gives

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

for $M = e_{m+n,j+m}$, $M' = I_{m+n}$, and $\mathcal{C}$ from Eq. (53). From the previous argument, we can choose L_ν, τ and $\mathcal{H}_0$ the same way, and choose $\tau' = 1$ and $\mathcal{H}_0' = \|E\|$, which trivially satisfy the condition on M'. The same choice of R_1, R_2 and R_3 also satisfy Eqs. (68) and (69). Therefore

$$\left\| \left(\tilde{V}_S \tilde{V}_S^T - V_S V_S^T \right)_{j,\cdot} \right\| \leq C_0 r \tau_1 (R_1 + R_3) + \frac{1}{m+n},$$

which holds uniformly over $j \in [n]$, proving Eq. (37)..

Let us prove Eq. (38). 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 4.1 gives

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

We choose $\mathrm{L}_\nu = \lambda_s = \sigma_s$, $\tau = \frac{1}{\sqrt{2}}\tau_2$, $\tau' = \frac{1}{\sqrt{2}}\tau_1$, $\mathcal{H}_0 = \mathcal{H}$. 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 4.5 and Lemma 4.7, so we have

$$\left|\left(\tilde{A}_s\tilde{A}_s^T - A_sA_s^T\right)_{jk}\right| \leq C_0 r\tau_1\tau_2\sigma_s(R_1+R_3) + \frac{1}{m+n},$$

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. $\square$

5 Proof of main results: the random case

In this section, we prove Theorem 3.3. 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 3.2. These terms are:

- $\|E\|$. There are tight bounds in the literature. For E following the Model (40), with the assumption $M \leq (m+n)^{1/2}\log^{-5}(m+n)$, the moment argument in [35] can be used.
- $\|U^TEV\|_\infty = \max_{i,j}|u_i^TEv_j|$. These terms can be bounded with a simple Bernstein 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.
- τ_1, τ_2 and $\mathcal{H}$ 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 5.1. *Consider the objects in Setting 3.1. Let $E \in \mathbb{R}^{m\times n}$ be a random matrix satisfying Model (40) with parameters M and ς. Suppose $M \leq (m+n)^{1/2}\log^{-3}(m+n)$. Then with probability $1 - O((m+n)^{-2})$, all of the following hold:*

$$\|E\| \leq 2\varsigma\sqrt{m+n}, \tag{75}$$

$$\max_{i\neq j}(|u_iEE^Tu_j| + |v_iE^TEv_j|) \leq 2\|E\|^2 \leq 8\varsigma^2(m+n). \tag{76}$$

$$\max_{i,j}|u_i^TEv_j| \leq 2\varsigma(\sqrt{\log(m+n)} + M\|U\|_\infty\|V\|_\infty\log(m+n)). \tag{77}$$

Proof. Eq. (75) follows from the moment argument in [35]. Eq. (76) follows from Eq. (75). It remains to check Eq. (77). Fix $i, j \in [r]$. Write

$$u_i^TEv_j = \sum_{k\in[m],h\in[n]} u_{ik}v_{jh}E_{kh} = \sum_{(k,h)\in[m]\times[n]} Y_{kh},$$

where we temporarily let $Y_{kh} := u_{ik}v_{jh}E_{kh}$ for convenience. We have $|Y_{kh}| \leq \|U\|_\infty\|V\|_\infty|E_{kh}|$. Let $X_{kh} := Y_{kh}/(\varsigma\|U\|_\infty\|V\|_\infty)$, then $\{X_{kh} : (k,h) \in [m]\times[n]\}$ are independent random variables and for each $(k,h) \in [m]\times[n]$,

$$\mathbf{E}\left[X_{kh}\right] = 0, \quad \mathbf{E}\left[|X_{kh}|^2\right] \leq 1, \quad \mathbf{E}\left[|X_{kh}|^l\right] \leq M^{l-2} \text{ for all } l \in \mathbb{N}.$$

We also have

$$\sum_{k,h} \mathbf{E}\left[|X_{kh}|^2\right] = \frac{\sum_{k,h} u_{ik}^2 v_{jh}^2 \mathbf{E}\left[|E_{kh}|^2\right]}{\varsigma^2 \|U\|_\infty^2 \|V\|_\infty^2} \le \frac{\varsigma^2 \sum_{k,h} u_{ik}^2 v_{jh}^2}{\|U\|_\infty^2 \|V\|_\infty^2} = \frac{1}{\|U\|_\infty^2 \|V\|_\infty^2}$$

By Bernstein's inequality [13], we have for all $t > 0$

$$\mathbf{P}\left(\left|\sum_{k,h} X_{kh}\right| \ge t\right) \le \exp\left(\frac{-t^2}{\sum_{k,h} \mathbf{E}\left[|X_{kh}|^2\right] + \frac{2}{3}Mt}\right) \le \exp\left(\frac{-t^2}{\|U\|_\infty^{-2}\|V\|_\infty^{-2} + \frac{2}{3}Mt}\right).$$

We rescale $Y_{kh} = \varsigma\|U\|_\infty\|V\|_\infty X_{kh}$ and replace t with $t/(\varsigma\|U\|_\infty\|V\|_\infty)$, the above becomes

$$\mathbf{P}\left(\left|\sum_{k,h} Y_{kh}\right| \ge t\right) \le \exp\left(\frac{-t^2}{\varsigma^2 + \frac{2}{3}M\|U\|_\infty\|V\|_\infty t}\right).$$

Let $N = m + n$ and $t = 2\varsigma(\sqrt{\log N} + M\|U\|_\infty\|V\|_\infty \log N)$, we have

$$t^2 \ge 4\varsigma^2 \log N, \qquad t^2 \ge 2M\|U\|_\infty\|V\|_\infty t \log N,$$

thus

$$t^2 \ge \frac{12}{7}\left(\varsigma^2 + \frac{2}{3}M\|U\|_\infty\|V\|_\infty t\right)\log N.$$

Combining everything above, we get

$$\mathbf{P}\left(|u_i^T E v_j| \ge 2\varsigma(\sqrt{\log N} + M\|U\|_\infty\|V\|_\infty \log N)\right) \le N^{-12/7}.$$

By a union bound over $(i, j) \in [r] \times [r]$, the proof of Eq. (77) and the lemma is complete. $\square$

The last bound is technically heavy, so we will dedicate a separate lemma for it, whose proof is postponed to Section B.2.

Lemma 5.2. *Let M and ς be positive real numbers and E be a $m \times n$ random matrix with independent entries following Model (40) with parameters M and ς. 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*

$$D_{U,V,p} := \frac{Mp^3\|U\|_{2,\infty}}{\sqrt{r(m+n)}} + \frac{p^{3/2}}{\sqrt{m+n}} + \frac{p\|V\|_{2,\infty}}{\sqrt{r}}. \tag{78}$$

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

$$\max_{0 \le \alpha \le t\log(m+n)} \frac{\|e_{n,k}^T (E^T E)^a E^T U\|}{(1.9\varsigma\sqrt{m+n})^{2a+1}} \vee \frac{\|e_{n,k}^T (E^T E)^a V\|}{(1.9\varsigma\sqrt{m+n})^{2a}} \le C\sqrt{r} D_{U,V,\log\log(m+n)}. \tag{79}$$

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

$$\max_{0 \le \alpha \le t\log(m+n)} \frac{\|e_{n,k}^T (E^T E)^a E^T U\|_{2,\infty}}{(1.9\varsigma\sqrt{m+n})^{2a+1}} \vee \frac{\|e_{n,k}^T (E^T E)^a V\|_{2,\infty}}{(1.9\varsigma\sqrt{m+n})^{2a}} \le C\sqrt{r} D_{U,V,\log(m+n)}. \tag{80}$$

We only use Eq. (80) to prove Theorem 3.3, but for completeness, we still include (79), which has a better bound at the cost of being non-uniform. This may have potential for other applications.

Let us prove the main theorem using these lemmas.

Proof of Theorem 3.3. Consider the objects from Setting 3.1. We aim to apply Theorem 3.2.

Firstly, we want to choose R_1, R_2 and R_3 to satisfy (35). Let R_1 and R_2 be given from hypothesis, and $R_3 := R_2^2$. Lemma 5.1 then guarantees (35) with probability $1 - O((m+n)^{-1})$.

Secondly, we want to choose τ_1, τ_2 and $\mathcal{H}$ to satisfy (34). There are terms of the same name in Theorem 3.3, so we denote them by τ_1' and τ_2' respectively. Note that $\tau_1' = \tau_{U,V,\log(m+n)}$ from Lemma 5.2. Let C be the constant from that lemma, then $\tau_1 := C\tau_1'$ and $\mathcal{H} := 2\varsigma\sqrt{m+n}$ satisfy the first half of Eq. (34) with probability $1 - O((m+n)^{-2})$. By symmetry, $\tau_2 := C\tau_2'$ and the same $\mathcal{H}$ satisfy the second half, also with probability $1 - O((m+n)^{-2})$.

We can now apply Theorem 3.2. Plugging each of the choices above into their corresponding places in the right-hand sides of Eqs. (36), (37) and (38), we get the desired bounds with probability $1 - O((m+n)^{-1})$. □

A Proof of the full matrix completion theorem

In this section, we prove Theorem 1.5, with the sampling density condition (21) replacing (18).

Proof of the full Theorem 1.5. Let $C_2 = 1/c$ for the constant c in Theorem 3.3. 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:

$$p \geq \frac{Cr^4 r_{\max}\mu_0^2 K_{A,Z}^2}{\varepsilon^2}\left(\frac{1}{m} + \frac{1}{n}\right), \tag{81}$$

$$p \geq C\left(\frac{1}{m} + \frac{1}{n}\right)\log^{10} N, \tag{82}$$

$$p \geq \frac{Cr^3 K_{A,Z}^2}{\varepsilon^2}\left(1 + \frac{\mu_0^2}{\log^2 N}\right)\left(1 + \frac{r^3\log N}{N}\right)\left(\frac{1}{m} + \frac{1}{n}\right)\log^6 N. \tag{83}$$

Let $\rho := \hat{p}/p$. From the sampling density assumption, a standard application of concentration bounds [22, 13] guarantees that, with probability $1 - O(N^{-2})$.

$$0.9 \leq 1 - \frac{1}{\sqrt{N}} \leq 1 - \frac{\log N}{\sqrt{pmn}} \leq \rho \leq 1 + \frac{\log N}{\sqrt{pmn}} \leq 1 + \frac{1}{\sqrt{N}} \leq 1.1. \tag{84}$$

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

$$\|E\| \leq 2\kappa\sqrt{N}, \tag{85}$$

with probability $1 - O(N^{-1})$. See [3, 35], [33, Lemma A.7] or [3] for detailed proofs. Therefore we can assume both Eqs. (84) and (85) at the cost of an $O(N^{-1})$ exceptional probability.

Let $C_0 := 40$. The index s chosen in the SVD step of **AR2** is the largest such that

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

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

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

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

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

We want to show that the first three steps of **AR2** recover A up to an absolute error 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 $\|A_s - A\|_\infty \leq \varepsilon/4$ when C is large enough. To this end, we establish:

$$\sigma_{s+1} \leq r\delta_{s+1} \leq r(\tilde{\delta}_{s+1} + 2\|E\|) \leq r(C_0\rho^{-1/2}\sqrt{r_{\max}} + 4)\kappa\sqrt{N} \leq 2rC_0K_{A,Z}\sqrt{r_{\max}N/p}. \quad (86)$$

For each fixed indices j, k, we have

$$|(A_s - A)_{jk}| = \left|U_{j,\cdot}^T\Sigma_{[s+1,r]}V_{k,\cdot}\right| \leq \sigma_{s+1}\|U\|_{2,\infty}\|V\|_{2,\infty} \leq 2rC_0K_{A,Z}\sqrt{\frac{r_{\max}N}{p}}\frac{r\mu_0}{\sqrt{mn}}$$
$$= \sqrt{\frac{4C_0^2r^4r_{\max}\mu_0^2K_{A,Z}^2}{p}\left(\frac{1}{m} + \frac{1}{n}\right)} \leq \varepsilon/4.$$

where the last inequality comes from the assumption (81) 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 $\|\tilde{A}_s - A_s\|_\infty \leq \varepsilon/4$ with probability $1 - O(N^{-1})$. We aim to use Theorem 3.3, 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:

$$\sigma_s \geq \delta_s \geq \tilde{\delta}_s - 2\|E\| \geq C_0\rho^{-1/2}\kappa\sqrt{r_{\max}N} - 2\|E\| \geq .9C_0\kappa\sqrt{r_{\max}N}. \quad (87)$$

Consider the condition (41). If it holds, then we can apply Theorem 3.3. We want

$$\frac{\kappa\sqrt{N}}{\sigma_s} \vee \frac{r\kappa(\sqrt{\log N} + K\|U\|_\infty\|V\|_\infty\log N)}{\delta_s} \vee \frac{\kappa\sqrt{rN}}{\sqrt{\delta_s\sigma_s}} \leq \frac{1}{16}$$

By Eq. (87), we can replace all three denominators above with $.9C_0\kappa\sqrt{r_{\max}N}$. Additionally, $\|U\|_\infty \leq \|U\|_{2,\infty} \leq \sqrt{\frac{r\mu_0}{m}}$ and $\|V\|_\infty \leq \|V\|_{2,\infty} \leq \sqrt{\frac{r\mu_0}{n}}$, so we can replace them with these upper bounds. We also replace K with $p^{-1/2}$ (its definition). We want

$$\frac{\kappa\sqrt{N} \;\vee\; \kappa\sqrt{rN} \;\vee\; r\kappa(\sqrt{\log N} + \frac{r\mu_0}{\sqrt{pmn}}\log N)}{.9C_0\kappa\sqrt{r_{\max}N}} \leq \frac{1}{16},$$

which is equivalent to

$$\frac{1 \;\vee\; \sqrt{r} \;\vee\; r(\sqrt{\frac{\log N}{N}} + \frac{r\mu_0}{\sqrt{pmnN}}\log N)}{.9C_0\sqrt{r_{\max}}} \leq \frac{1}{16}$$

which easily holds. Therefore we can apply Theorem 3.3. We get, for a constant C_1,

$$\|\tilde{A}_s - A_s\|_\infty \leq C_1 \tau_{UV}\tau_{VU} \cdot r\sigma_s R_s + \frac{1}{N}.$$

Let us simplify the first term in the product, $\tau_{UV}\tau_{VU}$.

$$\begin{aligned}
\tau_{UV} &= \frac{K\|U\|_{2,\infty}\log^3 N}{\sqrt{rN}} + \frac{\log^{3/2} N}{\sqrt{N}} + \frac{\|V\|_{2,\infty}\log N}{\sqrt{r}} \\
&\leq \frac{\sqrt{\mu_0}\log^3 N}{\sqrt{pmN}} + \frac{\log^{3/2} N}{\sqrt{N}} + \frac{\sqrt{\mu_0}\log N}{\sqrt{n}} \leq \frac{\log^{3/2} N}{\sqrt{N}} + \frac{\sqrt{2\mu_0}\log N}{\sqrt{n}},
\end{aligned}$$

where the first inequality comes from (82) if C is large enough. Similarly,

$$\tau_{VU} \leq N^{-1/2}\log^{3/2} N + m^{-1/2}\sqrt{2\mu_0}\log N.$$

Therefore,

$$\begin{aligned}
\tau_{UV}\tau_{VU} &\leq \frac{\log^3 N}{N} + \frac{\sqrt{\mu_0}\log^{5/2} N}{\sqrt{N}} \cdot \frac{\sqrt{2m}+\sqrt{2n}}{\sqrt{mn}} + \frac{2\mu_0\log^2 N}{\sqrt{mn}} \\
&\leq \log^2 N \frac{\log N + 4\sqrt{\mu_0}\sqrt{\log N} + 4\mu_0}{2\sqrt{mn}} \leq \log^2 N \frac{\log N + 4\mu_0}{\sqrt{mn}}.
\end{aligned}$$

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

$$\begin{aligned}
r\sigma_s R_s &\leq r\sigma_s\left(\frac{\kappa\sqrt{N}}{\sigma_s} + \frac{r\kappa(\sqrt{\log N} + \frac{r\mu_0}{\sqrt{mn}}K\log N}{\delta_s} + \frac{r\kappa^2 N}{\delta_s\sigma_s}\right) \\
&= r\left(\kappa\sqrt{N} + \frac{r\kappa\sigma_s}{\delta_s}\left(\sqrt{\log N} + \frac{r\mu_0\log N}{\sqrt{pmn}}\right) + \frac{r\kappa^2 N}{\delta_s}\right) \\
&\leq r\left(\kappa\sqrt{N} + r^2\kappa\left(\sqrt{\log N} + \frac{r\mu_0\log N}{\sqrt{pmn}}\right) + \frac{r\kappa^2 N}{.9C_0\kappa\sqrt{rN}}\right) \\
&\leq r^{3/2}\kappa\left(\sqrt{2N} + r^{3/2}\left(\sqrt{\log N} + \frac{r\mu_0\log N}{\sqrt{pmn}}\right)\right).
\end{aligned}$$

Under the condition (83), we have

$$pmn \geq Cr^3\mu_0^2\log^4 N \implies \frac{r\mu_0\log N}{\sqrt{pmn}} < .1\sqrt{\log N},$$

so the above is simply upper bounded by

$$\frac{\sqrt{2}r^{3/2}K_{A,Z}}{\sqrt{p}}\left(\sqrt{N} + r^{3/2}\sqrt{\log N}\right).$$

Multiplying the two terms, we have by Theorem 3.3,

$$\begin{aligned}
\|\tilde{A}_s - A_s\|_\infty &\leq \log^2 N \cdot \frac{\log N + 4\mu_0}{\sqrt{mn}} \cdot \frac{\sqrt{2}r^{3/2}K_{A,Z}}{\sqrt{p}}\left(\sqrt{N} + r^{3/2}\sqrt{\log N}\right) \\
&\leq \sqrt{\frac{2r^3K_{A,Z}^2\log^6 N}{p}\left(1 + \frac{4\mu_0^2}{\log^2 N}\right)\left(1 + \frac{r^3\log N}{N}\right)\left(\frac{1}{m} + \frac{1}{n}\right)} \leq \varepsilon/4.
\end{aligned} \tag{88}$$

where the last inequality comes from the condition (83) 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 (84) and the triangle inequality,

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

This is the desired bound. The total exceptional probability is $O(N^{-1})$. The proof is complete. □

B Proofs of technical lemmas

B.1 Proof of bound for contour integrals of polynomial reciprocals

In this section, we prove Lemma 4.2, which provides the necessary bounds on the integral coefficients to advance the second step of the main proof (Section 4.2). Recall that the integrals we are interested in have the form

$$\mathcal{C}_\nu(\mathbf{I}) := \oint_{\Gamma_S} \frac{z^\nu \mathrm{d}z}{2\pi i} \frac{1}{z^{\gamma+1}} \prod_{k=1}^{\beta} \frac{\lambda_{i_k}}{z - \lambda_{i_k}}, \qquad \text{where } \nu \in \{0,1\} \text{ and } \beta \le \gamma + 1. \tag{89}$$

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

$$\mathcal{C}_\nu(\mathbf{I}) = \prod_{i=1}^{l} a_i^{m_i} \prod_{j=1}^{k} b_j^{n_j} C(n_0; A, \mathbf{m}; B, \mathbf{n}), \tag{90}$$

where

$$C(n_0; A, \mathbf{m}; B, \mathbf{n}) := \oint_{\Gamma_A} \frac{\mathrm{d}z}{2\pi i} \frac{1}{z^{n_0}} \prod_{j=1}^{k} \frac{1}{(z - b_j)^{n_j}} \prod_{i=1}^{l} \frac{1}{(z - a_i)^{m_i}}, \tag{91}$$

where $n_0 = \gamma + 1 - \nu$. The m_i's and n_j's satisfy $\sum_i m_i + \sum_j n_j \le \gamma + 1$. 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 B.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 $m + n + n_0 \ge 2$, where $m = \sum_i m_i$ and $n := \sum_{i\ge 1} n_i$. 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:*

$$d \le a, \qquad a \le \min_i |a_i|, \qquad d \le \min_{i,j} |a_i - b_j|. \tag{92}$$

Suppose that $0 \le m_i' \le m_i$ for each $i \in [l]$ and that $m' := \sum_{i=1}^{k} m_i' \le n_0$. Then for $C(n_0; A, \mathbf{m}; B, \mathbf{n})$ defined Eq. (91), we have

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \le \binom{m + n + n_0 - 2}{m - 1} \frac{1}{a^{n_0 - m'} d^{m+n-1}} \prod_{i=1}^{l} \frac{1}{|a_i|^{m_i'}} \tag{93}$$

Proof. Firstly, given the sets A and B and the notations and conditions in Lemma B.1, the weak bound below holds

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m+n+n_0-2}{m-1} \frac{1}{d^{m+n+n_0-1}}. \tag{94}$$

We omit the details of the proof, which is a simple induction argument. We now use Eq. (94) to prove the following:

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m+n+n_0-2}{m-1} \frac{1}{a^{n_0} d^{m+n-1}}. \tag{95}$$

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 B.1, such that $m + n + n_0 = N$, Eq. (95) 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. (95) automatically holds by being the same as Eq. (94). 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

$$C(n_0; A, \mathbf{m}; B, \mathbf{n}) = \frac{1}{a_1} \Big[C(n_0 - 1; A, \mathbf{m}; B, \mathbf{n}) - C(n_0; A, \mathbf{m}^{(1)}; B, \mathbf{n}) \Big] \tag{96}$$

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

$$|C(n_0 - 1; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m+n+n_0-3}{m-1} \frac{1}{a^{n_0-1} d^{m+n-1}}. \tag{97}$$

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

$$\left| C(n_0; A, \mathbf{m}^{(1)}; B, \mathbf{n}) \right| \leq \binom{m+n+n_0-3}{m-2} \frac{1}{a^{n_0} d^{m+n-2}} \leq \binom{m+n+n_0-3}{m-2} \frac{1}{a^{n_0-1} d^{m+n-1}}. \tag{98}$$

Notice that the binomial coefficients in Eqs. (97) and (98) sum to the binomial coefficient in Eq. (95), we get $P_1(N)$, which proves Eq. (95) by induction.

Now we can prove Eq. (93). The logic is almost identical, with Eq. (95) playing the role of Eq. (94) 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 B.1, such that $m + n + n_0 = N$, Eq. (93) 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. (95). By symmetry among the indices, assume $m'_1 \geq 1$. This also means $n_0 \geq 1$. Consider Eq. (96) 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

$$|C(n_0 - 1; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{m+n+n_0-3}{m-1} \frac{1}{a^{n_0-m'} d^{m+n-1}} \frac{1}{|a_1|^{m'_1-1}} \prod_{i=2}^{l} \frac{1}{|a_i|^{m'_i}}. \tag{99}$$

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. (96):

$$\begin{aligned}\left|C(n_0; A, \mathbf{m}^{(1)}; B, \mathbf{n})\right| &\leq \binom{m+n+n_0-3}{m-2} \frac{1}{a^{n_0-m'+1} d^{m+n-2}} \frac{1}{|a_1|^{m'_1-1}} \prod_{i=2}^{l} \frac{1}{|a_i|^{m'_i}} \\ &\leq \binom{m+n+n_0-3}{m-2} \frac{1}{a^{n_0-m'} d^{m+n-1}} \frac{1}{|a_1|^{m'_1-1}} \prod_{i=2}^{l} \frac{1}{|a_i|^{m'_i}}.\end{aligned}$$

Summing up the bounds by summing the binomial coefficients, we get exactly $P_2(N)$, so Eq. (93) is proven by induction. □

Lemma B.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 B.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*

$$m' + n' \leq n_0 \quad \text{for} \quad m' := \sum_i m'_i, \quad n' := \sum_j n'_j.$$

Then for $C(n_0; A, \mathbf{m}; B, \mathbf{n})$ defined in Eq. (91), we have

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \leq \binom{n+n_0-n'+m-2}{m-1} \frac{(1+d/a)^{n'}}{a^{n_0-m'-n'} d^{m+n-1}} \prod_{i=1}^{l} \frac{1}{|a_i|^{m'_i}} \prod_{j=1}^{k} \frac{1}{|b_j|^{n'_j}}. \tag{100}$$

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 B.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 4.2.

Proof of Eq. (58) *of Lemma 4.2.* First rewrite the integral into the forms of (89), then (90) and (91). 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 $m' + n' = \beta \le \gamma + 1 = n_0$, so we can apply Lemma B.2 to get

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

or equivalently,

$$|\mathcal{C}_0(\mathbf{I})| \le \left(1 + \frac{\Delta_S(\mathbf{I})}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})} \binom{\gamma + \beta_S(\mathbf{I}) - 1}{\beta_S(\mathbf{I}) - 1} \frac{1}{\lambda_S(\mathbf{I})^{\gamma+1-\beta} \Delta_S(\mathbf{I})^{\beta-1}}.$$

Note that since $\beta_S(\mathbf{I}) \le \beta \le \gamma + 1$ and $\mathrm{L}_0 = 2$,

$$\binom{\gamma + \beta_S(\mathbf{I}) - 1}{\beta_S(\mathbf{I}) - 1} = \frac{\gamma + \beta_S(\mathbf{I}) - 1}{\gamma} \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \le \mathrm{L}_0 \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1},$$

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 $m_1 \ge 1$. Applying Lemma B.2 for the same parameters as in the previous case, except that $m'_1 = m_1 - 1$, we get

$$|C(n_0; A, \mathbf{m}; B, \mathbf{n})| \le |a_1| \binom{n_0 + m - 2}{m - 1} \frac{(1 + d/a)^{n'}}{a^{n_0 - m + 1 - n} d^{m+n-1}} \prod_{i=1}^{l} \frac{1}{|a_i|^{m_i}} \prod_{j=1}^{k} \frac{1}{|b_j|^{n_j}},$$

which translates to

$$|\mathcal{C}_1(\mathbf{I})| \le \lambda_S(\mathbf{I}) \binom{\gamma + \beta_S(\mathbf{I}) - 2}{\beta_S(\mathbf{I}) - 1} \left(1 + \frac{\Delta_S(\mathbf{I})}{\lambda_S(\mathbf{I})}\right)^{\beta_{S^c}(\mathbf{I})} \frac{1}{\lambda_S(\mathbf{I})^{\gamma+1-\beta} \Delta_S(\mathbf{I})^{\beta-1}},$$

which is the desired bound since $\mathrm{L}_1 = \lambda_S(\mathbf{I})$.

Let us now prove Eq. (59). We can assume $\beta_S(\mathbf{I}) \ge 1$, 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. (58) 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 $\lambda_S \le \lambda_S(\mathbf{I})$ and $\Delta_S \le \Delta_S(\mathbf{I})$.

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. (58) and (59) hold. The proof is complete. □

B.2 Proof of semi-isotropic bounds for powers of random matrices

In this section, we prove Lemma 5.2, 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 B.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 (40) with parameters M and $\varsigma = 1$, meaning E has independent entries and*

$$\mathbf{E}\left[E_{ij}\right] = 0, \quad \mathbf{E}\left[\|E\|_{ij}^2\right] \leq 1, \quad \mathbf{E}\left[\|E\|_{ij}^p\right] \leq M^{p-2} \quad \text{for all } \; p.$$

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

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

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

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

Lemma B.4. *Let E be a $m \times n$ random matrix following the model in Lemma B.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*

$$m + n \geq 2^8 M^2 p^6 (2a)^4,$$

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

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

Let us prove the main objective of this section, Lemma 5.2, before delving into the proof of the technical lemmas.

Proof of Lemma 5.2. Consider Eq. (79) and assume $M \le \log^{-2-\varepsilon}(m+n)\sqrt{m+n}$. Fix $k \in [n]$. It suffices to prove the following two bounds uniformly over all $a \in [\lfloor t \log(m+n) \rfloor]$:

$$\left\| e_{n,k}^T (E^T E)^a E^T U \right\| \le C\sqrt{r} D_{U,V,\log\log(m+n)} (1.9\varsigma\sqrt{m+n})^{2a+1} \tag{101}$$

$$\left\| e_{n,k}^T (E^T E)^a V \right\| \le C\sqrt{r} D_{U,V,\log\log(m+n)} (1.9\varsigma\sqrt{m+n})^{2a}. \tag{102}$$

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

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

so we can apply both Lemmas B.3 and B.4.

Let us prove Eq. (101) for a. Applying Lemma B.3 for the random matrix E/ς and $D = 2^{13}$ gives, with probability $1 - \log^{-4.04}(m+n)$,

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

where the second inequality is due to $\alpha \le (\sqrt{2}/1.9)^\alpha$. 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. (101) follows.

Let us prove Eq. (102). Applying Lemma B.3 for the random matrix E/ς and $D = 2^{10}$ gives, with probability $1 - \log^{-8}(m+n)$,

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

proving Eq. (102) after a union bound, similar to the previous case. Combining the two cases, Eq. (79) is proven.

Let us now prove Eq. (80). Since the 2-to-∞ norm is the the largest norm among the rows, it suffices to prove Eq. (79) 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. (79), but with probability at least $1 - (m+n)^{-4.04}$. Applying another union bound over $k \le [n]$ gives Eq. (80) with probability at least $1 - (m+n)^{-3}$. The proof is complete. $\square$

Now let us handle the technical lemmas B.3 and B.4. The odd case (Lemma B.3) is more difficult, so we will handle it first to demonstrate our technique. The argument for the even case (Lemma B.4) is just a simpler version of the same technique.

B.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

$$\mathbf{E}\left[\left\| e_{n,1}^T (E^T E)^a E^T U \right\|^{2p} \right] = \mathbf{E}\left[\left(\sum_{l=1}^{r} (e_{n,1}^T (E^T E)^a E^T u_l)^2 \right)^p \right] = \sum_{l_1,\dots,l_p \in [r]} \mathbf{E}\left[\prod_{h=1}^{p} (e_{n,1}^T (E^T E)^a E^T u_{l_h})^2 \right]. \tag{103}$$

Temporarily let $\mathcal{W}$ be the set of walks $W = (j_0 i_0 j_1 i_1 \dots i_a)$ of length $2a+1$ on the complete bipartite graph $M_{m,n}$ such that $j_0 = 1$. Here the two parts of M are $I = \{1', 2', \dots, m'\}$ and $J = \{1, 2, \dots, n\}$, where the prime symbol serves to distinguish two vertices on different parts with the same number. Let $E_W = E_{i_0 j_0} E_{i_0 j_1} \dots E_{i_{a-1} j_a} E_{i_a j_a}$. We can rewrite the final expression in the above as

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

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

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

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

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

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 $M_{m,n}$, 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:

$$\sum_{P \in \mathcal{P}} |\mathbf{E}\left[E_P\right]| \prod_{h=1}^{p} \|U_{\cdot, i_{(2h-1)a}}\| \|U_{\cdot, i_{(2h)a}}\| = r^p \sum_{P \in \mathcal{P}} u_{P^{\text{end}}} |\mathbf{E}\left[E_P\right]| = r^p \sum_{S \in \mathcal{S}} \sum_{P \in \mathcal{P}(S)} u_{P^{\text{end}}} |\mathbf{E}\left[E_P\right]|, \quad (104)$$

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

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

Fix $P \in \mathcal{P}$. Let us bound $\mathbf{E}[E_P]$. 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

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

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

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

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 $l = 1$, $\|u\|_l^l \le \sqrt{m}$ by Cauchy-Schwarz. When $l \ge 2$, we have $\|u\|_l^l \le \|u\|_\infty^{l-2}\|u\|_2^2 = \|u\|_\infty^{l-2}$. Thus

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

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 (104):

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

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 $M_1 := M^{2p(2a+1)}\|u\|_\infty^{2p}$. Then we can rewrite the above as:

$$M_1 \sum_{x,y,z,t\in\mathcal{A}} M^{-2(z+t)} m^{z-x/2-y} n^{t-1} \|u\|_\infty^{-x-2y} |\mathcal{S}(x,y,z,t)|, \tag{105}$$

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

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

where the last bound is due to $z + t = |V(S)| \le |E(S)| + 1 \le p(2a+1) + 1$, 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 B.5. *When each edge is repeated at least twice, we have $z - x/2 - y + t - 1 \le 2ap$.*

Proof of Claim B.5. Let $S = (S_1, \dots, S_{2p})$, where $S_r = j_{r0}i_{r0}j_{r1}i_{r1}\dots j_{ra}i_{ra}$. 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

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

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 $k \ge 0$, $S^{(k)}$ 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

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

proving the claim. If there is some vertex in $A_1(S^{(k)})$, assume it is some i_{rs}, then we must have $s \le a-1$ 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 $|A_1(S^{(k+1)})| \le |A_1(S^{(k)})|$, 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 B.5 shows that we can define the set $\mathcal{A}$ of *eligible sizes* as follows:

$$\mathcal{A} = \left\{(x,y,z,t) \in \mathbb{N}^4_{\ge 0} : \quad 1 \le t; \quad 1 \le x + y \le z; \quad x + 2y \le 2p; \quad z - x/2 - y + t - 1 \le 2ap\right\}. \tag{106}$$

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

Claim B.6. *Given a tuple $(x,y,z,t) \in \mathcal{A}$, where $\mathcal{A}$ is defined in Eq. (106), we have*

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

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 $r = 1, 2, \dots, 2p$:
   (a) Relabel S_r as $1k_1k_2\dots k_{2a}$.
   (b) For $s = 1, 2, \dots, 2a$:
      - 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 [35] 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 B.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*

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

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

$$|\mathcal{S}(x,y,z,t)| \le \binom{2p(a+1)}{z}\binom{2pa}{t-1}\sum_{u+v=z+t-1}\binom{2p(a+1)-z}{u}\binom{2pa-t+1}{v}z^{f(z,u)}t^{g(t,v)}, \quad (107)$$

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

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

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

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

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

Back to the proof of Claim B.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 [35] 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 B.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 B.6 is proven. □

Back to the proof of Lemma B.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 B.6. Also let

$$M_2 = M_1(2p(a+1))!(2pa)! = M^{2p(2a+1)}(2p(a+1))!(2pa)!\|u\|_\infty^{2p}.$$

Replacing the appropriate terms in the bound in Claim B.6 with these short forms, we get another series of upper bounds for the last double sum in Eq. (104):

$$\begin{aligned}
&M_2 \sum_{x,y} \|u\|_\infty^{-x-2y} \sum_{l=x+y}^{\lfloor 2pa+x/2+y \rfloor} M^{-2(l+1)} F_l \sum_{z+t=l+1} \frac{m^{z-x/2-y} n^{t-1}}{z!(t-1)!} \\
&\leq M_2 \sum_{x,y} \|u\|_\infty^{-x-2y} \sum_{l=x+y}^{\lfloor 2pa+x/2+y \rfloor} \frac{M^{-2(l+1)} F_l}{(l - \lfloor \frac{x}{2} \rfloor - y)!} \sum_{z+t=l+1} \binom{l - \lfloor \frac{x}{2} \rfloor - y}{z - \lfloor \frac{x}{2} \rfloor - y} m^{z - \lfloor \frac{x}{2} \rfloor - y} n^{t-1} \\
&\leq M_2 \sum_{x,y} \|u\|_\infty^{-x-2y} \sum_{l=x+y}^{\lfloor 2pa+x/2+y \rfloor} \frac{M^{-2(l+1)} F_l}{(l - \lfloor \frac{x}{2} \rfloor - y)!} (m+n)^{l - \lfloor \frac{x}{2} \rfloor - y}.
\end{aligned}$$

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

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

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.

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

Therefore, under the assumption that $m+n \geq 256 M^2 p^6 (2a+1)^4$, 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

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

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

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

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. (104):

$$\begin{aligned}
&M_2 (m+n)^{2pa} \sum_{d=0}^{p-1} 4(p-d) \|u\|_\infty^{-2(p-d)} \cdot \frac{(2M^{-2})^{N-d+1}(N-d+1)^{2d}(8p+8d-2)^{2p+2d-1}}{(2pa)! \cdot (2d)! \cdot (N-d)!} \\
&\leq M_3 (m+n)^{2pa} \sum_{d=0}^{p-1} \|u\|_\infty^{2d} \cdot \frac{2^{-d} M^{2d} (N-d+1)^{2d} (8p+8d-2)^{2p+2d-1}}{(2d)! \cdot (N-d)!},
\end{aligned} \tag{108}$$

where

$$M_3 = 4p\frac{M_2\|u\|_\infty^{-2p}(2M^{-2})^{N+1}}{(2pa)!} = 2^{p(2a+1)+3}pM^{-2}(2p(a+1))!.$$

Let us bound the sum at the end of Eq. (108). Temporarily let A_d be the term corresponding to d and $x := 2^{-1/2}M\|u\|_\infty$. 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

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

Plugging this into Eq. (108), we get another upper bound for (104):

$$M_4(16N^{3/2}M\|u\|_\infty+1)^{2p}(m+n)^{2ap},$$

where

$$M_4 := M_3\frac{(16p)^{2p}}{8pN!} = 2^{p(2a+1)+3}pM^{-2}(2p(a+1))!\frac{(16p)^{2p}}{8p(2ap+p)!} \leq \frac{2^{2ap}2^{10p}p^{3p}(a+1)^p}{8M^2}.$$

To sum up, we have

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

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

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

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

B.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 B.3 to get the bound

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

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

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

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 $P \in \mathcal{P}(S)$,

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

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

$$\sum_{S\in\mathcal{S}} \sum_{P\in\mathcal{P}(S)} v_{P^{\text{end}}} |\mathbf{E}\left[E_P\right]| \le M_1 \sum_{S\in\mathcal{S}} M^{-2|V(S)|} m^{|V_I(S)|} n^{|V_J(S)|-1}, \quad \text{where } M_1 := M^{4pa+2}\|v\|_\infty^{2p}.$$

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 B.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

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

We have

$$\begin{aligned}
\sum_{S\in\mathcal{S}} \sum_{P\in\mathcal{P}(S)} v_{P^{\text{end}}} |\mathbf{E}\left[E_P\right]| &\le M_1 \sum_{l=0}^{2ap} M^{-2(l+1)}[(2ap)!]^2 F_l \sum_{z+t=l+1} \frac{m^z n^{t-1}}{z! \cdot (t-1)!} \\
&= M_2 \sum_{l=0}^{2ap} \frac{M^{-2l} F_l}{l!} \sum_{z+t=l+1} \binom{l}{z} m^z n^{t-1} = M_2 \sum_{l=0}^{2ap} \frac{M^{-2l} F_l}{l!} (m+n)^l,
\end{aligned}$$

where $M_2 := M_1[(2pa)!]^2 M^{-2} = M^{4ap}[(2pa)!]^2\|v\|_\infty^{2p}$. Let C_l be the term corresponding to l in the last sum above. An anlogous calculation from the proof of Lemma B.3 shows that under the assumption that $m + n \ge 256M^2p^6(2a)^4$, $C_l \ge 2C_{l-1}$ for each l, so $\sum_{l=0}^{2pa} C_l \le 2C_{2pa}$, where

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

Therefore

$$\begin{aligned}
\mathbf{E}\left[\left\|e_{n,1}^T (E^T E)^a V\right\|^{2p}\right] &\le r^p \sum_{S\in\mathcal{S}} \sum_{P\in\mathcal{P}(S)} v_{P^{\text{end}}} |\mathbf{E}\left[E_P\right]| \\
&\le 2r^p M_2 \frac{M^{-4ap}2^{2ap+1}(8p-2)^{2p-1}}{[(2ap)!]^2}(m+n)^{2ap} = 4\left(2^3 p r^{1/2}\|v\|_\infty [2(m+n)]^a\right)^{2p}.
\end{aligned}$$

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] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48 3:1452–1474, 2017.

[2] Jinho Baik, Gerard Ben Arous, and Sandrine Peche. Phase transition of the largest eigenvalue for non-null complex covariance matrices. *Annals of Probability*, 33, 09 2005.

[3] 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.

[4] Robert M. Bell and Yehuda Koren. Lessons from the netflix prize challenge. *SIGKDD Explor. Newsl.*, 9(2):75–79, December 2007.

[5] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

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

[7] 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.

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

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

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

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

[12] Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. The largest eigenvalues of finite rank deformation of large wigner matrices: Convergence and nonuniversality of the fluctuations. *The Annals of Probability*, 37, 06 2007.

[13] 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.

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

[15] 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.

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

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

[18] Delphine Féral and Sandrine Peche. The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in Mathematical Physics*, 272, 06 2006.

[19] Farzan Haddadi and Arash Amini. Eigenvectors of deformed wigner random matrices. *IEEE Transactions on Information Theory*, 67(2):1069–1079, 2021.

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

[21] 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.

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

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

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

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

[26] Raghunandan Keshavan and Sewoong Oh. A gradient descent algorithm on the grassman manifold for matrix completion. 910, 10 2009.

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

[28] 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.

[29] Sandrine Peche. Péché s.: The largest eigenvalue of small rank perturbations of hermitian random matrices. probab. theory relat. fields 134, 127-173. *Probability Theory and Related Fields*, 134:127–173, 01 2006.

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

[31] 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.

[32] 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.

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

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

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

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