Near-optimal fitting of ellipsoids to random points

Aaron Potechin POTECHIN@UCHICAGO.EDU
University of Chicago

Paxton Turner PAXTONTURNER@G.HARVARD.EDU
Harvard University

Prayaag Venkat PVENKAT@G.HARVARD.EDU
Harvard University

Alexander S. Wein ASWEIN@UCDAVIS.EDU
UC Davis

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Given independent standard Gaussian points $v_1, \dots, v_n$ in dimension d, for what values of (n, d) does there exist with high probability an origin-symmetric ellipsoid that simultaneously passes through all of the points? This basic problem of fitting an ellipsoid to random points has connections to low-rank matrix decompositions, independent component analysis, and principal component analysis. Based on strong numerical evidence, Saunderson, Parrilo, and Willsky (Saunderson, 2011; Saunderson et al., 2013) conjectured that the ellipsoid fitting problem transitions from feasible to infeasible as the number of points n increases, with a sharp threshold at $n \sim d^2/4$. We resolve this conjecture up to logarithmic factors by constructing a fitting ellipsoid for some $n = d^2/\mathrm{polylog}(d)$. Our proof demonstrates feasibility of the least squares construction of (Saunderson, 2011; Saunderson et al., 2013) using a convenient decomposition of a certain non-standard random matrix and a careful analysis of its Neumann expansion via the theory of graph matrices.

Keywords: High-dimensional probability, semi-definite programming, phase transitions, convex geometry

1. Introduction

Let $v_1, \dots, v_n \in \mathbb{R}^d$ be a collection of points. We say that this collection has the *ellipsoid fitting property* if there exists a symmetric matrix $X \in \mathbb{R}^{d \times d}$ such that $X \succeq 0$ and $v_i^T X v_i = 1$ for all $i \in [n]$. That is, the eigenvectors and eigenvalues of the matrix X describe the directions and reciprocals of the squared-lengths of the principal axes of an origin-symmetric ellipsoid that passes through all of $v_1, \dots, v_n$. From the definition, it is clear that testing whether the ellipsoid fitting property holds for a given set of points reduces to solving a certain semidefinite program. It is known that if $v_1, \dots, v_n$ satisfy the ellipsoid fitting property, then $\pm v_1, \dots, \pm v_n$ lie on the boundary of their convex hull[1] and that the converse holds when $n \leq d+1$ (Corollary 3.6 of Saunderson et al. (2012)).

In this paper, we study whether *random* points satisfy the ellipsoid fitting property. Specifically, let $v_1, \dots, v_n \sim \mathcal{N}(0, I_d)$ be i.i.d. standard Gaussian vectors in $\mathbb{R}^d$. Treating $n = n(d)$ as a function of d, we ask: what is the largest value of n for which n standard Gaussian vectors have the ellipsoid

1. A point v_i lies on the boundary of the convex hull of $\pm v_1, \dots, \pm v_n$ if there exists $x \in \mathbb{R}^d$ such that $\langle x, v_i \rangle = 1$ and $|\langle x, v_j \rangle| \leq 1$ for all $j \neq i$.

fitting property with high probability[2] as $d \to \infty$? Since the probability of the ellipsoid fitting property is non-increasing as a function of n, it is natural to ask if it exhibits a sharp phase transition from 1 to 0 asymptotically as n increases.

If $n \leq d+1$, then with probability 1, the points $\pm v_1, \ldots, \pm v_n$ have the aforementioned convex hull property and hence satisfy the ellipsoid fitting property. However, it turns out that for random points, the ellipsoid fitting property actually holds for much larger values of n. Intriguing experimental results due to Saunderson et al. Saunderson (2011); Saunderson et al. (2012, 2013) suggest that the ellipsoid fitting property undergoes a sharp phase transition at the threshold $n \sim d^2/4$. Formally, we restate their conjecture:

Conjecture 1 *Let $\epsilon > 0$ be a constant and $v_1, \ldots, v_n \sim \mathcal{N}(0, I_d)$ be i.i.d. standard Gaussian vectors in $\mathbb{R}^d$.*

1. *If $n \leq (1-\epsilon)\frac{d^2}{4}$, then $v_1, \ldots, v_n$ have the ellipsoid fitting property with probability $1 - o(1)$.*
2. *If $n \geq (1+\epsilon)\frac{d^2}{4}$, then $v_1, \ldots, v_n$ have the ellipsoid fitting property with probability $o(1)$.*

By genericity of the random linear constraints and the fact that any $d \times d$ PSD matrix (in fact, symmetric matrix) is described by $d(d+1)/2$ parameters, it can be verified that the system of random linear constraints alone (without the PSD constraint) becomes infeasible with probability 1 if and only if $n > d(d+1)/2$ (see Lemma 40). Fascinatingly, Conjecture 1 posits the existence of a range of values $n \in \left(\frac{d^2}{4}, \frac{d(d+1)}{2}\right)$ for which with high probability, there exists a *symmetric* matrix satisfying the linear constraints, but no such *positive semidefinite* matrix exists. Saunderson et al. Saunderson (2011); Saunderson et al. (2013) made partial progress towards resolving the positive part of this conjecture: they showed that for any $\epsilon > 0$, when $n < d^{6/5-\epsilon}$, the ellipsoid fitting property holds with high probability. A special case of Theorem 1.4 of Ghosh, Jeronimo, Jones, Potechin, and Rajendran Ghosh et al. (2020), developed in the context of certifying upper bounds on the Sherrington–Kirkpatrick Hamiltonian, guarantees that for any $\epsilon > 0$, when $n < d^{3/2-\epsilon}$, there exists with high probability a fitting ellipsoid X whose diagonal entries are all equal to $1/d$.

The ellipsoid fitting problem, a basic question in high-dimensional probability and convex geometry, is further motivated by connections to other problems in machine learning and theoretical computer science. First, Conjecture 1 was first formulated by Saunderson et al. Saunderson (2011); Saunderson et al. (2012, 2013) in the context of decomposing an observed $n \times n$ data matrix as the sum of a diagonal matrix and a random rank-r matrix. They proposed a convex-programming heuristic, called "Minimum-Trace Factor Analysis (MTFA)" for solving this problem and showed it succeeds with high probability if the ellipsoid fitting property for n standard Gaussian vectors in $d = n - r$ dimensions holds with high probability.

Second, Podosinnikova et al. Podosinnikova et al. (2019) identified a close connection between the ellipsoid fitting problem and the overcomplete independent component analysis (ICA) problem, in which the goal is to recover a mixing component of the model when the number of latent sources n exceeds the dimension d of the observations. They show that the ability of an SDP-based algorithm to recover a mixing component is related to the feasibility of a variant of the ellipsoid fitting problem in which the norms of the random points fluctuate with higher variance than in our model. They give experimental evidence that the SDP succeeds when $n < d^2/4$, the same phase transition behavior described in Conjecture 1, and show rigorously that it succeeds for some $n = \Omega(d \log d)$.

2. Here and throughout, *high probability* means probability tending to 1 as $d \to \infty$.

Third, the ellipsoid fitting property for random points is directly related to the ability of a canonical SDP relaxation to certify lower bounds on the discrepancy of nearly-square random matrices. The discrepancy of random matrices is a topic of recent interest, with connections to controlled experiments Turner et al. (2020), the Ising Perceptron model from statistical physics Aubin et al. (2019), and the negatively-spiked Wishart model Bandeira et al. (2020); Venkat (2022). A result implicit in the work of Saunderson, Chandrasekaran, Parrilo, and Willsky Saunderson et al. (2012) states that if the ellipsoid fitting property for n Gaussian points in dimension d holds with high probability, then the SDP fails to certify a non-trivial lower bound on the discrepancy of a $(n-d) \times n$ matrix with i.i.d. standard Gaussian entries (see Appendix C for further discussion). In addition, this provides further evidence of the algorithmic phase transition for the detection problem in the negatively-spiked Wishart model that was previously predicted by the low-degree likelihood ratio method Bandeira et al. (2020).

Finally, a current active area of research in theoretical computer science aims to give rigorous evidence for information-computation gaps in average-case problems by characterizing the performance of powerful classes of algorithms, such as the Sum-of-Squares (SoS) SDP hierarchy. Often, the most challenging technical results in this area involve proving lower bounds against these SDP-based algorithms. Moreover, there are relatively few examples for which predicted phase transition behavior has been sharply characterized (see e.g. Barak et al. (2019); Ghosh et al. (2020); Hopkins et al. (2017); Hsieh and Kothari (2022); Jones et al. (2022); Kothari and Manohar (2021); Mohanty et al. (2020); Schoenebeck (2008)), all proven using the same technique of "pseudo-calibration". We remark that proving the positive side of Conjecture 1 amounts to proving the feasibility of an SDP with random linear constraints. This also arises in average-case SoS lower bounds, although the linear constraints for average-case SoS lower bounds are generally very intricate.

The main contribution of our work is to resolve the positive side of Conjecture 1 up to logarithmic factors. (Recall that the negative side of Conjecture 1 has already been resolved up to a factor of 2.)

Theorem 2 *There is a universal constant $C > 0$ so that if $n \leq d^2/\log^C(d)$, then $v_1, \ldots, v_n \sim \mathcal{N}(0, I_d)$ have the ellipsoid fitting property with high probability.*

As a first corollary of Theorem 2, we conclude that MTFA in this setting succeeds provided $r \leq n - \sqrt{n}\,\mathrm{polylog}(n)$, improving on the bound $r \leq n - \omega(n^{2/3})$ from a combination of the results of Saunderson et al. Saunderson (2011); Saunderson et al. (2013) and Ghosh et al. Ghosh et al. (2020). Second, Theorem 2 implies the following "finite-size" phase transition result: a canonical SDP cannot distinguish between an $m \times n$ matrix with i.i.d. standard Gaussian entries and one with a planted Boolean vector in its in kernel when $m \leq n - \sqrt{n}\,\mathrm{polylog}(n)$ (see Appendix C), again improving on the bound $m \leq n - \omega(n^{2/3})$ that follows from Ghosh et al. (2020).

Experimental results It is natural to wonder whether our proof of Theorem 2 can be sharpened to make further progress on Conjecture 1. Our proof is based on a least-squares construction that was first studied in Saunderson (2011); Saunderson et al. (2013) (see Section 2). Although the least-squares construction always satisfies the linear constraints, in Section 3 we corroborate experimental evidence of Saunderson et al. suggesting that it fails to be positive semidefinite strictly below the conjectured $n \sim d^2/4$ threshold. We also introduce a new method called the "identity-perturbation" construction that also always satisfies the linear constraints and appears to improve on

the least-squares construction in experiments, while having similar time complexity. Our simulations in Section 3 provide numerical evidence that the positive semi-definiteness of the least-squares and identity-perturbation constructions undergo sharp phase transitions at roughly $n \approx d^2/17$ and $n \approx d^2/10$, respectively. We did not run eperiments on the pseudo-calibration construction of Ghosh et al. (2020) because this construction has the drawback that it involves logarithmic degree polnomials of the input and is thus very hard to compute.

These results suggest that a full resolution of Conjecture 1 requires either sharply analyzing the pseudocalibration construction of Ghosh et al. (2020) (if it achieves the threshold $n \approx \frac{d^2}{4}$, which is unknown), inventing a new construction and analyzing it, or reasoning indirectly about the ellipsoid fitting property without considering any explicit candidate.

Related work We now discuss two closely related works that study a simpler variant of the ellipsoid fitting problem. In this variant, the constraints $v_i^T X v_i = 1$ in the definition of the ellipsoid fitting property are replaced by $\langle X, G_i \rangle = 1$, where $G_1, \ldots, G_n \in \mathbb{R}^{d \times d}$ have i.i.d. standard Gaussian entries. Amelunxen, Lotz, McCoy, and Tropp Amelunxen et al. (2014) give a general framework for characterizing phase transition behavior of convex programs with random constraints. Interestingly, their framework shows that the conclusion of Conjecture 1 is true for the simpler variant. Moreover, they explain that the occurrence of the phase transition at $n \sim d^2/4$ arises from the fact that $d(d+1)/4$ is the "statistical dimension" of the cone of $d \times d$ PSD matrices. The known proofs of these results are either based on conic geometry or Gaussian process techniques that crucially rely on the fact that the entries of the constraint matrices are i.i.d. and Gaussian. Despite the strikingly similar phase transition behavior for the two models of random constraints, it appears unlikely that these techniques can be used to resolve Conjecture 1. In this simpler i.i.d. setting of Amelunxen et al., Hsieh and Kothari Hsieh and Kothari (2022) show that when $n \leq d^2/\operatorname{polylog}(d)$, the ellipsoid fitting SDP (which corresponds to the degree-2 SoS SDP) equipped with some additional symmetry constraints (corresponding to the degree-4 SoS SDP) is still feasible with high probability.

Ghosh et al. Ghosh et al. (2020) consider the original setting in which the constraint matrices are of the form $v_i v_i^T$ and also impose the constraint that the diagonal entries of X satisfy $X_{ii} = 1/d$ for all $i \in [d]$. They show that for any $\epsilon > 0$ this SDP, even when augmented with more constraints corresponding to higher degree SoS, remains feasible with high probability for some $n = \Omega(d^{3/2-\epsilon})$ and conjecture that this should even hold for some $n = \Omega(d^{2-\epsilon})$. The proofs of the results of Ghosh et al. Ghosh et al. (2020) and Hsieh and Kothari Hsieh and Kothari (2022) are based on the pseudo-calibration technique. Due to technical complications that arise when analyzing higher degree SoS pseudocalibration constructions, Ghosh et al. (2020) can only prove feasibility for some $n = \Omega(d^{3/2-\epsilon})$. However, as we detail in Appendix I, these technical complications do not arise when specialized to the degree-2 case, which gives an alternative proof of Theorem 2.

On a technical level, our proof heavily relies on the recently introduced machinery of *graph matrices* Ahn et al. (2016), a powerful tool for obtaining norm bounds of structured random matrices using a certain graphical calculus (see Section B). Ours is among the first works to apply graph matrices outside of the Sum-of-Squares lower bound literature, and we expect graph matrices to be useful for other probabilistic applications beyond average-case complexity theory.

Independent work Shortly after a revised version of this paper was posted on the arXiv, independent work of Kane and Diakonikolas Kane and Diakonikolas (2022) analyzed the identity-perturbation construction and showed it improves on the logarithmic factor in Theorem 2. Their short proof crucially uses the fact that the norms and directions of a standard Gaussian are inde-

pendent. Our proof is more technically involved but can be adapted to handle non-Gaussian distributions whose coordinates are independent and sufficiently well-concentrated. Our work analyzes the least-squares construction, and its analysis also carries over easily to the analysis of the identity perturbation construction, as we demonstrate in Section J.

2. Technical overview

We now give an overview of the proof of Theorem 2. To begin, we introduce some convenient notation. Define the linear operator $\mathcal{A} : \mathbb{R}^{d\times d} \to \mathbb{R}^n$ by $\mathcal{A}(X) := (v_1^T X v_1, \dots, v_n^T X v_n)^T$ and let $\mathcal{A}^\dagger$ be its pseudoinverse. The fitting ellipsoid in Theorem 2 is obtained via the *least-squares construction*:

$$X_{\text{LS}} = \mathcal{A}^\dagger(1_n), \tag{1}$$

which is the minimum Frobenius norm solution to the linear constraints. This construction was first studied by Saunderson et al. Saunderson (2011); Saunderson et al. (2013). Our analysis builds on their work and also introduces additional probabilistic and linear-algebraic ideas, such as the application of graph matrices, leading to nearly-sharp bounds for the ellipsoid fitting problem.

To prove Theorem 2, it suffices to verify that $\mathcal{A}(X_{\text{LS}}) = 1_n$ and $X_{\text{LS}} \succeq 0$ with high probability, for appropriate values of n and d. The first condition can be easily verified: with probability 1, the $n \times n$ matrix $\mathcal{A}\mathcal{A}^*$ is invertible (Lemma 40), so we may write $\mathcal{A}^\dagger = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}$ and compute that indeed $\mathcal{A}(X_{\text{LS}}) = 1_n$, where the adjoint $\mathcal{A}^* : \mathbb{R}^n \to \mathbb{R}^{d\times d}$ satisfies $\mathcal{A}^*(c) = \sum_{i=1}^n c_i v_i v_i^T$.

The challenging part of the proof is to verify that $X_{\text{LS}} \succeq 0$ with high probability. We now give some intuition for why this condition holds. First, observe that if we take $X_0 = \frac{1}{d} I_d$, then a simple application of a tail bound for the χ^2 distribution and a union bound over the n constraints yields $\|\mathcal{A}(X_0) - 1_n\|_\infty = O(\sqrt{\log(n)/d})$ with high probability. In words, X_0 defines an ellipsoid that approximately fits the points $v_1, \dots, v_n$ and whose eigenvalues are well-separated from 0.

Second, there is a sense in which X_{LS} is (approximately) a projection of X_0 onto the affine subspace $\{X \in \mathbb{R}^{d\times d} : \mathcal{A}(X) = 1_n\}$. Recall that X_{LS} can be expressed as the solution of the following optimization problem:

$$\min_{X\in\mathbb{R}^{d\times d},\, \mathcal{A}(X)=1_n} \|X\|_F^2 .$$

In fact, since the above minimization is over X that satisfy $\mathcal{A}(X) = 1_n$, X_{LS} is also the solution of

$$\min_{X\in\mathbb{R}^{d\times d},\, \mathcal{A}(X)=1_n} \left\| X - \frac{1}{dn} \sum_{i=1}^n v_i v_i^T \right\|_F^2 .$$

For $n \gg d$, it is the case that $\frac{1}{dn}\sum_{i=1}^n v_i v_i^T \approx X_0$ with high probability. Thus, we interpret X_{LS} as an (approximate) projection of X_0 onto the affine subspace $\{X \in \mathbb{R}^{d\times d} : \mathcal{A}(X) = 1_n\}$.

We now provide an outline of the proof that $X_{\text{LS}} \succeq 0$ and describe some of its challenges. A basic approach is to center around the deterministic matrix $M = (d^2 + d) I_n + d 1_n 1_n^T$. A straightforward rearrangement yields

$$X_{\text{LS}} = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1} 1_n = \mathcal{A}^*(I_n - M^{-1}\tilde{\Delta})^{-1} M^{-1} 1_n,$$

where $\tilde{\Delta} = M - \mathcal{A}\mathcal{A}^*$. To invert the matrix $I_n - M^{-1}\tilde{\Delta}$, we may expand it as a Neumann series

$$(I_n - M^{-1}\tilde{\Delta})^{-1} = I_n + M^{-1}\tilde{\Delta} + \sum_{i=2}^{\infty}(M^{-1}\tilde{\Delta})^i, \tag{2}$$

that converges if $\left\|M^{-1}\tilde{\Delta}\right\|_{op} < 1$. Observe that if the vector

$$u = (\mathcal{A}\mathcal{A}^*)^{-1}1_n = (I_n - M^{-1}\tilde{\Delta})^{-1}M^{-1}1_n$$

has non-negative coordinates with high probability, then we may immediately conclude that $X_{\text{LS}} \succeq 0$ since $\mathcal{A}^*(u) = \sum_i u_i v_i v_i^T$ is automatically PSD.

While it is possible to show that when $n \ll d^{3/2}$, the vector u indeed has positive coordinates, this argument suffers from a significant problem. It turns out that there is a phase transition at $n \asymp d^{3/2}$, beyond which the vector u switches from having non-negative coordinates to having both positive and negative ones. One reason for this is that the approximation $M^{-1}\tilde{\Delta} \approx I_n$ breaks down at the same $n \asymp d^{3/2}$ barrier. A previous version of this paper contained an error related to this non-negativity phenomenon. In Section K, we describe how this error can be fixed. However, we now present a cleaner approach that avoids this issue altogether.

To handle the positive and negative coordinates of u requires a different approach that more precisely takes into account its correlations with $\mathcal{A}^*$. We achieve this by first removing a rank-two component from $\mathcal{A}\mathcal{A}^*$ that prevents it from being close to the identity when $n \gg d^{3/2}$. Define

$$B = \mathcal{A}\mathcal{A}^* - (w1_n^T + 1_n w^T + d1_n 1_n^T)$$

where $w \in \mathbb{R}^n$ is defined by $w_i = \|v_i\|^2 - d$. As we show in Lemma 10, B is close to $(d^2 + d)I_n$ for all $n \leq d^2/\log^C(d)$. For this reason, B is well-behaved and amenable to Neumann expansion arguments.

Next, since $\mathcal{A}\mathcal{A}^*$ is the sum of B and a low rank matrix, we obtain a convenient expression for $(\mathcal{A}\mathcal{A}^*)^{-1}$ using the Woodbury matrix formula Woodbury (1950), which results in the following useful decomposition of the vector u (see Lemma 4):

$$u = (\mathcal{A}\mathcal{A}^*)^{-1}1_n = \rho \cdot \left(\lambda_1 B^{-1}1_n + \lambda_2 B^{-1}w\right),$$

where $\rho, \lambda_1, \lambda_2$ are certain scalar random variables. We show that $\rho > 0$, $\lambda_1 = 1 + 1_n^T B^{-1} w \sim 1$, and $\lambda_2 = -1_n^T B^{-1} 1_n = o(n/d^2) \ll \lambda_1$ with high probability. The proof then reduces to showing that

$$\mathcal{A}^*(B^{-1}1_n) \succeq (1 - o(1))\frac{n}{d^2}I_d \tag{3}$$

$$\|\mathcal{A}^*(B^{-1}w)\|_{op} = o(1). \tag{4}$$

Intuitively, (3) has non-negative coordinates since B is close to a multiple of the identity for the entire range $n \leq d^2/\log^C(d)$ and we have $B^{-1}1_n \approx 1_n$. With the same intuition, we expect that $B^{-1}w$ behaves like a multiple of w, which has i.i.d. centered coordinates. If w were independent of $\mathcal{A}^*$, significant cancellation would happen among the rank one vectors $\{v_i v_i^T\}_{i=1}^n$, yielding the bound (4) (by matrix Bernstein or its variants, see e.g. Tropp (2012)).

However, making this argument precise to take into account interactions between $\mathcal{A}^*$ and $B^{-1}w$ is a considerable technical challenge. To handle this, we expand B^{-1} as a Neumann series, similarly to (2). We then analyze terms of this series using the framework of graph matrices Ahn et al. (2016). Graph matrices provide a powerful tool for controlling the operator norm of certain matrices whose entries can be expressed as low-degree polynomials in i.i.d. random variables. Graph matrices serve to transform the analytic problem of controlling the operator norm of a random matrix X into a more tractable combinatorial one that involves studying certain weights of graphs associated to X. This part of the argument forms the bulk of our analysis and is detailed in Section B.

3. Future work

Towards the positive side of Conjecture 1 Towards understanding whether an explicit construction can be used resolve the positive side of Conjecture 1, we now discuss the following "identity perturbation" construction that is inspired by previous work Saunderson et al. (2013):

$$X_{\text{IP}} = \frac{1}{d} I_d + \mathcal{A}^*(\alpha) = \frac{1}{d} I_d + \sum_{i=1}^{n} \alpha_i v_i v_i^T,$$

where $\alpha \in \mathbb{R}^n$ is defined to be the unique solution of $\mathcal{A}(\mathcal{A}^*(\alpha)) = 1_n - \mathcal{A}(\frac{1}{d} I_d)$.[3] By definition of α, it always holds that $\mathcal{A}(X_{\text{IP}}) = 1_n$. In words, X_{IP} is obtained from the approximately fitting ellipsoid $\frac{1}{d} I_d$ by adding multiples of the constraint matrices $\{v_i v_i^T\}_{i=1}^n$ so that it exactly satisfies the linear constraints.

Our experimental results are depicted in Figure 1. For each (n, d) with $1 \leq d \leq n \leq 200$, we generated 10 independent instances of the ellipsoid fitting problem with n points in $\mathbb{R}^d$ and computed the fraction of instances for which each of the three constructions (original SDP, least-squares, and identity perturbation) was a valid fitting ellipsoid. The color of each cell corresponds to the fraction of "successful" instances, increasing in the following order: black (zero), red, orange, yellow, white (one). In each plot, the green curve corresponds to a function of the form $n(d) = cd^2$ for some constant c. See Appendix E for further details.

In summary, there appear to be two constants $c_{\text{LS}} \approx 1/17$ and $c_{\text{IP}} \approx 1/10$ such that the probabilities of PSD-ness of X_{LS} and X_{IP} undergo phase transitions from 1 to 0 asymptotically at $n = c_{\text{LS}} d^2$ and $n = c_{\text{IP}} d^2$, respectively. We emphasize that $c_{\text{LS}} < c_{\text{IP}} < 1/4$, meaning that there actually appear to be *three* distinct phase transitions related to the ellipsoid fitting problem. These results suggest that it is unlikely that the positive side of Conjecture 1 can be resolved by a sharper analysis of either of these two natural constructions.

In this work, we show that both the least-squares and identity perturbation constructions are positive semidefinite provided that $n \leq d^2/\text{polylog}(d)$. However, we still believe it is an interesting problem to sharply characterize the behavior of X_{LS} and X_{IP}. Given that X_{IP} appears to outperform X_{LS}, we now explain how one might approach this problem for X_{IP}. Again the central challenge is to show that $X_{\text{IP}} \succeq 0$ with high probability. Observe that $\alpha = (\mathcal{A}\mathcal{A}^*)^{-1} b$ and so $X_{\text{IP}} \succeq 0$ is implied by $\left\| \mathcal{A}^*((\mathcal{A}\mathcal{A}^*)^{-1} b) \right\|_{op} \leq 1/d$. We immediately recognize that to proceed with the analysis, we must invert $\mathcal{A}\mathcal{A}^*$ as in the analysis of X_{LS}. Applying the Neumann series expansion thus encounters the same bottlenecks as in the analysis of X_{LS}. We leave the problem of precisely characterizing

3. A similar construction is suggested in Saunderson et al. (2013), although no specific initialization is given.

the eigenvalues and eigenvectors of the random inner-product matrix $\mathcal{A}\mathcal{A}^*$ as a direction for future research.

Additionally, we remark that computing either of X_{LS} or X_{IP} amounts to applying the inverse of a certain $n \times n$ matrix to a vector. In contrast, testing whether the ellipsoid fitting property holds for a given set of points involves solving a semidefinite program, which requires a large polynomial runtime. To the best of our knowledge, it is an open question to find a faster algorithm achieving the conjectured threshold $n \sim d^2/4$, even in simulations.

Towards the negative side of Conjecture 1 As noted earlier, a simple dimension-counting argument (see Lemma 40) shows that when $n > d(d+1)/2$, the linear constraints alone are infeasible with probability 1. Any proof of the failure of the ellipsoid fitting property with high probability for $n > cd^2$ for a constant $c \in (1/4, 1/2)$ would likely yield significant insight into Conjecture 1.

Applications to other random SDPs In Appendix C, we prove a negative result showing that a certain SDP (which corresponds to the degree-2 SoS SDP relaxation) cannot certify a non-trivial lower bound on the discrepancy of random Gaussian matrices with m rows and n columns when $m < n - \sqrt{n}\,\text{polylog}(n)$. As we have mentioned, for simpler variants of the ellipsoid fitting problem, there are results of this type for SDPs corresponding to higher-degree SoS relaxations (e.g. Ghosh et al. (2020); Hsieh and Kothari (2022)). Is it true that higher-degree SoS SDPs also fail to certify non-trivial discrepancy lower bounds in the regime described above?

More generally, can one apply either the least-squares or identity-perturbation constructions to prove average-case SDP lower bounds for other problems? We expect that these constructions are tractable to analyze for SDPs with a PSD constraint and "simple" random linear constraints, such as the degree-2 SoS SDP relaxation of the clique number (see e.g. Section 2.2 of Barak et al. (2019)) and SoS relaxations of random systems of polynomial equations of the type in Hsieh and Kothari (2022).

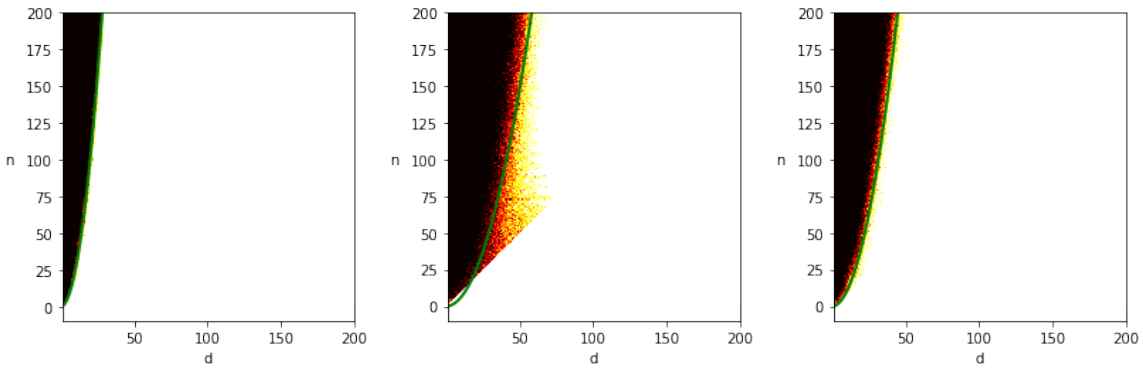

Figure 1: In each plot, the green curve corresponds to $n(d) = cn^2$. **(Left)** Ellipsoid fitting SDP, $c = 1/4$, **(Middle)**: Least-squares, $c = 1/17$, **(Right)**: Identity perturbation, $c = 1/10$

4. Proof of Theorem 2

As discussed in Section 2, it suffices to show that $X = X_{\mathrm{LS}} \succeq 0$. We make the simplification that $n = d^2/\mathrm{polylog}(d)$, as recorded in the following remark.

Remark 3 *By monotonicity (with respect to n) of the probability of the ellipsoid fitting property holding, it suffices to fix $n = d^2/\log^C(d)$ for some sufficiently large constant $C > 0$ to be determined. In fact, all of our technical lemmas below hold under the more general assumption that $d \le n \le d^2/\log^C(d)$.*

We proceed to showing $X \succeq 0$ by first separating out the low-rank and high-rank terms from $\mathcal{A}\mathcal{A}^*$ and then expanding the inverse as a Neumann series. Define the vector $w \in \mathbb{R}^n$ by $w_i = \|v_i\|_2^2 - d$ for every $i \in [n]$. Next, define the rank 2 matrix $W = w1_n^T + 1_n w^T + d1_n1_n^T \in \mathbb{R}^{n\times n}$, the high-rank matrix $\Gamma = \mathcal{A}\mathcal{A}^* - W - \alpha I_n \in \mathbb{R}^{n\times n}$, where $\alpha = d^2 + d$, and $B = \Gamma + \alpha I_n$. Then, we have the following decomposition:

$$\begin{aligned}\mathcal{A}\mathcal{A}^* &= \left(\mathcal{A}\mathcal{A}^* - (w1_n^T + 1_n w^T + d1_n1_n^T) - \alpha I_n\right) + (w1_n^T + 1_n w^T + d1_n1_n^T) + \alpha I_n \\ &= \Gamma + W + \alpha I_n \\ &= B + W.\end{aligned}$$

The following lemma is a consequence of the Woodbury matrix identity Woodbury (1950). We defer the proof to Appendix G.

Lemma 4 *Let $B = \Gamma + \alpha I_n$. We have*

$$(\mathcal{A}\mathcal{A}^*)^{-1}1_n = \frac{1}{s^2 - ru} \cdot \left((1 + 1_n^T B^{-1} w)B^{-1}1_n - (1_n^T B^{-1}1_n)B^{-1}w\right) \tag{5}$$

where r, s, u are defined as

$$\begin{pmatrix} r & s \\ s & u \end{pmatrix} := \begin{pmatrix} 1_n^T B^{-1}1_n & 1 + 1_n^T B^{-1} w \\ 1 + 1_n^T B^{-1} w & -d + w^T B^{-1} w \end{pmatrix}.$$

By Lemma 4, we have that

$$\begin{aligned} X &= \frac{1}{s^2 - ru}\mathcal{A}^*\left((1 + 1_n^T B^{-1} w)B^{-1}1_n - (1_n^T B^{-1}1_n)B^{-1}w\right) \\ &= \frac{1}{s^2 - ru}\left((1 + 1_n^T B^{-1} w)\,\mathcal{A}^*(B^{-1}1_n) - (1_n^T B^{-1}1_n)\mathcal{A}^*(B^{-1}w)\right).\end{aligned}$$

Clearly, $X \succeq 0$ follows if the next two conditions are satisfied:

$$s^2 - ru \ge 0, \tag{6}$$

and

$$(1 + 1_n^T B^{-1} w)\,\mathcal{A}^*(B^{-1}1_n) - (1_n^T B^{-1}1_n)\mathcal{A}^*(B^{-1}w) \succeq 0. \tag{7}$$

We verify that these two conditions are satisfied with high probability by invoking the following lemmas.

Lemma 5 *There is some constant $C > 0$ such that if $d \le n \le d^2/\log^C(d)$, then $1_n^T B^{-1} 1_n = \Theta(n/d^2)$ with high probability.*

Lemma 6 *There is some constant $C > 0$ such that if $d \le n \le d^2/\log^C(d)$, then $w^T B^{-1} w = \tilde{O}(n/d)$ with high probability.*

Lemma 7 *There is some constant $C > 0$ such that if $d \le n \le d^2/\log^C(d)$, then $|1_n^T B^{-1} w| = o(1)$ with high probability.*

Lemma 8 *There is some constant $C > 0$ such that if $d \le n \le d^2/\log^C(d)$, then $\mathcal{A}^*(B^{-1}1_n) \succeq (1 - o(1))\frac{n}{d^2} I_d$ with high probability.*

Lemma 9 *There is some constant $C > 0$ such that if $d \le n \le d^2/\log^C(d)$, then $\left\|\mathcal{A}^*(B^{-1}w)\right\|_{op} = o(1)$ with high probability.*

The proofs of Lemmas 5, 6, and 8 are contained in the next section. The proofs of Lemmas 7 and 9 are postponed to Section A.

For Condition (6), if $n \le d^2/\log^C(d)$ for a sufficiently large constant C, we have that with high probability

$$s^2 - ru \ge -ru = 1_n^T B^{-1} 1_n (d - w^T B^{-1} w) = \Theta(n/d^2)(d - \tilde{O}(n/d)) = \Theta(n/d) \ge 0, \quad (8)$$

for sufficiently large n, d, by Lemmas 5 and 6. For Condition (7), if $n \le d^2/\log^C(d)$ for a sufficiently large constant C, we have that with high probability

$$\begin{aligned}
(1 + 1_n^T B^{-1} w)\mathcal{A}^*(B^{-1}1_n) - (1_n^T B^{-1} 1_n)\mathcal{A}^*(B^{-1}w) & \\
&\succeq (1 - o(1))\mathcal{A}^*(B^{-1}1_n) - \Theta\left(\frac{n}{d^2}\right)\left\|\mathcal{A}^*(B^{-1}w)\right\|_{op} I_d \\
&\succeq \left((1 - o(1))(1 - o(1))\frac{n}{d^2} - \Theta\left(\frac{n}{d^2}\right)\left\|\mathcal{A}^*(B^{-1}w)\right\|_{op}\right) I_d \\
&= \left((1 - o(1))(1 - o(1))\frac{n}{d^2} - o\left(\frac{n}{d^2}\right)\right) I_d \succeq 0,
\end{aligned}$$

for sufficiently large n, d, by Lemmas 7, 5, 8, and 9.

5. Proofs of technical lemmas

The proofs of the remaining technical lemmas all make use of the following result, whose proof is postponed to Section B.2.

Lemma 10 *There is some constant $C > 0$ such that if $d \le n \le d^2/\log^C(d)$, then with high probability, $\|B - \alpha I_n\|_{op} = \tilde{O}(d\sqrt{n})$.*

We now show that Lemmas 5 and 6 follow from Lemma 10.

Proof [Proof of Lemma 5] By assumption on n and Lemma 10, with high probability, it holds that

$$0 \preceq (\alpha - \tilde{O}(d\sqrt{n}))I_n \preceq B \preceq (\alpha + \tilde{O}(d\sqrt{n}))I_n.$$

This implies that

$$(\alpha + \tilde{O}(d\sqrt{n}))^{-1} I_n \preceq B^{-1} \preceq (\alpha - \tilde{O}(d\sqrt{n}))^{-1} I_n.$$

The proof is complete by combining the previous line with the following fact:

$$\lambda_{min}(B^{-1})\|1_n\|^2 \leq 1_n^T B^{-1} 1_n \leq \lambda_{max}(B^{-1})\|1_n\|^2.$$

■

Proof [Proof of Lemma 6] As in the proof of Lemma 5, by assumption on n and Lemma 10, it holds with high probability that:

$$w^T B^{-1} w = \Theta\left(\frac{1}{d^2}\right) \cdot \|w\|_2^2 . \tag{9}$$

To complete the proof, it suffices to show that $\|w\|_2^2 = \sum_{i=1}^n (\|v_i\|_2^2 - d)^2 = \tilde{O}(nd)$ with high probability.

Note that for fixed i, we have conservatively that

$$|\|v_i\|_2^2 - d| \leq C(\log n)\sqrt{d} \tag{10}$$

with probability $n^{-C\cdot\Omega(1)}$ by Bernstein's inequality Vershynin (2018). Now for a large enough constant $C > 0$, using the union bound we have that (10) holds for all $1 \leq i \leq n$. Immediately we have $\|w\|_2^2 = \tilde{O}(nd)$, proving Lemma 6 after combining with (9). ■

5.1. Proof of Lemma 8

Define the matrix $\Delta = -\Gamma = \alpha I_n - B \in \mathbb{R}^{n\times n}$. By Lemma 10, we have that $\|\Delta\|_{op} = \tilde{O}(\max(n, d\sqrt{n}))$ with high probability. By our assumption that $n = O(d^2/\operatorname{polylog}(d))$, we have $\left\|\alpha^{-1}\Delta\right\|_{op} < 1$ (for d large enough). We may then conduct the following (convergent) Neumann series expansion:

$$\begin{aligned} B^{-1} &= (\alpha I_n - \Delta)^{-1} \\ &= \alpha^{-1}(I_n - \alpha^{-1}\Delta)^{-1} \\ &= \alpha^{-1}\sum_{k=0}^{\infty}(\alpha^{-1}\Delta)^k. \end{aligned}$$

Thus, we have that

$$\lambda_{min}(\mathcal{A}^*(B^{-1}1_n)) \geq \alpha^{-1}\lambda_{min}(\mathcal{A}^*(1_n)) - \sum_{k=1}^{\infty}\alpha^{-(k+1)}\left\|\mathcal{A}^*(\Delta^k 1_n)\right\|_{op}.$$

It is a standard fact from random matrix theory (see e.g. Theorem 4.7.1 of Vershynin (2018)) that when $n = \omega(d)$, then with high probability:

$$\lambda_{min}(\mathcal{A}^*(1_n)) = \lambda_{min}\left(\sum_{i=1}^n v_i v_i^T\right) = (1 - o(1))n.$$

To complete the proof, it suffices to show that with high probability:

$$\sum_{k=1}^{\infty} \alpha^{-(k+1)} \left\| \mathcal{A}^*(\Delta^k 1_n) \right\|_{op} = o\left(\frac{n}{d^2}\right).$$

To this end, introduce a truncation parameter $T \in \mathbb{N}$ and write:

$$\sum_{k=1}^{\infty} \alpha^{-(k+1)} \left\| \mathcal{A}^*(\Delta^k 1_n) \right\|_{op} = \sum_{k=1}^{T-1} \alpha^{-(k+1)} \left\| \mathcal{A}^*(\Delta^k 1_n) \right\|_{op} + \sum_{k=T}^{\infty} \alpha^{-(k+1)} \left\| \mathcal{A}^*(\Delta^k 1_n) \right\|_{op}.$$

Now, take $T = 2$ and recall $n \leq d^2/\operatorname{polylog}(d)$. The proof is complete by invoking Lemma 12 below with $k = 1$ to control the first summation and Lemma 11 below with $T = 2$ to control the second summation.

Although Lemma 11 below is only required with $T = 2$ in order to prove Lemma 8, its general form with $T \geq 2$ is crucial to the proofs of Lemmas 7 and 9.

Lemma 11 *Suppose $T \geq 1$. There is some constant $C > 0$ such that if $d \leq n \leq d^2/\log^C(d)$ then with high probability, it holds that*

$$\sum_{k=T}^{\infty} \alpha^{-(k+1)} \left\| \mathcal{A}^*(\Delta^k 1_n) \right\|_{op} = \tilde{O}\left(\frac{\sqrt{n}}{d}\right)^{T+1}.$$

Proof Note that

$$\left\| \mathcal{A}^*(\Delta^k 1_n) \right\|_{op} \leq \|\mathcal{A}^*\|_{2\to op} \|\Delta\|_{op}^k \|1_n\|_2.$$

By Lemma 10, $\alpha^{-1}\|\Delta\|_{op} = \tilde{O}(\sqrt{n}/d)$ with high probability by assumption on n. Combining these with the fact that $\|\mathcal{A}^*\|_{2\to op} = O(d)$ with high probability when $n = o(d^2)$ (see Lemma 3 of Saunderson (2011)), we may conclude by the geometric decay of the terms in the series that

$$\sum_{k=T}^{\infty} \alpha^{-(k+1)} \left\| \mathcal{A}^*(\Delta^k 1_n) \right\|_{op} = (\alpha^{-T}\|\Delta\|_{op}^T) \cdot \tilde{O}(\alpha^{-1} d\sqrt{n}) = \tilde{O}\left(\frac{\sqrt{n}}{d}\right)^{T+1}.$$

■

Lemma 12 *Let $k \in \mathbb{Z}_{\geq 1}$ be fixed. Then with probability $1 - n^{-\Omega(1)}$, it holds that*

$$\left\| \mathcal{A}^*(\Delta^k 1_n) \right\|_{op} \leq (\log n)^{O(k)} \cdot \sqrt{d} n^{3/4} \cdot O(\sqrt{n} d)^k.$$

The proof of this lemma is deferred to Section B.

As mentioned previously, the remaining proofs of Lemmas 7 and 9 are postponed to Section A.

Acknowledgments

Aaron Potechin was supported in part by NSF grant CCF-2008920. Prayaag Venkat was supported by an NSF Graduate Fellowship under grant DGE1745303 and Boaz Barak's Simons Investigator Fellowship, NSF grant DMS-2134157, DARPA grant W911NF2010021, and DOE grant DE-SC0022199, support from Oracle Labs and past support by the NSF, as well as the Packard and Sloan foundations and the BSF. Alexander S. Wein was supported by a Simons-Berkeley Research Fellowship and NSF grants CCF-2007443 and CCF-2106444.

We thank Sinho Chewi for helpful discussions during the early stages of this project and Tselil Schramm for helpful conversations and making us aware of the work of Amelunxen et al. Amelunxen et al. (2014). We also thank Yue Lu, Subhabrata Sen and Nati Srebro for helpful discussions. Prayaag Venkat and Alex Wein thank the Simons Institue for hosting them for the Fall 2021 program on Computational Complexity of Statistical Inference, during which part of this work was done.

References

Kwangjun Ahn, Dhruv Medarametla, and Aaron Potechin. Graph matrices: norm bounds and applications. *arXiv preprint arXiv:1604.03423*, 2016.

Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

Benjamin Aubin, Will Perkins, and Lenka Zdeborová. Storage capacity in symmetric binary perceptrons. *Journal of Physics A: Mathematical and Theoretical*, 52(29):294003, 2019.

Afonso S Bandeira, Dmitriy Kunisky, and Alexander S Wein. Computational hardness of certifying bounds on constrained PCA problems. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151, 2020.

Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019.

Moses Charikar, Alantha Newman, and Aleksandar Nikolov. Tight hardness results for minimizing discrepancy. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 1607–1614. SIAM, 2011.

Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. Sum-of-squares lower bounds for Sherrington-Kirkpatrick via planted affine planes. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–965. IEEE, 2020.

Samuel Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018.

Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

Jun-Ting Hsieh and Pravesh K Kothari. Algorithmic thresholds for refuting random polynomial systems. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1154–1203. SIAM, 2022.

Chris Jones, Aaron Potechin, Goutham Rajendran, Madhur Tulsiani, and Jeff Xu. Sum-of-squares lower bounds for sparse independent set. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 406–416. IEEE, 2022.

Daniel M Kane and Ilias Diakonikolas. A nearly tight bound for fitting an ellipsoid to gaussian random points. *arXiv preprint arXiv:2212.11221*, 2022.

Pravesh K Kothari and Peter Manohar. A stress-free sum-of-squares lower bound for coloring. *arXiv preprint arXiv:2105.07517*, 2021.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2022.

Cheng Mao and Alexander S Wein. Optimal spectral recovery of a planted vector in a subspace. *arXiv preprint arXiv:2105.15081*, 2021.

Sidhanth Mohanty, Prasad Raghavendra, and Jeff Xu. Lifting sum-of-squares lower bounds: degree-2 to degree-4. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 840–853, 2020.

Aleksandar Nikolov. The Komlós conjecture holds for vector colorings. *arXiv preprint arXiv:1301.4039*, 2013.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Anastasia Podosinnikova, Amelia Perry, Alexander S Wein, Francis Bach, Alexandre d'Aspremont, and David Sontag. Overcomplete independent component analysis via SDP. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2583–2592. PMLR, 2019.

Prasad Raghavendra, Satish Rao, and Tselil Schramm. Strongly refuting random CSPs below the spectral threshold. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 121–131, 2017.

James Saunderson, Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.

James Saunderson, Pablo A Parrilo, and Alan S Willsky. Diagonal and low-rank decompositions and fitting ellipsoids to random points. In *52nd IEEE Conference on Decision and Control*, pages 6031–6036. IEEE, 2013.

James Francis Saunderson. *Subspace identification via convex optimization*. PhD thesis, Massachusetts Institute of Technology, 2011.

Grant Schoenebeck. Linear level Lasserre lower bounds for certain k-CSPs. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 593–602. IEEE, 2008.

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

Paxton Turner, Raghu Meka, and Philippe Rigollet. Balancing gaussian vectors in high dimension. In *Conference on Learning Theory*, pages 3455–3486. PMLR, 2020.

Prayaag Venkat. Efficient algorithms for certifying lower bounds on the discrepancy of random matrices. *arXiv preprint arXiv:2211.07503*, 2022.

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

Max A Woodbury. *Inverting modified matrices*. Statistical Research Group, 1950.

Appendix A. Proofs of remaining technical lemmas

A.1. Proof of Lemma 9

Let $T \in \mathbb{N}$ be a truncation parameter. Using the same power series expansion as in the proof of Lemma 8 and the triangle inequality, we have that

$$\left\|\mathcal{A}^*(B^{-1}w)\right\|_{op} \leq \sum_{k=0}^{T-1} \alpha^{-(k+1)} \left\|\mathcal{A}^*(\Delta^k w)\right\|_{op} + \sum_{k=T}^{\infty} \alpha^{-(k+1)} \left\|\mathcal{A}^*(\Delta^k w)\right\|_{op}.$$

Now, let $C > 0$ be an absolute constant whose value we determine in the following, let $T = C\log(d)$ be an integer and let $n = d^2/\log^C(d)$. Our choice of $C > 0$ depends on Lemmas 13 and 14 that are stated below. There exists a sufficiently large choice of absolute constant $C > 0$ such that invoking Lemma 13 with $T = C\log(d)$ ensures the second summation above is $o(1)$ with high probability. There also exists a sufficiently large choice of absolute constant $C > 0$ such that a union bound and invocation of Lemma 14, for all $k \in \{0, \ldots, T-1\}$ ensures the first summation above is $o(1)$ with high probability. Setting C to be the maximum of these two choices completes the proof.

Lemma 13 *Suppose $T \geq 1$. There is some constant $C > 0$ such that if $d \leq n \leq d^2/\log^C(d)$ then with high probability, it holds that*

$$\sum_{k=T}^{\infty} \alpha^{-(k+1)} \left\|\mathcal{A}^*(\Delta^k w)\right\|_{op} = \sqrt{d} \cdot \tilde{O}\left(\frac{\sqrt{n}}{d}\right)^T.$$

Proof Note that

$$\left\|\mathcal{A}^*(\Delta^k w)\right\|_{op} \leq \|\mathcal{A}^*\|_{2\to op} \|\Delta\|_{op}^k \|w\|_2.$$

By Lemma 10 and assumption on n, $\alpha^{-1}\|\Delta\|_{op} = \tilde{O}(\sqrt{n}/d)$ with high probability. A standard calculation (see the proof of Lemma 6) reveals that $\|w\|_2 = \tilde{O}(\sqrt{nd})$ with high probability. Combining these with the fact that $\|\mathcal{A}^*\|_{2\to op} = O(d)$ with high probability when $n = o(d^2)$ (see Lemma 3 of Saunderson (2011)), we may conclude that

$$\sum_{k=T}^{\infty} \alpha^{-(k+1)} \left\|\mathcal{A}^*(\Delta^k w)\right\|_{op} = (\alpha^{-T}\|\Delta\|_{op}^T) \cdot \tilde{O}(\alpha^{-1} d^{3/2}\sqrt{n}) = \sqrt{d} \cdot \tilde{O}\left(\frac{\sqrt{n}}{d}\right)^T.$$

■

Lemma 14 *Let $k \in \mathbb{Z}_{\geq 0}$. Then with probability $1 - n^{-\Omega(1)}$, it holds that*

$$\left\|\mathcal{A}^*(\Delta^k w)\right\|_{op} \leq (\log n)^{O(k)} \cdot d\sqrt{n} \cdot O(\sqrt[4]{n} d^{3/2})^k.$$

The proof of this lemma is deferred to Section B.

A.2. Proof of Lemma 7

Let $T \in \mathbb{N}$ be a truncation parameter. Using the same power series expansion as in the proof of Lemma 8 and the triangle inequality, we have that

$$|1_n^T B^{-1} w| \leq \sum_{k=0}^{T-1} \alpha^{-(k+1)} |1_n^T \Delta^k w| + \sum_{k=T}^{\infty} \alpha^{-(k+1)} |1_n^T \Delta^k w|.$$

The argument requires Lemmas 15 and 16 stated below. Now, let $C > 0$ be some constant whose value we determine in the following, let $T = C\log(d)$ and let $n = d^2/\log^C(d)$. There exists a sufficiently large choice of C such that invoking Lemma 15 with $T = C\log(d)$ an integer ensures the second summation above is $o(1)$ with high probability. There also exists a sufficiently large choice of C such that invoking Lemma 16 for all $k \in \{0, \ldots, T-1\}$ with $\epsilon = o(1/T) = o(1/\log(d))$ ensures the first summation above is $o(1)$ with high probability. Setting C to be the maximum of these two choices completes the proof.

Lemma 15 *Suppose $T \geq 1$. There is some constant $C > 0$ such that if $d \leq n \leq d^2/\log^C(d)$ then with high probability, it holds that*

$$\sum_{k=T}^{\infty} \alpha^{-(k+1)} |1_n^T \Delta^k w| = \sqrt{d} \cdot \tilde{O}\left(\frac{\sqrt{n}}{d}\right)^T.$$

Proof Note that

$$|1_n^T \Delta^k w| \leq \|1_n\|_2 \|\Delta\|_{op}^k \|w\|_2.$$

By Lemma 10 and assumption on n, we have $\alpha^{-1}\|\Delta\|_{op} = \tilde{O}(\sqrt{n}/d)$ with high probability when $n = o(d^2)$. A standard calculation (see the proof of Lemma 6) reveals that $\|w\|_2 = \tilde{O}(\sqrt{nd})$ with high probability. Combining these, we may conclude that

$$\sum_{k=T}^{\infty} \alpha^{-(k+1)} |1_n^T \Delta^k w| = (\alpha^{-T}\|\Delta\|_{op}^T) \cdot \tilde{O}(\alpha^{-1} n\sqrt{d}) = \sqrt{d} \cdot \tilde{O}\left(\frac{\sqrt{n}}{d}\right)^T.$$

■

Lemma 16 *Let $k \in \mathbb{Z}_{\geq 0}$. Then with probability $1 - n^{-\Omega(1)}$, it holds that*

$$|1_n^T \Delta^k w| \leq \left\| \mathcal{A}^*(\Delta^k w) \right\|_{op} \leq (\log n)^{O(k)} \cdot d\sqrt{n} \cdot O(\sqrt[4]{n} d^{3/2})^k.$$

The proof of this lemma is nearly identical to that of Lemma 9 and is deferred to Section B.

Appendix B. Graph matrices

B.1. Background

We use the theory of *graph matrices* to derive operator norm bounds on various random matrices that arise in our analysis. Graph matrices provide a natural basis for decomposing matrices whose entries depend on random inputs, where this dependence has lots of symmetry but may be nonlinear. For our setting, we can define graph matrices as follows. These definitions are a special case of the definitions in Ahn et al. (2016) and are equivalent to the definitions in Ghosh et al. (2020) except that instead of summing over ribbons, we sum over injective maps. This gives a constant factor difference (see Remark 2.17 of Ahn et al. (2016)) in the final norm bounds.

In our analysis, many of the matrices we study, such as Δ^k, are $n \times n$ and have entries that are sums of terms of the form

$$M_{i_1,i_r} = \sum_{\substack{i_2,\ldots,i_{r-1} \\ k_1,\ldots,k_s}} \prod_{(x,y)\in E} f_{x,y}(v_{i_x,k_y}) \tag{11}$$

where $E \subset [r] \times [s]$, v_{i_x,k_y} is the k_y coordinate of v_{i_x}, $\{f_{x,y}\}$ are low-degree Hermite polynomials, and the indices of summation obey certain restrictions, including that $i_2, \ldots, i_{r-1}$ are distinct as well as $k_1, \ldots, k_s$.

The framework of graph matrices provides a convenient way of encoding these restrictions and attaining good norm bounds. Concretely, each matrix as in term (11) can be represented by a 'shape' consisting of a graph with r circle vertices, s square vertices, and integer edge labels. For a term like (11) which is an $n \times n$ matrix, there are two distinguished circle vertices that represent i_1 and i_r. The edges in the shape are specified by $E \subset [r] \times [s]$, and the vertices specify (distinct) indices of summation. The remaining circle vertices each represent an index of summation over $1 \leq i \leq n$ (i.e., one of $i_2, \ldots, i_{r-1}$) and a square vertex is used to represent an index of summation over $1 \leq k \leq d$ (i.e., one of $k_1, \ldots, k_s$, each of which indexes the dimension). The integer edge labels of the shape denote the degree of the Hermite polynomial that is applied to the random variable v_{i_x,k_y}. We make this precise with the following definitions.

Definition 17 (Normalized Hermite polynomials, see e.g. O'Donnell (2014), Chapter 11.2) *Define the sequence of normalized Hermite polynomials $h_0, h_1, h_2, \ldots$ by*

$$h_j(z) = \frac{1}{\sqrt{j!}} \cdot H_j(z),$$

where H_j are defined uniquely by the following formal power series in z:

$$\exp(tz - \frac{1}{2}t^2) = \sum_{j=0}^{\infty} \frac{1}{j!} H_j(z) t^j.$$

The first few Hermite polynomials are

$$h_0(z) = 1, \;\; h_1(z) = z, \;\; h_2(z) = \frac{1}{\sqrt{2}}(z^2 - 1), \;\; h_3(z) = \frac{1}{\sqrt{6}}(z^3 - 3z), \;\; \ldots.$$

Recall $\mathbb{E}_{Z \sim N(0,1)}[h_j(Z)h_k(Z)] = \delta_{jk}$, where δ_{jk} denotes the Kronecker function.

Definition 18 *A shape α is a graph that consists of the following:*

1. *A set of vertices $\mathcal{V}(\alpha)$. Each vertex is either a square or circle. We take $\mathcal{V}_\circ(\alpha)$ to be the set of circle vertices in $\mathcal{V}(\alpha)$ and we take $\mathcal{V}_\square(\alpha)$ to be the set of square vertices in $\mathcal{V}(\alpha)$.*

2. *Distinguished tuples of vertices U_α, V_α (which may intersect), which we call the left and right vertices of α, respectively. We also define the set of middle vertices as $W_\alpha = \mathcal{V}(\alpha) \setminus (U_\alpha \cup V_\alpha)$[4]. We take $U_{\alpha,\circ}$ to be the circle vertices of U_α (in the same order) and we take $U_{\alpha,\square}$ to be the square vertices of U_α (in the same order). Similarly, we take $V_{\alpha,\circ}$ to be the circle vertices of V_α (in the same order) and we take $V_{\alpha,\square}$ to be the square vertices of V_α (in the same order). We always take $U_\alpha = (U_{\alpha,\circ}, U_{\alpha,\square})$ and $V_\alpha = (V_{\alpha,\circ}, V_{\alpha,\square})$ so that circle vertices preceed square vertices in order.*

3. *A set $E(\alpha)$ of edges, where each edge is between a circle vertex and a square vertex. For each edge $e \in E(\alpha)$, we have a label $l_e \in \mathbb{Z}_{\geq 1}$. We define $|E(\alpha)| := \sum_{e \in E(\alpha)} l_e$. If a shape contains a multi-edge (i.e., two or more edges with the same endpoints), we call it improper (and proper otherwise). In a multi-edge, each copy of the edge has its own label. We represent an edge with endpoints u and v and label l by the notation $\{u, v\}_l$; we use the simpler notation $\{u, v\}$ when $l = 1$.*

Definition 19 *Given a shape α, we define M_α to be the $\frac{n!d!}{(n-|U_{\alpha,\circ}|)!(d-|U_{\alpha,\square}|)!} \times \frac{n!d!}{(n-|V_{\alpha,\circ}|)!(d-|V_{\alpha,\square}|)!}$ matrix with entries*

$$M_\alpha(A, B) = \sum_{\substack{\pi_\circ : \mathcal{V}_\circ(\alpha) \to [n], \pi_\square : \mathcal{V}_\square(\alpha) \to [d]: \\ \pi_\circ, \pi_\square \text{ are injective} \\ \pi_\circ(U_{\alpha,\circ}) = A_\circ, \pi_\square(U_{\alpha,\square}) = A_\square, \\ \pi_\circ(V_{\alpha,\circ}) = B_\circ, \pi_\square(V_{\alpha,\square}) = B_\square}} \left(\prod_{e = \{u,v\} \in E(\alpha) : u \in \mathcal{V}_\circ(\alpha), v \in \mathcal{V}_\square(\alpha)} h_{l_e}(v_{\pi_\circ(u), \pi_\square(v)}) \right) \tag{12}$$

where $A = (A_\circ, A_\square)$ is an ordered tuple such that $A_\circ$ is an ordered tuple of $|U_{\alpha,\circ}|$ elements from $[n]$ and $A_\square$ is an ordered tuple of $|U_{\alpha,\square}|$ elements from $[d]$, and $B = (B_\circ, B_\square)$ is an ordered tuple such that $B_\circ$ is an ordered tuple of $|V_{\alpha,\circ}|$ elements from $[n]$ and $B_\square$ is an ordered tuple of $|V_{\alpha,\square}|$ elements from $[d]$.

In the next section, we illustrate this definition by deriving the graph matrix representations of various matrices arising in our analysis. The proofs of Lemmas 8 and 9 boil down to obtaining norm bounds on $\mathcal{A}^*(\Delta^k z)$ for $z \in \{1_n, w\}$. Such a matrix is $d \times d$ and can be expressed as a sum of terms that are similar to (11):

$$M_{k_1,k_s} = \sum_{\substack{i_1, i_2, \ldots, i_{r-1}, i_r \\ k_2, \ldots, k_{s-1}}} \prod_{(x,y) \in E} f_{x,y}(v_{i_x, k_y}) \tag{13}$$

4. We abuse notation slightly by identifying tuples with the set composed of the union of their elements.

where again this is a graph matrix, and restrictions on the indices are encoded by an associated shape as described in Definition 19. The difference between this and (11) is that the distinguished vertices are now both squares instead of circles. Also, note that the restrictions imply that $i_1, \dots, i_r$ are distinct, as well as $k_2, \dots, k_{s-1}$.

B.2. Graph matrix representations

In this section, we derive the graph matrix representations of various matrices that arise in our analysis. For the purposes of computing $\mathcal{A}\mathcal{A}^*$, we can view $\mathcal{A}$ as an $n \times d^2$ matrix A with rows indexed by $i \in [n]$ and columns indexed by an ordered pair of indices in $(j,k) \in [d] \times [d]$, with entry $A_{i,(j,k)} = (v_i)_j (v_i)_k$. Given this entry-wise expression, the correctness of the graph matrix representation of A below can be directly verified by inspecting Equation (12) for the shapes below.

We decompose A as $A = M_{\alpha_{A1}} + M_{\alpha_{A2}}$ and $A^* = M_{\alpha_{A1}}^T + M_{\alpha_{A2}}^T$ for the following shapes α_{A1} and α_{A2} where we make the dimensions of $M_{\alpha_{A1}}$ and $M_{\alpha_{A2}}$ match by filling in the missing columns with zeros. These shapes are illustrated in Figures 2 and 3. Note that α_{A2} is improper. For each shape α considered below, its vertices $\mathcal{V}(\alpha)$ are given by $U_\alpha \cup V_\alpha \cup W_\alpha$:

- $U_{\alpha_{A1}} = (u)$ where u is a circle vertex, $V_{\alpha_{A1}} = (x_1, x_2)$ where x_1, x_2 are square vertices, $W_{\alpha_{A1}} = \{\}$ and $E(\alpha_{A1}) = \{\{u, x_1\}, \{u, x_2\}\}$. The matrix $M_{\alpha_{A1}}$with zeros filled in for the columns of $M_{\alpha_{A2}}$ has dimensions $n \times d^2$. Its $(i, (j,k))$ entry, for $i \in [n]$ and $(j,k) \in [d] \times [d]$ with $j \neq k$, is given by:

  $$M_{\alpha_{A1}}(i, (j,k)) = h_1((v_i)_j) h_1((v_i)_k) = (v_i)_j (v_i)_k.$$

  Its $(i, (j,j))$ entry, for $i \in [n]$ and $j \in [d]$, is zero.

- $U_{\alpha_{A2}} = (u)$ where u is a circle vertex, $V_{\alpha_{A2}} = (x, x)$ where x is a square vertex, $W_{\alpha_{A2}} = \{\}$ and $E(\alpha_{A2}) = \{\{u, x\}, \{u, x\}\}$. The matrix $M_{\alpha_{A2}}$with zeros filled in for the columns of $M_{\alpha_{A1}}$ has dimensions $n \times d^2$. Its $(i, (j,j))$ entry, for $i \in [n]$ and $j \in [d]$, is given by:

  $$M_{\alpha_{A1}}(i, (j,j)) = h_1((v_i)_j) h_1((v_i)_j) = (v_i)_j^2.$$

  Its $(i, (j,k))$ entry, for $i \in [n]$ and $j, k \in [d]$ with $j \neq k$, is zero.

Multiplying A and A^*, we see that $\mathcal{A}\mathcal{A}^* = AA^* \in \mathbb{R}^{n \times n}$ has (i,j) entry $(\mathcal{A}\mathcal{A}^*)_{ij} = \langle v_i, v_j \rangle^2$. We then obtain the following graph matrix representation:

$$\mathcal{A}\mathcal{A}^* = M_{\alpha_1} + M_{\alpha_2} + M_{\alpha_3} + M_{\alpha_{4'}}$$

where α_1, α_2, α_3, and $\alpha_{4'}$ are the following shapes (note that α_2, α_3, and $\alpha_{4'}$ are improper):

- $U_{\alpha_1} = (u)$ and $V_{\alpha_1} = (v)$ where u, v are circle vertices, $W_{\alpha_1} = \{x_1, x_2\}$ where x_1, x_2 are square vertices, and $E(\alpha_1) = \{\{u, x_1\}, \{u, x_2\}, \{x_1, v\}, \{x_2, v\}\}$; see Figure 7. The matrix M_{α_1} has dimensions $n \times n$. Its (i,j) entry, for $i, j \in [n]$ with $i \neq j$ is given by:

  $$M_{\alpha_1}(i,j) = \sum_{k,l \in [d], k \neq l} h_1((v_i)_k) h_1((v_j)_k) h_1((v_i)_l) h_1((v_j)_l) = \sum_{k,l \in [d], k \neq l} (v_i)_k (v_j)_k (v_i)_l (v_j)_l.$$

  If $i = j$, then note that $M_{\alpha_1}(i,j) = 0$.

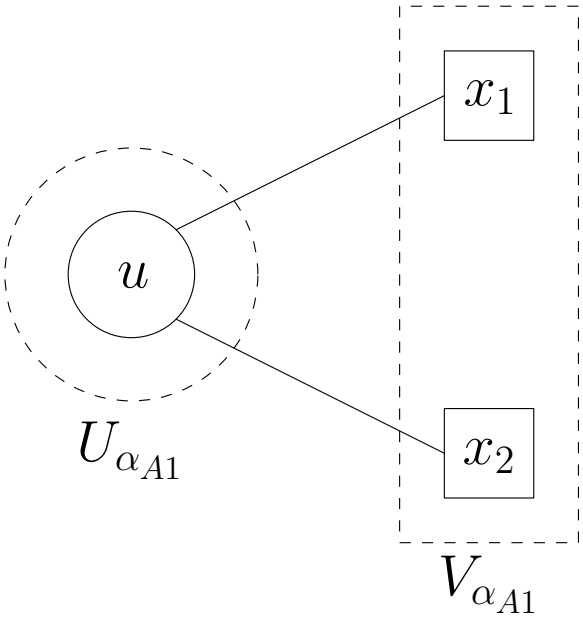

Figure 2: Shape α_{A1}.

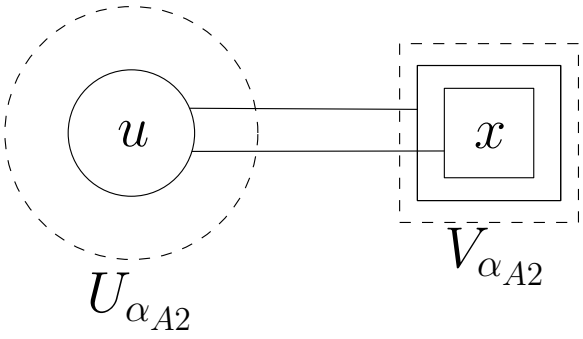

Figure 3: Shape α_{A2}. Here, we depict the shape α_{A2} by drawing the two identical copies of the square vertex x as two overlapping squares sharing the label x. Note that the edge $\{u, x\}$ is a multi-edge, so the shape is improper.

- $U_{\alpha_2} = (u)$ and $V_{\alpha_2} = (v)$ where u, v are circle vertices, $W_{\alpha_2} = \{x\}$ where x is a square vertex, and $E(\alpha_2) = \{\{u, x\}, \{u, x\}, \{x, v\}, \{x, v\}\}$; see Figure 4. The matrix M_{α_2} has dimensions $n \times n$. Its (i, j) entry, for $i, j \in [n]$ with $i \neq j$ is given by:

$$M_{\alpha_2}(i, j) = \sum_{k \in [d]} h_1((v_i)_k)^2 h_1((v_j)_k)^2 = \sum_{k \in [d]} (v_i)_k^2 (v_j)_k^2.$$

  If $i = j$, then note that $M_{\alpha_2}(i, j) = 0$.

- $U_{\alpha_3} = V_{\alpha_3} = (u)$ where u is a circle vertex, $W_{\alpha_3} = \{x_1, x_2\}$ where x_1, x_2 are square vertices, and $E(\alpha_3) = \{\{u, x_1\}, \{u, x_1\}, \{u, x_2\}, \{u, x_2\}\}$; see Figure 5.

- $U_{\alpha_{4'}} = V_{\alpha_{4'}} = (u)$ where u is a circle vertex, $W_{\alpha_{4'}} = \{x\}$ where x is a square vertex, and $E(\alpha_{4'}) = \{\{u, x\}, \{u, x\}, \{u, x\}, \{u, x\}\}$; see Figure 6.

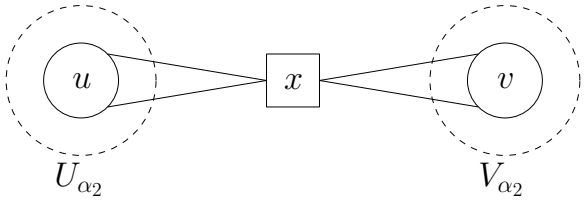

Figure 4: Shape α_2, one of the improper shapes appearing in $\mathcal{AA}^*$.

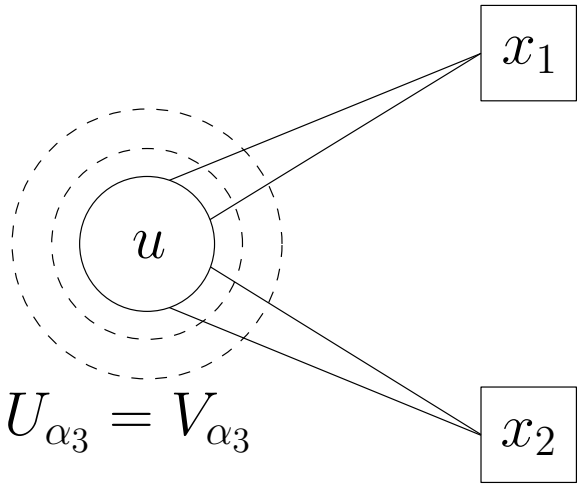

Figure 5: Shape α_3, one of the improper shapes appearing in $\mathcal{AA}^*$.

Finally, we express the vectors $w, 1_n \in \mathbb{R}^n$ as $n \times 1$ graph matrices. Recall that $w_i = \|v_i\|_2^2 - d$ and that $h_2(z) = \frac{1}{\sqrt{2}}(z^2 - 1)$. So, w is represented by the shape α_w with leading coefficient $\sqrt{2}$, and 1_n is represented by the shape α_{1_n} with leading coefficient 1:

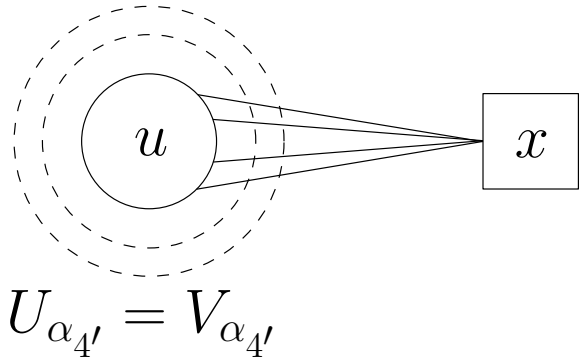

Figure 6: Shape $\alpha_{4'}$, one of the improper shapes appearing in $\mathcal{AA}^*$.

- $U_{\alpha_w} = (u)$ where u is a circle vertex, $V_{\alpha_w} = \emptyset$, $W_{\alpha_w} = \{x\}$ where x is a square vertex, and $E(\alpha_w) = \{\{u, x\}_2\}$.
- $U_{\alpha_{1_n}} = (u)$ where u is a circle vertex, $V_{\alpha_{1_n}} = \emptyset$, and $E(\alpha_{1_n}) = \emptyset$.

B.2.1. RESOLVING MULTI-EDGES

As we demonstrate later, it is important for the purposes of our analysis that all shapes we work with are proper. To shift from improper shapes (i.e. ones with multi-edges) to proper shapes (i.e. ones without multi-edges), we record the following proposition.

Proposition 20 *Let α be a shape which contains two or more copies of an edge e. Consider two such copies of e that have labels $i, j \in \mathbb{Z}_{\geq 1}$, respectively. Then, we have*

$$M_\alpha = \sum_{k=0}^{\infty} c_k M_{\alpha_k},$$

where α_k is the shape that is identical to α, except that the two labeled copies of e are replaced by a single copy of e with label k, and $\{c_k : k \in \mathbb{Z}_{\geq 0}\}$ are coefficients that satisfy:

$$h_i(x)h_j(x) = \sum_{k=0}^{\infty} c_k h_k(x).$$

That is, the coefficients are obtained by writing the polynomial $h_i \cdot h_j$ in the Hermite basis. In particular, it holds that $c_k = 0$ unless $i + j + k$ is even and $k \leq i + j$. In other words, for each term we obtain, the parity of k is the same as the parity of $i + j$. We regard any edge with label 0 as a non-edge and say that such an edge vanishes.

Note that we may convert two or more parallel labeled edges into a single labeled edges by repeated application of Proposition 20. The proof of this result follows from Definition 19. The parity result follows from elementary calculations involving the Hermite polynomials, which we defer to Section H.

Given Proposition 20, we replace the improper shapes from the previous section with proper shapes to obtain the following graph matrix representation:

$$\mathcal{A}\mathcal{A}^* = M_{\alpha_1} + 2M_{\alpha_{2a}} + \sqrt{2}M_{\alpha_{2b}} + \sqrt{2}M_{\alpha_{2c}} + dM_{\alpha_{2d}} + 2M_{\alpha_{3a}} + 2\sqrt{2}(d-1)M_{\alpha_{3b}} + (d^2-d)M_{\alpha_{3c}} + \sqrt{24}M_{\alpha_4} + 6\sqrt{2}M_{\alpha_{3b}} + 3dM_{\alpha_{3c}}$$

where α_1, α_{2a}, α_{2b}, α_{2c}, α_{2d}, α_{3a}, α_{3b}, α_{3c}, and α_4 are the following proper shapes that we define below. First, α_1 is the same as above since it is already proper.

Second, observe that α_2 has two sets of double edges. Using the following identities (see also Section H):

$$h_1(x)^2 = \sqrt{2}h_2(x) + 1$$
$$h_1(x)^2 h_1(y)^2 = 2h_2(x)h_2(y) + \sqrt{2}h_2(x) + \sqrt{2}h_2(y) + 1,$$

we can replace each of these double edges by a linear combination of an edge with label 2 and a non-edge. We then write $M_{\alpha_2} = 2M_{\alpha_{2a}} + \sqrt{2}M_{\alpha_{2b}} + \sqrt{2}M_{\alpha_{2c}} + d \cdot M_{\alpha_{2d}}$ as a linear combination of $2 \times 2 = 4$ graph matrices associated with the shapes $\alpha_{2a}, \alpha_{2b}, \alpha_{2c}, \alpha_{2d}$ defined as follows:

- $U_{\alpha_{2a}} = (u)$ and $V_{\alpha_{2a}} = (v)$ where u, v are circle vertices, $W_{\alpha_{2a}} = \{x\}$ where x is a square vertex, and $E(\alpha_{2a}) = \{\{u, x\}_2, \{x, v\}_2\}$. The matrix $M_{\alpha_{2a}}$ has dimensions $n \times n$. Its (i, j) entry, for $i, j \in [n]$ with $i \neq j$ is given by:

$$M_{\alpha_{2a}}(i, j) = \sum_{k \in [d]} h_2((v_i)_k) h_2((v_j)_k).$$

  If $i = j$, then note that $M_{\alpha_{2a}}(i, j) = 0$.

- $U_{\alpha_{2b}} = (u)$ and $V_{\alpha_{2b}} = (v)$ where u, v are circle vertices, $W_{\alpha_{2b}} = \{x\}$ where x is a square vertex, and $E(\alpha_{2b}) = \{\{u, x\}_2\}$.

- $U_{\alpha_{2c}} = (u)$ and $V_{\alpha_{2c}} = (v)$ where u, v are circle vertices, $W_{\alpha_{2c}} = \{x\}$ where x is a square vertex, and $E(\alpha_{2c}) = \{\{x, v\}_2\}$.

- $U_{\alpha_{2d}} = (u)$ and $V_{\alpha_{2d}} = (v)$ where u, v are circle vertices, $W_{\alpha_{2d}} = \{\}$, and $E(\alpha_{2d}) = \{\}$. Note that we have made the following simplification in describing α_{2d}, which arises when we replace each of the double edges in α_2 by non-edges. This will leave α_{2d} with an isolated middle square vertex x. However, observe that from Equation (12), we may equivalently delete the isolated vertex x and multiply the resulting graph matrix by a d factor. In summary, we work with the definition of α_{2d} which does not contain a middle square vertex, but which has an associated scalar coefficient of d.

Third, using the same approach as in re-expressing α_2, we write $M_{\alpha_3} = 2M_{\alpha_{3a}} + 2\sqrt{2}(d-1)M_{\alpha_{3b}} + d(d-1)M_{\alpha_{3c}}$ as a linear combination of 3 graph matrices associated with the shapes $\alpha_{3a}, \alpha_{3b}, \alpha_{3c}$ defined as follows:

- $U_{\alpha_{3a}} = V_{\alpha_{3a}} = (u)$ where u is a circle vertex, $W_{\alpha_{3a}} = \{x_1, x_2\}$ where x_1, x_2 are square vertices, and $E(\alpha_{3a}) = \{\{u, x_1\}_2, \{u, x_2\}_2\}$.

- $U_{\alpha_{3b}} = V_{\alpha_{3b}} = (u)$ where u is a circle vertex, $W_{\alpha_{3b}} = \{x\}$ where x is a square vertex, and $E(\alpha_{3b}) = \{\{u, x\}_2\}$.
- $U_{\alpha_{3c}} = V_{\alpha_{3c}} = (u)$ where u is a circle vertex, $W_{\alpha_{3c}} = \{\}$, and $E(\alpha_{3c}) = \{\}$.

Fourth, for $\alpha_{4'}$, we use the following identity to replace its quadruple edge by a linear combination of an edge with label 4, an edge with label 2, and a non-edge (see also Section H):

$$h_1(x)^4 = (x^4 - 6x^2 + 3) + 6(x^2 - 1) + 3 = \sqrt{24}h_4(x) + 6\sqrt{2}h_2(x) + 3.$$

We write $M_{\alpha_{4'}} = \sqrt{24}M_{\alpha_4} + 6\sqrt{2}M_{\alpha_{3b}} + 3dM_{\alpha_{3c}}$ as a linear combination of 3 graph matrices associated with the shapes $\alpha_4, \alpha_{3b}, \alpha_{3c}$, where α_4 is defined as follows:

- $U_{\alpha_4} = V_{\alpha_4} = (u)$ where u is a circle vertex, $W_{\alpha_4} = \{x\}$ where x is a square vertex, and $E(\alpha_4) = \{\{u, x\}_4\}$.

We may further simplify by observing that $M_{\alpha_{2d}} = 1_n 1_n^T - I_n$ and $M_{\alpha_{3c}} = I_n$, which leads to following graph matrix representation involving only proper shapes:

$$\mathcal{A}\mathcal{A}^* = (d^2+d)I_n + d1_n 1_n^T + M_{\alpha_1} + 2M_{\alpha_{2a}} + \sqrt{2}M_{\alpha_{2b}} + \sqrt{2}M_{\alpha_{2c}} + 2M_{\alpha_{3a}} + (2\sqrt{2}d+4\sqrt{2})M_{\alpha_{3b}} + \sqrt{24}M_{\alpha_4}.$$

In order to decompose $B = \mathcal{A}\mathcal{A}^* - W$ in terms of graph matrices, we first decompose W as follows:

$$W = w1_n^T + 1_n w^T + d1_n 1_n^T = \sqrt{2}M_{\alpha_{2b}} + \sqrt{2}M_{\alpha_{2c}} + 2\sqrt{2}M_{\alpha_{3b}} + d1_n 1_n^T.$$

Combining these decompositions, we have:

$$B = \mathcal{A}\mathcal{A}^* - W = (d^2 + d)I_n + M_{\alpha_1} + 2M_{\alpha_{2a}} + 2M_{\alpha_{3a}} + (2\sqrt{2}d + 2\sqrt{2})M_{\alpha_{3b}} + \sqrt{24}M_{\alpha_4}, \tag{14}$$

$$\Delta = -M_{\alpha_1} - 2M_{\alpha_{2a}} - 2M_{\alpha_{3a}} - (2\sqrt{2}d + 2\sqrt{2})M_{\alpha_{3b}} - \sqrt{24}M_{\alpha_4}. \tag{15}$$

Define the index set $\mathcal{I} = \{1, 2a, 3a, 3b, 4\}$, which collects the indices of non-identity shapes appearing in B; see Figures 7, 8, 9, 10, 11. For a given index $i \in \mathcal{I}$, we define λ_i to be the scalar coefficient appearing in front of M_{α_i} in the expression for B above.

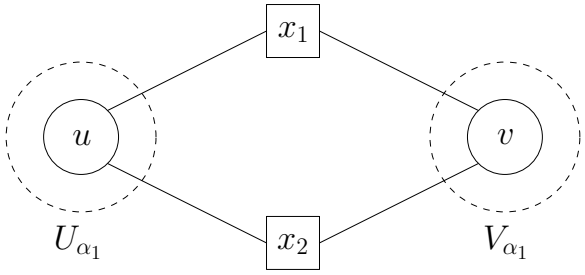

Figure 7: Shape α_1

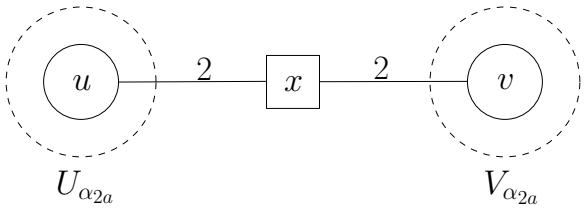

Figure 8: Shape α_{2a}

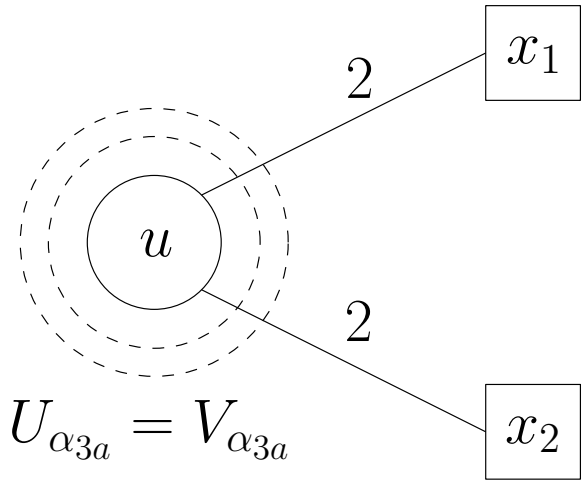

Figure 9: Shape α_{3a}

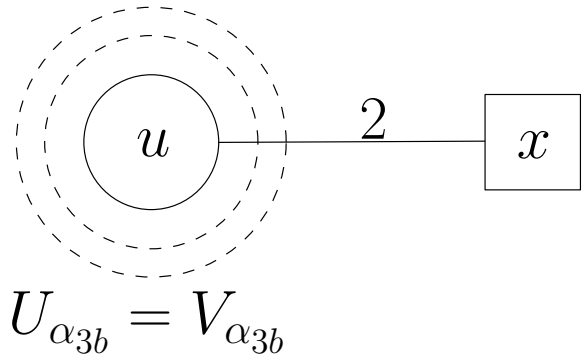

Figure 10: Shape α_{3b}

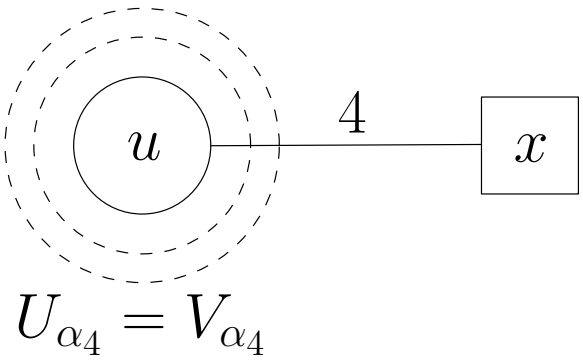

Figure 11: Shape α_4

B.3. Graph matrix norm bounds

As mentioned earlier, graph matrices admit norm bounds that only depend on certain combinatorial parameters associated with the graph, as expressed by the following theorem that follows from Ahn et al. (2016). We defer its proof to Appendix F. To complete the proofs of Lemmas 10, 12, 14, and 16, we will derive graph matrix representations of the relevant matrices, estimate their associated combinatorial parameters, and then invoke the theorem below. To state the theorem, we borrow some notions from Ahn et al. (2016).

Given a shape α and a vertex $a \in \mathcal{V}(\alpha)$, we define

$$\varphi(a) = \begin{cases} 1 & \text{if } a \text{ is a circle,} \\ \log_n(d) & \text{if } a \text{ is a square.} \end{cases}$$

For a (potentially empty) subset of vertices $S \subset \mathcal{V}(\alpha)$, define $\varphi(S) = \sum_{s \in S} \varphi(s)$.

Definition 21 *Define the min-vertex separator $S_{\min}$ of α to be a (potentially empty) subset of vertices of α with the smallest value of $\varphi(S)$ such that all paths from U_α to V_α intersect S. Here, we allow for paths of length 0, so any separator between U_α and V_α must contain $U_\alpha \cap V_\alpha$. We also define $\mathrm{Iso}(\alpha)$ to consist of all isolated vertices lying in $\mathcal{V}(\alpha) \backslash (U_\alpha \cup V_\alpha)$.*

Theorem 22 *Given $D_V, D_E \in \mathbb{N}$ such that $D_E \geq D_V \geq 2$ and $\epsilon > 0$, with probability at least $1 - \epsilon$, for all shapes α on square and circle vertices such that $|\mathcal{V}(\alpha)| \leq D_V$, $|E_\alpha| \leq D_E$, $|U_\alpha| \leq 1$, and $|V_\alpha| \leq 1$,*

$$\|M_\alpha\|_{op} \leq \left((2D_E + 2)\ln(D_V) + \ln(11n) + \ln\left(\frac{1}{\epsilon}\right) \right)^{|\mathcal{V}(\alpha)| + |E(\alpha)|} n^{\frac{\varphi(\mathcal{V}(\alpha)) - \varphi(S_{min}) + \varphi(\mathrm{Iso}(\alpha))}{2}}$$

where S_{min} is a min-vertex separator of α.

In our applications of this result, we invoke it with $\epsilon = 1/\operatorname{poly}(n)$ on shapes α for which $|\mathcal{V}(\alpha)|, |E_\alpha| \leq \operatorname{polylog}(n)$, so that the norm bounds take the form:

$$\|M_\alpha\|_{op} = (\log n)^{O(|\mathcal{V}(\alpha)| + |E(\alpha)|)} \cdot n^{\frac{\varphi(\mathcal{V}(\alpha)) - \varphi(S_{\min}) + \varphi(\mathrm{Iso}(\alpha))}{2}}.$$

With Theorem 22 and the graph matrix representations from Section B.2 in hand, we can immediately prove Lemma 10. The proof also demonstrates how the usage of graph matrices allows us to reduce the challenging problem of bounding the norm of a "complicated" random matrix to a significantly simpler combinatorial problem.

Proof [Proof of Lemma 10] Recall from Equation (14) that we have

$$\|B - \alpha I_n\|_{op} \leq \|M_{\alpha_1}\|_{op} + 2\|M_{\alpha_{2a}}\|_{op} + 2\|M_{\alpha_{3a}}\|_{op} + (2\sqrt{2}d + 2\sqrt{2})\|M_{\alpha_{3b}}\|_{op} + \sqrt{24}\|M_{\alpha_4}\|_{op}.$$

To complete the proof, we will invoke Theorem 22 to upper bound the norm of each of the 5 graph matrices above. In the following, for each of the 5 shapes, we identify the min-vertex separator and then estimate the combinatorial parameters appearing in the bound in Theorem 22.

Recall that for $a \in \mathcal{V}(\alpha)$, we have $\varphi(a) = \log_n(n) = 1$ if a is a circle and $\varphi(a) = \log_n(d) \approx 1/2$ if a is a square (since we consider the regime $n \leq d^2/\operatorname{polylog}(d)$) and that $\mathrm{Iso}(\alpha)$ is the set of isolated vertices that do not lie in U_α or V_α.

- **Term M_{α_1}:** Consider the following vertex separators: $\{u\}, \{v\}, \{x_1, x_2\}$. By inspection, any other vertex separator contains one of these three. The weights of $\{u\}$ and $\{v\}$ are both 1. The weight of $\{x_1, x_2\}$ is $2\log_n(d) > 1$. Thus we may choose u as a min-vertex separator without loss of generality. Thus for α_1, we have

$$\frac{\varphi(\mathcal{V}(\alpha)) - \varphi(S_{\min}) + \varphi(\mathrm{Iso}(\alpha))}{2} = \frac{(2\log_n(d) + 2) - 1 + 0}{2} = \frac{2\log_n(d) + 1}{2},$$

  leading to a norm bound $\|M_{\alpha_1}\|_{op} = \tilde{O}(d\sqrt{n})$ with high probability by Theorem 22.

- **Term $M_{\alpha_{2a}}$:** Every vertex is a separator of $U_{\alpha_{2a}}$ and $V_{\alpha_{2a}}$. Since x has weight $\log_n(d) < 1$ and u, v have weight 1 in α_2, the minimum weight vertex separator is x. Thus for α_{2a}, we have

$$\frac{\varphi(\mathcal{V}(\alpha)) - \varphi(S_{\min}) + \varphi(\mathrm{Iso}(\alpha))}{2} = \frac{(2 + \log_n(d)) - \log_n(d) + 0}{2} = 1,$$

  leading to a high probability norm bound $\|M_{\alpha_{2a}}\|_{op} = \tilde{O}(n)$ by Theorem 22.

The remaining shapes represent matrices that are diagonal; thus u is the min-vertex separator.

- **Term $M_{\alpha_{3a}}$:** By similar arguments to the above, Theorem 22 yields $\|M_{\alpha_{3a}}\| = \tilde{O}(d)$.

- **Term $M_{\alpha_{3b}}$:** We obtain a norm bound $\tilde{O}(n^{(1+\log_n d - 1 + 0)/2} = \tilde{O}(d^{1/2})$, so its contribution to Δ has operator norm at most $\tilde{O}(d^{3/2})$ (see (15)).

- **Term M_{α_4}:** Similarly, we obtain the norm bound $\tilde{O}(d^{1/2})$.

Assembling these bounds completes the proof. ■

B.4. Tools for dealing with products of shapes

As mentioned earlier, our analysis involves large powers the matrix Δ. In the following, we will derive a graph matrix representation for Δ. Unfortunately, explicitly writing down such a representation for Δ^k for arbitrary $k \in \mathbb{N}$ is complicated. To overcome this issue, we now introduce some definitions and technical results that allow us to express in a systematic way the graph matrix representation of a product of matrices in terms of the representations of the individual matrices. The following result follows directly from the formula in Definition 19.

Proposition 23 (Multiplication rule) *Given shapes α and β such that V_α and U_β match (i.e. V_α and U_β have the same number of circle and square vertices), the product $M_\alpha M_\beta$ is a linear combination of graph matrices M_γ of shapes γ of the following form:*

1. *Glue α and β together by setting $V_\alpha = U_\beta$; these vertices now become middle vertices of γ[5]. We set $U_\gamma = U_\alpha$ to be the left side of γ and we set $V_\gamma = V_\beta$ to be the right side of γ.*

2. *The possible realizations of γ are obtained by considering all possible ways in which the vertices in $\mathcal{V}(\beta) \setminus U_\beta$ may intersect with the vertices in $\mathcal{V}(\alpha) \setminus V_\alpha$. For a possible intersection of $\mathcal{V}(\beta) \setminus U_\beta$ and $\mathcal{V}(\alpha) \setminus V_\alpha$ to give rise to γ, it must satisfy the following constraints:*

   (a) *For a given intersection of $\mathcal{V}(\beta) \setminus U_\beta$ and $\mathcal{V}(\alpha) \setminus V_\alpha$ and a vertex v which is in this intersection, we say that the occurrence of v in $\mathcal{V}(\beta) \setminus U_\beta$ is identified with the occurrence of v in $\mathcal{V}(\alpha) \setminus V_\alpha$. Circle vertices can only be identified with other circle vertices and square vertices can only be identified with other square vertices.*

   (b) *The vertices in $\mathcal{V}(\alpha) \setminus V_\alpha$ must remain distinct as well as the vertices in $\mathcal{V}(\beta) \setminus U_\beta$. In other words, each vertex can only be identified with at most one other vertex (which must be in the other shape).*

We illustrate Proposition 23 in Figure 12 by multiplying two shapes that arise in the multiplication Δw.

B.4.1. ACTION OF $\mathcal{A}^*$ AS A GRAPH MATRIX

In this section, we record for later use how to express the action of the linear operator $\mathcal{A}^*$ in terms of graph matrices. Taking the transpose of the shapes α_{A1}, α_{A2} from Section B.2 and applying Proposition 20 to resolve multi-edges, we define the following shapes $\alpha_{A^*,1}$, $\alpha_{A^*,2}$, and $\alpha_{A^*,3}$; see also Figures 13, 14, 15.

1. $U_{\alpha_{A^*,1}} = (u)$ and $V_{\alpha_{A^*,1}} = (v)$ where u, v are square vertices, $W_{\alpha_{A^*,1}} = \{x\}$ where x is a circle vertex, and $E(\alpha_{A^*,1}) = \{\{u, x\}, \{x, v\}\}$. This shape has an associated coeffcient of 1.

2. $U_{\alpha_{A^*,2}} = V_{\alpha_{A^*,2}} = (u)$ where u is a square vertex, $W_{\alpha_{A^*,2}} = \{x\}$ where x is a circle vertex, and $E(\alpha_{A^*,2}) = \emptyset$. This shape has an associated coeffcient of 1.

3. $U_{\alpha_{A^*,3}} = V_{\alpha_{A^*,3}} = (u)$ where u is a square vertex, $W_{\alpha_{A^*,3}} = \{x\}$ where x is a circle vertex, and $E(\alpha_{A^*,3}) = \{\{u, x\}_2\}$. This shape has an associated coeffcient of $\sqrt{2}$.

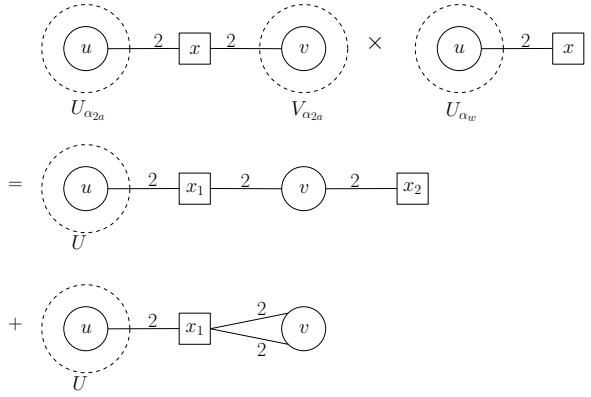

Figure 12: Multiplication of $M_{\alpha_{2a}}$ and M_{α_w}. Since $V_{\alpha_w} = \emptyset$, it is not depicted above; as result, the shapes in the resulting product also have $V = \emptyset$. Also, note that the second shape of the resulting product has a multi-edge. Using Proposition 20 as before, we can express this multi-edge as a linear combination of 3 labeled edges, with labels 0, 2, and 4, respectively.

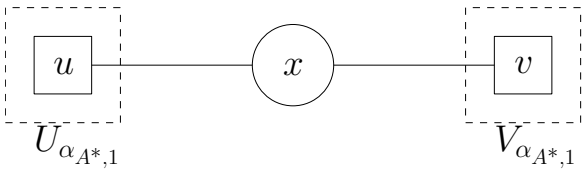

Figure 13: Shape $\alpha_{A^*,1}$

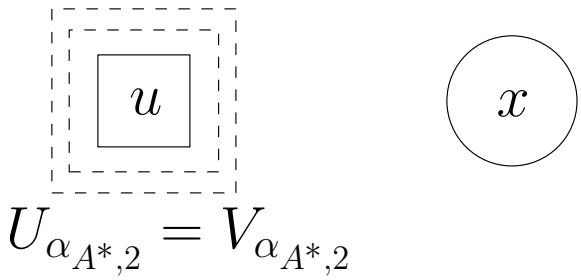

Figure 14: Shape $\alpha_{A^*,2}$

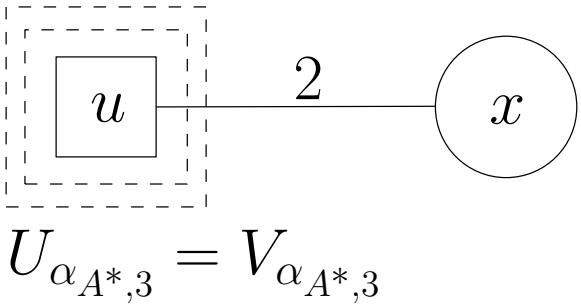

Figure 15: Shape $\alpha_{A^*,3}$

Let $x \in \mathbb{R}^n$ be a $(n \times 1)$ graph matrix represented by the shape β. The $d \times d$ matrix $\mathcal{A}^*(x)$ can be represented as a linear combination of graph matrices in the following way.

1. First, "re-shape" each of the $\alpha_{A^*,j}$ for $j = 1, 2, 3$ into shapes that represent $d^2 \times n$ matrices by redefining $U_{\alpha_{A^*,j}} \leftarrow U_{\alpha_{A^*,j}} \cup V_{\alpha_{A^*,j}}$ and $V_{\alpha_{A^*,j}} \leftarrow \{x\}$.
2. Invoke Proposition 23 to multiply each of these shapes by x.
3. Reshape the resulting shapes, which represent $d^2 \times 1$ matrices, into shapes representing $d \times d$ matrices by defining the left vertex set to contain only u and the right vertex set to contain only v (for $\alpha_{A^*,1}$) or the left vertex set and right vertex set to contain only u (for $\alpha_{A^*,2}, \alpha_{A^*,3}$) and defining all other vertices to be middle vertices.

See Figures 16 and 17 for an example of such a multiplication arising in the product $\mathcal{A}^*(\Delta w)$.

B.5. Norm bound strategy using graph matrices

We need to bound norms of matrices of the following forms:

1. $\{\mathcal{A}^*(\Delta^k w) : k \in \mathbb{N}\}$
2. $\{\mathcal{A}^*(\Delta^k 1_n) : k \in \mathbb{N}\}$
3. $\{1_n^T \Delta^k w : k \in \mathbb{N}\}$. (Here, we regard the scalars as 1×1 matrices.)

For a fixed k, consider one of the matrices in 1–3 above. In Section B.2, we expressed $\mathcal{A}^*$ and Δ as a linear combination of shapes. Thus, for fixed k, any matrix in 1–3 above can be written as linear combination of terms, where each term is a product of shapes, say

$$M_{\beta_0} \cdot M_{\beta_1} \cdots M_{\beta_k} \cdot M_{\beta_{k+1}}, \tag{16}$$

where each β_i is a (proper) shape from Section B.2. For each term as in (16), we further decompose it into a linear combination of several sub-terms represented by new (proper) shapes using Proposition 20. Let α_P denote a shape that arises as a subterm. We also define an scalar c_P associated with α_P that is used to form the its coefficient in the aforementioned linear combination. The shape α_P and scalar c_P are constructed by the following procedure:

5. Note that if, say, V_α has repeated vertices, then U_β must have the same number and type of repeated vertices. Otherwise, the dimensions of the two matrices are not compatible for multiplication.

1. We start with β_0. If β_0 is $\alpha_{A^*,1}$, $\alpha_{A^*,2}$, or $\alpha_{A^*,3}$, we call the circle vertex x to be the (initial) *loose end*. If $\beta_0 = \alpha_{1_n^T}$, we call the single vertex in this shape, which is a circle, the (initial) *loose end*. In all cases, we set $U_{\alpha_P} = U_{\beta_0}$ and $V_{\alpha_P} = V_{\beta_0}$. Note that $U_{\alpha_P} = V_{\alpha_P} = \emptyset$ if $\beta_0 = \alpha_{1_n^T}$.

2. We now do the following for each $j \in [k]$

   (a) Append the shape β_j by identifying $U_{\beta_j} = (u)$ with the current loose end and making $V_{\beta_j} = (v)$ the new loose end.

   (b) For each vertex in $\mathcal{V}(\beta_j) \setminus U_{\beta_j}$, either leave it alone or identify it with an existing vertex of the same type (circle or square) which is not u and has not yet been identified with a vertex in $\mathcal{V}(\beta_j) \setminus U_{\beta_j}$.

   (c) If this creates two parallel edges with integer labels N and M, we either (i) remove these parallel edges if $N + M$ is even or (ii) replace those parallel edges with a single labeled edge that has the same parity as $N + M$ and lies in $[N + M]$.

   In either case (i) or (ii), assign the edge (or empty edge) a coefficient according to the rule in Proposition 20.

3. Finally, apply the same procedure as described in 2(a–c) to β_{k+1}. Concretely, if the last term in the product is 1_n (so $\beta_{k+1} = \alpha_{1_n}$), we stop here. If the last term in the product is w (so $\beta_{k+1} = \alpha_w$), we append α_w to the existing shape by identifying the current loose end with $U_{\alpha_w} = (u)$. Then, for each set of resulting parallel edges, we remove or replace them as described in 2(c) and assign them coefficients.

4. Form the scalar c_P by multiplying together all coefficients of the labeled edges (including non-edges) that are output by the conversion procedure in Steps 2(c) and 3.

As we described, the matrices $\mathcal{A}^*(\Delta^k w)$, $\mathcal{A}^*(\Delta^k 1_n)$, and $1_n^T \Delta^k w$ are linear combinations of terms of the form α_P. We now describe the coefficients associated to a particular term α_P in this linear combination. To do so, we introduce the following definitions.

Definition 24 *For each shape β_j for $j \in [k]$, we define its **coefficient** $c(\beta_j)$ to be its coefficient in the graph matrix decomposition of Δ.*

Definition 25 *Let shapes $\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}$ be as described above.*

1. *An **identification pattern** P on $\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}$ specifies which vertices are identified with each other (according to Proposition 23) and which labelled edge is chosen when we convert parallel labeled edges into a single labeled edge (according to Proposition 20) as in step 2(c) above.*

2. *We define $\mathcal{P}_{\beta_0,\beta_1,\ldots,\beta_k,\beta_{k+1}}$ to be the set of all identification patterns on the shapes $\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}$.*

3. *Given an identification pattern P, we define α_P to be the shape resulting from P and we define c_P to be the coefficient, so that the resulting term is $c_P \cdot \prod_{j=1}^{k} c(\beta_j) \cdot M_{\alpha_P}$.*

In other words, c_P captures the part of the coefficient of M_{α_P} which comes from converting parallel labeled edges into a single labeled edge (or non-edge). Note that the (constant) coefficients coming from β_0 and β_{k+1} are also absorbed into c_P. In our argument it is particularly important to keep track of any non-edges that result from resolving two or more parallel labeled edges into a single labeled edge, which we make precise in the definition below.

Definition 26 (Vanishing edges) *Consider an identification pattern P on $\beta_0, \beta_1, \ldots, \beta_{k+1}$. The **vanishing edges** are the edges with label 0 (i.e., non-edges) that result from resolving parallel edges according to Proposition 20, as in step 2(c) above. Moreover, we say that a non-edge in α_P **vanishes** if it is in the set of vanishing edges.*

Next, we define a method for concisely summarizing certain information about a given identification pattern. Given $j \neq j'$, an identification pattern P, and a vertex $y \in \beta_j$, we say below that y *appears* in $\beta_{j'}$ if P identifies y with a vertex $y' \in \beta_{j'}$.

Definition 27 *For a given identification pattern P, we define a **decoration** $\tau : \cup_{j=0}^{k+1} \mathcal{V}(\beta_j) \to \{\emptyset, L, R, LR\}$ to summarize information about P in the following way. For each j and each vertex $y \in \beta_j$, define:*

$$\tau(y) = \begin{cases} L & \text{if } y \text{ appears in } \beta_{j'} \text{ for } j' < j \text{ and does not appear in any } \beta_{j''} \text{ for } j'' > j, \\ R & \text{if } y \text{ appears in } \beta_{j'} \text{ for } j' > j \text{ and does not appear in any } \beta_{j''} \text{ for } j'' < j, \\ LR & \text{if } y \text{ appears in } \beta_{j'} \text{ for } j' < j \text{ and also appears in some } \beta_{j''} \text{ for } j'' > j, \\ \emptyset & \text{otherwise}. \end{cases}$$

In particular, note that:

- *For any $j \geq 1$ and $y \in U_{\beta_j}$, y automatically appears in β_{j-1}, so $\tau(y) \in \{L, LR\}$.*
- *For any $j < k+1$ and $y \in V_{\beta_j}$, y automatically appears in β_{j+1}, so $\tau(y) \in \{R, LR\}$.*
- *For any $y \in \mathcal{V}(\beta_0)$, $\tau(y) \in \{\emptyset, R\}$.*
- *For any $y \in \mathcal{V}(\beta_{k+1})$, $\tau(y) \in \{\emptyset, L\}$.*

See Figures 16 and 17 for an example of an identification pattern and the associated decoration that can arise when multiplying shapes from the product $\mathcal{A}^*(\Delta w)$. With these definitions and for a fixed k, each of $\mathcal{A}^*(\Delta^k 1_n)$, $\mathcal{A}^*(\Delta^k w)$, and $1_n^T(\Delta^k)w$ can each be expressed as a summation of the following form:

$$\sum_{\beta_1,\ldots,\beta_k \in \{\alpha_i : i \in \mathcal{I}\}} \sum_{P \in \mathcal{P}_{\beta_0,\beta_1,\ldots,\beta_k,\beta_{k+1}}} \left(c_P \cdot \prod_{j=1}^{k} c(\beta_j) \right) M_{\alpha_P}. \tag{17}$$

To bound the norm of this expression, we apply the triangle inequality and bound separately $\|M_{\alpha_P}\|_{op}$ for each identification pattern P.

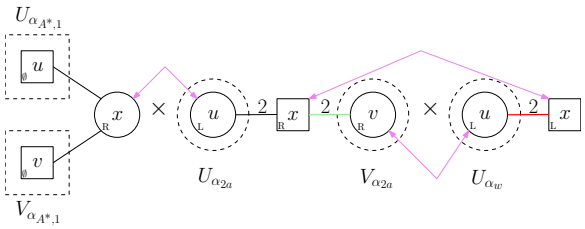

Figure 16: Example of an identification pattern for multiplying $M_{\alpha_{A^*,1}}$, $M_{\alpha_{2a}}$, and M_{α_w}. The violet arrows denote which vertices are identified with each other. As in Figure 12, if these vertices are identified with each other, the red and green edges become a multi-edge which gets converted to a linear combination of labeled edges using Proposition 20. The identification pattern that is consistent with the violet vertex identifications and which also picks edge label 0 to replace the multi-edge results in a shape that is depicted in Figure 17. The decorations of each vertex are written in the bottom left corner of each square or circle. The red edge is a right-critical edge and the green edge is a left-critical edge.

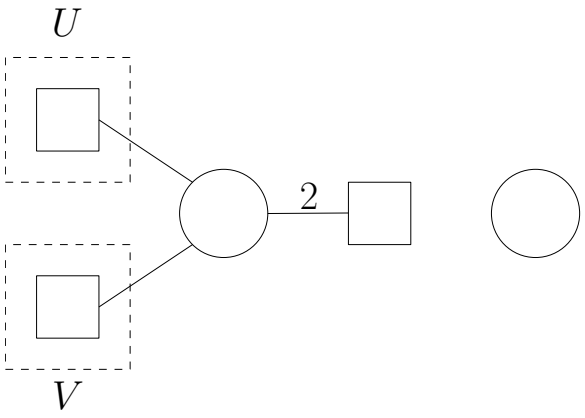

Figure 17: Resulting shape from the identification pattern in Figure 16. Observe that the minimum weight vertex separator can be taken to be one of the square vertices in U or V.

B.5.1. UPPER BOUNDING $\|M_{\alpha_P}\|_{op}$ VIA WEIGHTS

In order to upper bound $\|M_{\alpha_P}\|_{op}$, we will consider the contribution to $\|M_{\alpha_P}\|_{op}$ from each of the shapes $\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}$. In order to invoke the bound in Theorem 22, we must be able to calculate the min-vertex separator of α_P and status (i.e. square vs. circle and isolated vs. non-isolated) of each of the vertices of α_P. Specifically, define the following short-hand notation for the dominant term in the norm bound of Theorem 22:

$$\mathcal{B}(\alpha_P) = n^{\frac{\varphi(V(\alpha_P)) - \varphi(S_{\text{min}}) + \varphi(\text{Iso}(\alpha_P))}{2}}$$

where S_{min} is a min-vertex separator of α_P.

To compute the combinatorial quantities that define $\mathcal{B}(P)$, we will design an ideal weight function $w_{\text{ideal},P}$ and an actual weight function $w_{\text{actual},P}$, each of which assigns a *weight* to each of the shapes $\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}$. Intuitively, the ideal weight function allows us to accurately estimate the right-hand side of the norm bound in Theorem 22. However, as we explain later, determining this ideal weight function exactly is intractable. Instead, we show that the actual weight function is a faithful "relaxation" of the ideal weight function, tractable to calculate and still leads to sufficiently good norm bounds.

More precisely, the following properties must be satisfied:

1. $\mathcal{B}(\alpha_P) = \prod_{j=0}^{k+1} w_{\text{ideal},P}(\beta_j)$ where $\beta_0 = \beta_0$ and $\beta_{k+1} = \beta_{k+1}$. This ensures that the product of the ideal weights over the shapes β_j faithfully estimates the dominant term of the norm bound in Theorem 22.

2. $\prod_{j=0}^{k+1} w_{\text{actual},P}(\beta_j) \geq \prod_{j=0}^{k+1} w_{\text{ideal},P}(\beta_j)$. In fact, for almost all shapes β_j we will have that $w_{\text{actual},P}(\beta_j) \geq w_{\text{ideal},P}(\beta_j)$. This ensures that the actual weight function gives a norm bound that is valid (i.e. no smaller than the "true" norm bound given by the ideal weight function).

3. For all $j \in [k]$, $|w_{\text{actual},P}(\beta_j) c(\beta_j)| \leq d^{\frac{3}{2}} \sqrt[4]{n}$. This ensures that the norm bound given by the actual weight function is sufficiently small to complete the proofs of our technical lemmas.

We note that the number of possibilities for $\beta_1, \ldots, \beta_k$, the number of possible identification patterns on $\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}$, the maximum coefficient c_P for any identification pattern P, and the ratio $\frac{\|M_{\alpha_P}\|_{op}}{\mathcal{B}(\alpha_P)}$ (assuming the probabilistic norm bound in Theorem 22 holds) are all at most $(\log n)^{O(k)}$. This follows from the observation that we have $k+2$ shapes $\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}$, each consisting of $O(1)$ vertices and edges, and a simple combinatorial fact which we defer to the appendix (Proposition 42).

Thus, if we can specify weight functions satisfying the above properies, then we may bound the expression in Equation 17 as follows:

$$\begin{aligned}
&\sum_{\beta_1,\ldots,\beta_k \in \{\alpha_i : i \in \mathcal{I}\}} \sum_{P \in \mathcal{P}_{\beta_0,\beta_1,\ldots,\beta_k,\beta_{k+1}}} |c_P| \left(\prod_{j=1}^{k} |c(\beta_j)|\right) \|M_{\alpha_P}\|_{op} \\
&= \sum_{\beta_1,\ldots,\beta_k} \sum_{P \in \mathcal{P}_{\beta_0,\beta_1,\ldots,\beta_k,\beta_{k+1}}} |c_P| \left(\prod_{j=1}^{k} |c(\beta_j)|\right) \left(\prod_{j=0}^{k+1} w_{\text{ideal},P}(\beta_j)\right) \frac{\|M_{\alpha_P}\|_{op}}{\mathcal{B}(\alpha_P)} \\
&\leq (\log n)^{O(k)} \sum_{\beta_1,\ldots,\beta_k} \sum_{P \in \mathcal{P}_{\beta_0,\beta_1,\ldots,\beta_k,\beta_{k+1}}} |c_P| \left(\prod_{j=1}^{k} |c(\beta_j)|\right) \left(\prod_{j=0}^{k+1} w_{\text{ideal},P}(\beta_j)\right) \\
&\leq (\log n)^{O(k)} \sum_{\beta_1,\ldots,\beta_k} \sum_{P \in \mathcal{P}_{\beta_0,\beta_1,\ldots,\beta_k,\beta_{k+1}}} |c_P| \left(\prod_{j=1}^{k} |c(\beta_j)|\right) \left(\prod_{j=0}^{k+1} w_{\text{actual},P}(\beta_j)\right) \\
&\leq (\log n)^{O(k)} \left(d^{\frac{3}{2}} \sqrt[4]{n}\right)^k \max_P \{w_{\text{actual},P}(\beta_0) w_{\text{actual},P}(\beta_{k+1})\}. \qquad (18)
\end{aligned}$$

B.5.2. MIN-VERTEX SEPARATOR OF α_P

Before specifying the weight functions, we recall that:

$$\mathcal{B}(\alpha_P) = d^{\frac{|\mathcal{V}_\square(\alpha_P) \cap \mathcal{V}_\square(\alpha_P)| + |\text{Iso}(\alpha_P) \cap \mathcal{V}_\square(\alpha_P)| - |S_{\min} \cap \mathcal{V}_\square(\alpha_P)|}{2}} \times n^{\frac{|\mathcal{V}_\circ(\alpha_P)| + |\text{Iso}(\alpha_P) \cap \mathcal{V}_\circ(\alpha_P)| - |S_{\min} \cap \mathcal{V}_\circ(\alpha_P)|}{2}}$$

where $S_{\min}$ is a min-vertex separator of α_P. We now determine the min-vertex separator for α_P so that we can apply the bound in Theorem 22. When $\beta_0 = \alpha_{1_n^T}$, $U_{\alpha_P} = V_{\alpha_P} = \emptyset$ so $S_{\min} = \emptyset$. As we now show, when β_0 is $\alpha_{A^*,1}$, $\alpha_{A^*,2}$, or $\alpha_{A^*,3}$, the min-vertex separator of α_P consists of a single square. See Figure 17 for an example.

Lemma 28 *If β_0 is $\alpha_{A^*,1}$, $\alpha_{A^*,2}$, or $\alpha_{A^*,3}$ then the min-vertex separator of α_P consists of a single square.*

Proof Since U_{α_P} consists of a single square and is a vertex separator, the minimum weight vertex separator is either a single square or no vertices at all. To show that the minimum weight vertex separator has at least one vertex, we prove that U_{α_P} must be connected to V_{α_P}.

To prove this, it is sufficient to prove the following lemma. Here by 'degree' of a vertex a in a shape, we mean the sum of all edge labels of edges incident to a.

Lemma 29 *If β_0 is $\alpha_{A^*,1}$, $\alpha_{A^*,2}$, or $\alpha_{A^*,3}$ then either $U_{\alpha_P} = V_{\alpha_P} = (u)$ where u is a square vertex, or $U_{\alpha_P} = (u)$, $V_{\alpha_P} = (v)$ where u and v are distinct square vertices and u and v are the only vertices with odd degree.*

With this lemma, the result follows easily. If $U_{\alpha_P} = V_{\alpha_P} = (u)$ then the result is trivial. If $U_{\alpha_P} = (u)$ and $V_{\alpha_P} = (v)$ where u and v are distinct square vertices and are the only vertices with odd degree then u and v must be in the same connected component of α_P due to the following fact.

Proposition 30 *For any undirected graph G with integer edge-labels, for any connected component C of G, $\sum_{v \in C} deg(v)$ is even.*

Proof This is the handshaking lemma and can be proved by observing that $\sum_{v \in C} deg(v) = 2|E(C)|$ which is even. ■

To be concrete, suppose that for the sake of contradiction u and v lie in distinct connected components C_u and C_v of α_P. Then the sum of degrees in C_u is odd by Lemma 29, which contradicts Proposition 30.

We proceed to prove Lemma 29.

Proof [Proof of Lemma 29] Let $U_{\alpha_P} = (u)$ and $V_{\alpha_P} = (v)$. We make the following observations about the process for building the shape α_P

1. In β_0, either $u = v$ (in which case u has even degree) or u and v are distinct square vertices which have odd degree. The circle vertex in β_0 always has even degree.
2. For all of the shapes $\beta_1, \ldots, \beta_k, \beta_{k+1}$, all of the vertices have even degree.
3. Whenever two vertices are identified, the parity of the degree of the resulting vertex is equal to the parity of the sum of the degrees of the original vertices.
4. No other operation affects the parities of the degrees of the vertices.

Together, these observations imply that either $u = v$ or u and v are the only vertices with odd degree. ■

The proof of Lemma 28 is now complete. ■

Corollary 31 *If β_0 is $\alpha_{A^*,1}$, $\alpha_{A^*,2}$, or $\alpha_{A^*,3}$ then*

$$\mathcal{B}(\alpha_P) = d^{\frac{|\mathcal{V}_\square(\alpha_P)|+|\mathrm{Iso}(\alpha_P)\cap\mathcal{V}_\square(\alpha_P)|-1}{2}} \cdot n^{\frac{|\mathcal{V}_\circ(\alpha_P)|+|\mathrm{Iso}(\alpha_P)\cap\mathcal{V}_\circ(\alpha_P)|}{2}}.$$

If β_0 is $\alpha_{1^T_n}$ then

$$\mathcal{B}(\alpha_P) = d^{\frac{|V_\square(\alpha_P)|+|\mathrm{Iso}(\alpha_P)\cap\mathcal{V}_\square(\alpha_P)|}{2}} \cdot n^{\frac{|V_\circ(\alpha_P)|+|\mathrm{Iso}(\alpha_P)\cap\mathcal{V}_\circ(\alpha_P)|}{2}}.$$

We remark that the factor of $\frac{1}{\sqrt{d}}$ appearing in the first expression turns out to be crucial to obtaining satisfactory norm bounds.

B.5.3. THE IDEAL WEIGHT FUNCTION

Given that we have identified the min-vertex separator of the shape α_P, we now specify an ideal weight function $w_{\text{ideal},P}$ such that $\mathcal{B}(\alpha_P) = \prod_{j=0}^{k+1} w_{\text{ideal},P}(\beta_j)$. First we introduce and formalize some intuitive terminology. We say that $v \in \mathcal{V}(\alpha_P)$ *appears* in a shape β_j (or that β_j *contains* v) if v is the result of identifying one or more vertices according to the identification pattern P, at least one of which lies in β_j. We order the indices $j_1 < \ldots < j_r$ (which we refer to as discrete *times*) of the shapes in $\beta_{j_1}, \ldots, \beta_{j_r}$ where v appears. We refer to j_1 as the *first time* v appears and j_r as the *last time* v appears.

Each vertex $v \in \mathcal{V}(\alpha_P)$ has an associated value coming from the expression in Corollary 31. This value, which we call $w_{\text{ideal},P}(v)$, is as follows:

$$w_{\text{ideal},P}(v) = \begin{cases} \sqrt{d} & \text{if } v \text{ is a square and not isolated,} \\ d & \text{if } v \text{ is a square and isolated}^{6}, \\ \sqrt{n} & \text{if } v \text{ is a circle and not isolated,} \\ n & \text{if } v \text{ is a circle and isolated.} \end{cases}$$

We "split" this value among (at most) 2 shapes that contain v by assigning the square root of the value of v to each of β_j and $\beta_{j'}$, where j is the smallest index for which v appears in β_j and j' is the largest index for which v appears in $\beta_{j'}$. If v only appears once, then we assign the full value of v to the shape in which it appears. Formally, we define this procedure in the following way. For $j \in [k+1]$, we let $w_{\text{ideal},P}(\beta_j)$ be the total weight which is assigned to β_j. This weight is

$$w_{\text{ideal},P}(\beta_j) = \prod_{v \in \mathcal{V}(\beta_j)} w_{\text{ideal},P}(v)^{1-\frac{1}{2}\mathbb{1}\{\tau(v)\in\{L,LR\}\}-\frac{1}{2}\mathbb{1}\{\tau(v)\in\{R,LR\}\}}.$$

For β_0, we adjust this to take the minimum weight vertex separator into account. In particular, if $\beta_0 = \alpha_{1_n^T}$ then:

$$w_{\text{ideal},P}(\beta_0) = \prod_{v \in \mathcal{V}(\beta_0)} w_{\text{ideal},P}(v)^{1-\frac{1}{2}\mathbb{1}\{\tau(v)=R\}},$$

while if β_0 is $\alpha_{A^*,1}$, $\alpha_{A^*,2}$, or $\alpha_{A^*,3}$ then

$$w_{\text{ideal},P}(\beta_0) = \frac{1}{\sqrt{d}} \prod_{v \in \mathcal{V}(\beta_0)} w_{\text{ideal},P}(v)^{1-\frac{1}{2}\mathbb{1}\{\tau(v)=R\}}.$$

From these definitions, we may immediately conclude that $\mathcal{B}(\alpha_P) = \prod_{j=0}^{k+1} w_{\text{ideal},P}(\beta_j)$ (i.e. Property 1 is satisfied). See the first row of Table 3 for an example of how w_{ideal} is computed for a particular shape and identification pattern arising in $\mathcal{A}^*(\Delta w)$.

B.5.4. THE LOCAL WEIGHT FUNCTION

While the ideal weight function yields the correct norm bound, it cannot be computed separately for each shape β_j because in order to determine if a vertex in β_j is isolated or not, we need to consider the entire identification pattern P. To handle this, we introduce a different weight function $w_{\text{local},P}$ which can be computed separately for each shape β_j by considering only the "local data" consisting of the decorations on vertices in $\mathcal{V}(\beta_j)$. To define $w_{\text{local},P}(v)$, for each j and each $v \in \mathcal{V}(\beta_j)$, we upper bound $w_{\text{ideal},P}(v)$ based on the local data of β_j. In particular, if a vertex is incident to an edge which cannot vanish based on the local data at β_j, then we know it cannot be isolated and $w_{\text{ideal},P}(v)$ is $\sqrt{d}$ or $\sqrt{n}$. For $j = 0$, we also know $w_{\text{ideal},P}(v) = \sqrt{d}$ for the vertices $v \in U_{\alpha_P} \cup V_{\alpha_P}$ since we never consider vertices in $U_{\alpha_P} \cup V_{\alpha_P}$ to be isolated. For other vertices, we conservatively upper bound $w_{\text{ideal},P}(v)$ by d or n. We introduce the following definitions in order to formally define $w_{\text{local},P}$.

6. Note that vertices in $U_{\alpha_P} \cup V_{\alpha_P}$ do not count as isolated.

Definition 32 *We say that an edge $e = \{u, v\}_l$ is **safe** for shape β_j if both of the following hold:*

1. *Either $\tau(u) \in \{\emptyset, R\}$ or $\tau(v) \in \{\emptyset, R\}$.*
2. *Either $\tau(u) \in \{\emptyset, L\}$ or $\tau(v) \in \{\emptyset, L\}$.*

Observe that a safe edge cannot vanish (i.e. it appears in α_P with a positive edge label). For every j and $v \in V(\beta_j)$, we define the "full" local weight as:

$$b_j(v) = \begin{cases} \sqrt{d} & \text{if } v \text{ is a square and incident to a safe edge for } \beta_j, \\ \sqrt{d} & j = 0 \text{ and } v \in U_{\alpha_P} \cup V_{\alpha_P}, \\ d & \text{if } v \text{ is any other square,} \\ \sqrt{n} & \text{if } v \text{ is a circle and incident to a safe edge for } \beta_j, \\ n & \text{if } v \text{ is any other circle.} \end{cases}$$

As before, if a vertex is identified with other vertices, then we "split" the full weight between the first and last times it appears. So we define, for every j and $v \in V(\beta_j)$, the "split" local weight as:

$$\overline{b}_j(v) = b_j(v)^{1 - \frac{1}{2}\mathbb{1}\{\tau(v) \in \{L, LR\}\} - \frac{1}{2}\mathbb{1}\{\tau(v) \in \{R, LR\}\}}. \tag{19}$$

Recall that for $1 \leq j \leq k$, $w_{\text{local},P}(\beta_j)$ is the product of the $\overline{b}_j(y)$'s, see (20). Using these definitions, we define the weight function $w_{\text{local},P}$ as:

$$w_{\text{local},P}(\beta_j) = \left(\frac{1}{\sqrt{d}}\right)^{\mathcal{I}(\beta_j)} \cdot \prod_{v \in \mathcal{V}(\beta_j)} \overline{b}_j(v), \tag{20}$$

where

$$\mathcal{I}(\beta_j) = \mathbb{1}\left\{j = 0 \text{ and } \beta_0 \in \{\alpha_{A^*,1}, \alpha_{A^*,2}, \alpha_{A^*,3}\}\right\}.$$

In other words, we divide by $\sqrt{d}$ for $j = 0$ if β_0 is $\alpha_{A^*,1}$, $\alpha_{A^*,2}$, or $\alpha_{A^*,3}$. We may immediately conclude from these definitions that property 2 is satisfied; in the next section we verify Property 3. We record for future use a simple upper bound on the weights of vertices.

Proposition 33 *Let $j \in \{0, \ldots, k+1\}$. If $v \in \mathcal{V}_\circ(\beta_j)$, then the contribution of v to $w_{local,P}(\beta_j)$ is at most $\sqrt{n}$. If $v \in \mathcal{V}_\square(\beta_j)$, then the contribution of v to $w_{local,P}(\beta_j)$ is at most $\sqrt{d}$. The same results also apply to $w_{ideal,P}$.*

Proof Consider the case $v \in \mathcal{V}_\circ(\beta_j)$. If v is not identified with any other vertex, then it cannot be isolated. Hence, it contributes no more than $\sqrt{n}$ to $w_{\text{local},P}(\beta_j)$. On the other hand, if v is identified with some other vertex, then the maximum possible weight of n is split between v and some other vertex. Thus, it contributes no more than $\sqrt{n}$ to $w_{\text{local},P}(\beta_j)$. ■

See the second row of Table 3 for an example of how w_{local} is computed for a particular shape and identification pattern arising in $\mathcal{A}^*(\Delta w)$.

B.6. Proof of Lemma 12

In this section, we show that the local weight function $w_{\text{local},P}$ from the previous section is sufficient to complete the proof of Lemma 12. As mentioned earlier, by definition of $w_{\text{local},P}$, it satisfies Property 2. It remains to verify the following:

1. If β_0 is any of $\alpha_{A^*,1}, \alpha_{A^*,2}, \alpha_{A^*,3}$ and $\beta_{k+1} = \alpha_{1_n}$, then $w_{\text{local},P}(\beta_0) w_{\text{local},P}(\beta_{k+1}) \le \max(\sqrt{d} n^{3/4}, n)$.
2. For all $j \in [k]$, $|w_{\text{local},P}(\beta_j) c(\beta_j)| \le d^{\frac{3}{2}} \sqrt[4]{n}$ (corresponding to Property 3).

Given these and Equation (18), we have the following bound with probability $1 - n^{-\Omega(1)}$:

$$\begin{aligned}\left\|\mathcal{A}^*(\Delta^k 1_n)\right\|_{op} &\le (\log n)^{O(k)} \cdot \left(d^{\frac{3}{2}} \sqrt[4]{n}\right)^k \cdot \max_{P \in \mathcal{P}_{\beta_0,\beta_1,\ldots,\beta_k,\beta_{k+1}}} |c_P| \cdot w_{\text{local},P}(\beta_0) \cdot w_{\text{local},P}(\beta_{k+1}) \\ &\le (\log n)^{O(k)} \cdot \left(d^{\frac{3}{2}} \sqrt[4]{n}\right)^k \cdot \max(\sqrt{d} n^{3/4}, n),\end{aligned}$$

which completes the proof of Lemma 12.

We now verify the two conditions on the local weight function. For the first condition, note that $w_{\text{local},P}(\beta_{k+1}) = w_{\text{local},P}(\alpha_{1_n}) = \sqrt{n}$. To handle the contribution from β_0, we enumerate the following cases:

- **Case $\beta_0 = \alpha_{A^*,1}$:** We know that the two squares u, v are not in $\text{Iso}(\alpha_P)$, so $b_0(u) = b_0(v) = \sqrt{d}$. Suppose at least one of the two edges in $\alpha_{A^*,1}$ are safe. Then, $b_0(x) = \sqrt{n}$ and $w_{\text{local},P}(\alpha_{A^*,1}) \le \frac{1}{\sqrt{d}} b_0(u) b_0(v) b_0(x)^{1/2} = \sqrt{d} n^{1/4}$. Otherwise, $\tau(u) = \tau(v) = R$, so
$$w_{\text{local},P}(\alpha_{A^*,1}) = \frac{1}{\sqrt{d}} b_0(u)^{1/2} b_0(v)^{1/2} b_0(x)^{1/2} \le \sqrt{n}.$$
- **Case $\beta_0 = \alpha_{A^*,2}$ or $\alpha_{A^*,3}$:** Again, we know $b_0(u) = \sqrt{d}$. So, $w_{\text{local},P}(\alpha_{A^*,2}) = \frac{1}{\sqrt{d}} b_0(u) b_0(x)^{1/2} \le \sqrt{n}$.

This completes the proof that $w_{\text{local},P}(\beta_0) w_{\text{local},P}(\beta_{k+1}) \le \max(\sqrt{d} n^{3/4}, n)$.

For the second condition, we fix $j \in [k]$. Below, for each shape α_i for $i \in \mathcal{I}$, we consider all possibilities of the decorations of the vertices of α_i, reducing the number of cases when possible by symmetry. The tables below handle the essential cases. In the tables below, we use the term 'Any' to denote that a vertex may have any of the decorations described above. We now analyze the possible cases for β_j, repeatedly making use of Proposition 33 when appropriate:

- **Case $\beta_j = \alpha_1$:** A priori, there are a total of $2 \cdot 4 \cdot 4 \cdot 2 = 64$ cases since $\tau(u) \in \{L, LR\}, \tau(x_1), \tau(x_2) \in \{\emptyset, L, R, LR\}$, and $\tau(v) \in \{R, LR\}$. By symmetry, each case reduces to one considered in Table 1. Note that the first row of Table 1 stands for 32 different cases. For this row, we slightly abuse notation and use $\bar{b}_j(y)$ to specify an upper bound on the split local weight of $y \in \{u, x_1, x_2, v\} \subset V(\alpha_1)$ for all of these 32 cases.

  By inspection of Table 1 and using that $n = d^2/\operatorname{polylog}(d)$ (see Remark 3), we conclude that
$$w_{\text{local},P}(\beta_j) = w_{\text{local},P}(\alpha_1) \le d\sqrt{n}. \tag{21}$$
  Also note that $c(\beta_j) = c(\alpha_1) = O(1)$.

- **Case $\beta_j = \alpha_{2a}$:** A priori, there are a total of $2 \cdot 4 \cdot 2 = 16$ cases since $\tau(u) \in \{L, LR\}, \tau(x) \in \{\emptyset, L, R, LR\}$, and $\tau(v) \in \{R, LR\}$. By symmetry, each case reduces to one considered in Table 2. Note that the first row of Table 2 stands for 4 different cases. For this row, we slightly abuse notation and use $\bar{b}_j(y)$ to specify an upper bound on the split local weight of $y \in \{u, x, v\} \subset V(\alpha_{2a})$ for all of these 4 cases.

  By inspection of Table 2 and using that $n = d^2/\operatorname{polylog}(d)$ (see Remark 3), we conclude that

$$w_{\text{local},P}(\beta_j) = w_{\text{local},P}(\alpha_{2a}) \leq n. \tag{22}$$

  Also note that $c(\beta_j) = c(\alpha_{2a}) = O(1)$.

- **Case $\beta_j = \alpha_{3a}$:** First, note that $w_{\text{local},P}(\beta_j) = b_j(u)^0 \bar{b}_j(x_1)\bar{b}_j(x_2) = \bar{b}_j(x_1)\bar{b}_j(x_2)$ because u is identified with vertices to the left and right of β_j. It is straightforward to enumerate the possible decorations of x_1, x_2 and verify that $\bar{b}_j(x_1)\bar{b}_j(x_2) \leq d$. Thus, $w_{\text{local},P}(\alpha_{3a}) \leq d$. Also, recall that $|c(\beta_j)| = |c(\alpha_{3a})| = O(1)$.

- **Case $\beta_j = \alpha_{3b}$:** Recall that α_{3b} has $U_{\alpha_{3b}} = V_{\alpha_{3b}} = (u)$. By the rules for graph matrix multiplication (see Section B.4), $u \in \alpha_{3b}$ is identified with a circle vertex in β_{j-1} and β_{j+1}. Hence $\tau(u) = LR$. By (19), we have $\bar{b}_j(u) = 1$. Moreover, $\bar{b}_j(x) \leq \sqrt{d}$. Therefore,

$$w_{\text{local},P}(\beta_j) = w_{\text{local},P}(\alpha_{3b}) \leq \sqrt{d}. \tag{23}$$

  Also recall that $|c(\beta_j)| = |c(\alpha_{3b})| = O(d)$.

- **Case $\beta_j = \alpha_4$:** The proof is very similar to the one for α_{3b}. Since $U_{\alpha_4} = V_{\alpha_4} = (u)$, we see that $\tau(u) = LR$, so $\bar{b}_j(u) = 1$. And automatically, $\bar{b}_j(x) \leq \sqrt{d}$. Hence

$$w_{\text{local},P}(\beta_j) = w_{\text{local},P}(\alpha_4) \leq \sqrt{d}. \tag{24}$$

  Also recall that $|c(\beta_j)| = |c(\alpha_4)| = O(1)$.

$\tau(u)$	$\tau(x_1)$	$\tau(x_2)$	$\tau(v)$	$\bar{b}_j(u)$	$\bar{b}_j(x_1)$	$\bar{b}_j(x_2)$	$\bar{b}_j(v)$	Product
LR	Any	Any	Any	1	$d^{1/2}$	$d^{1/2}$	$n^{1/2}$	$n^{1/2}d$
L	$\emptyset$	$\emptyset$	R	$n^{1/4}$	$d^{1/2}$	$d^{1/2}$	$n^{1/4}$	$n^{1/2}d$
L	$\emptyset$	L	R	$n^{1/4}$	$d^{1/2}$	$d^{1/4}$	$n^{1/4}$	$n^{1/2}d^{3/4}$
L	$\emptyset$	LR	R	$n^{1/4}$	$d^{1/2}$	1	$n^{1/4}$	$n^{1/2}d^{1/2}$
L	L	L	R	$n^{1/2}$	$d^{1/4}$	$d^{1/4}$	$n^{1/4}$	$n^{3/4}d^{1/2}$
L	L	R	R	$n^{1/4}$	$d^{1/4}$	$d^{1/4}$	$n^{1/4}$	$n^{1/2}d^{1/2}$
L	L	LR	R	$n^{1/2}$	$d^{1/4}$	1	$n^{1/4}$	$n^{3/4}d^{1/4}$
L	LR	LR	R	$n^{1/2}$	1	1	$n^{1/2}$	n

Table 1: Case work for α_1.

$\tau(u)$	$\tau(x)$	$\tau(v)$	$\bar{b}_j(u)$	$\bar{b}_j(x)$	$\bar{b}_j(v)$	Product
LR	Any	LR	1	$\sqrt{d}$	1	$\sqrt{d}$
LR	R	R	1	$d^{1/2}$	$n^{1/2}$	$n^{1/2}d^{1/2}$
LR	$\emptyset$	R	1	$d^{1/2}$	$n^{1/4}$	$n^{1/4}d^{1/2}$
LR	L	R	1	$d^{1/4}$	$n^{1/4}$	$n^{1/4}d^{1/4}$
L	$\emptyset$	R	$n^{1/4}$	$d^{1/2}$	$n^{1/4}$	$n^{1/2}d^{1/2}$
L	L	R	$n^{1/2}$	$d^{1/4}$	$n^{1/4}$	$n^{3/4}d^{1/4}$
L	LR	R	$n^{1/2}$	1	$n^{1/2}$	n

Table 2: Case work for α_{2a}.

B.7. Modifying the local weighting scheme

Unfortunately, while the local weight function is sufficient for proving Lemma 12, the bound it gives is too loose for terms arising in $\mathcal{A}^*(\Delta^k w)$ (for Lemma 14) and $1^T\Delta^k w$ (for Lemma 16). Specifically, it is too conservative when assigning weight to the vertices in α_w in the case that its single edge vanishes with respect to an identification pattern P (see Definition 26). To handle this bad case, we define a modified weight function $w_{\text{actual},P}$, for any given identification pattern P, by decreasing the weight on the square vertex of α_w when its edge vanishes. While this guarantees that $w_{\text{actual}}(\beta_{k+1})$ is small (making it possible to satisfy 3), previous arguments do not immediately imply that 2 holds in the case that the edge $\{u, x_1\}_2$ of α_w vanishes. We will show this is compensated for by an increase in the weight on squares in other shapes in a way that ensures 2 and 3 are satisfied simultaneously. To carry out this strategy, we introduce the notion of *critical edges*; see also Figure 16 for an example.

Definition 34 *We define the following two types of edges to be* ***right-critical edges****:*

1. *If the square vertex x_1 in α_w satisfies $\tau(x_1) = L$ then the edge in α_w is a right-critical edge.*
2. *If $\beta_j = \alpha_{2a}$, the circle vertex u in U_{β_j} satisfies $\tau(u) = L$, the circle vertex v in V_{β_j} satisfies $\tau(v) = R$, and the square vertex x satisfies $\tau(x) = LR$, then the edge $\{u, x\}_2$ in $\beta_j = \alpha_{2a}$ is a right-critical edge.*

Definition 35 *We define a* ***left-critical edge*** *of β_j to be an edge $e = \{u, v\}$ such that one of the following two cases holds:*

1. *$l_e = 2$, $\tau(u) \in \{R, LR\}$, and $\tau(v) \in \{R, LR\}$.*
2. *$l_e = 1$, $\tau(u) = \tau(v) = LR$.*

With these definitions in hand, our high-level strategy is as follows. If the right-critical edge in $\beta_{k+1} = \alpha_w$ does not vanish, then the proof strategy of Lemma 12 that employs the local weight scheme $w_{\text{local},P}$ of Section B.5.4 suffices to directly yield the bounds of Lemmas 14 and 16. If instead the right-critical edge e in $\beta_{k+1} = \alpha_w$ vanishes, we use Lemma 36 to pair β_j with a shape $\beta_{j'}$ that contains a left-critical edge. We adjust the weights of $\beta_{j'}$ and α_w directly according to the *actual* weight scheme defined in Section B.8 in order to satisfy 2 and 3. On the remaining shapes β_j where $j \in [k]\backslash\{j'\}$, we employ the local weight scheme of Section B.5.4 (i.e. the actual weights correspond with the local weights on these remaining shapes). Multiplying together the actual weights for all shapes then yields the bounds of Lemmas 14 and 16.

Lemma 36 *Given an identification pattern P on $\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1}$, if there is a $j \in [k+1]$ such that β_j has a vanishing right-critical edge e then there is a $j' < j$ such that $\beta_{j'}$ has no vanishing right-critical edge and has a left-critical edge e' whose square endpoint is identified with the square endpoint of e.*

Proof We prove this lemma by induction on j. Assume the result is true for $j = m$ and assume that β_j has a vanishing right-critical edge where $j = m + 1$. Recall that we say a vertex v of α_P *appears* in β_l if it is the result of identifying several vertices according to P, one of which lies in β_l. We say that an edge e of α_P *appears* in a shape β_l if both of its endpoints appear in β_l. Suppose that β_j contains a right-critical edge e. We claim that e does not appear in $\beta_{j'}$ for $j' > j$. If $\beta_j = \alpha_w$, this claim follows immediately because then $j = k + 1$. Now suppose that e is the second type of right-critical edge in Definition 34, in which case we have $\beta_j = \alpha_{2a}$. Since $\tau(u) = L$, it also holds in this case that e does not appear in $j' > j$.

Let j' denote the largest index such that $j' < j$ and e appears in $\beta_{j'}$ (such an edge must exist as otherwise e cannot vanish). We claim that e must be a left-critical edge in $\beta_{j'}$. To see this, observe that e appears in β_j where $j > j'$. If $l_e = 2$, then e is automatically a left-critical edge. If $l_e = 1$, then e must also appear in $\beta_{j''}$ for some $j'' < j'$ as otherwise e cannot vanish (two parallel edges with labels 1 and 2 give rise to a term with label 1 and a term with label 3, so they do not vanish). Thus, e is a left-critical edge in this case as well.

If $\beta_{j'}$ does not have a vanishing right-critical edge, then we are done. If $\beta_{j'}$ does have a vanishing right critical edge (in which case it must be α_{2a}) then by the inductive hypothesis there is a $j'' < j'$ which has a left-critical edge but does not have a vanishing right-critical edge, as needed.

■

B.8. Formal definition of w_{actual}

We now give a formal definition of w_{actual}. Let u_w, x_{extra} denote the vertices corresponding to u, x, respectively, in α_w. If $\beta_{k+1} = \alpha_w$, then define the *per-vertex actual weights* for β_{k+1} as follows:

1. If the edge $\{u_w, x_{\text{extra}}\}_2$ in α_w does not vanish, then set $w_{\text{actual}}(u_w) = \sqrt[4]{n}$ and $w_{\text{actual}}(x_{\text{extra}}) = \sqrt{d}$. Furthermore, set w_{actual} equal to w_{local} for all other vertices and shapes.

2. If the edge $\{u_w, x_{\text{extra}}\}_2$ in α_w vanishes, then set $w_{\text{actual},P}(u_w) = \sqrt{n}$ and $w_{\text{actual},P}(x_{\text{extra}}) = \frac{\sqrt{d}}{\sqrt[4]{n}}$. For the remaining shapes, define w_{actual} as below.

In the second case above, we modify w_{local} further to define w_{actual}. Let $j < k + 1$ be such that β_j has a left-critical edge and no vanishing right-critical edge (whose existence is guaranteed by Lemma 36). Note that β_j must be one of $\alpha_{A^*,3}, \alpha_1, \alpha_{2a}, \alpha_{3a}, \alpha_{3b}$ or α_4 by definition of left- and right-critical edges. We set w_{actual} to be equal to w_{local} on all shapes β_l for $l \neq k + 1, j$. To compensate for the reduction in weight on the shape β_{k+1} in the case that its edge vanishes, we define $w_{\text{actual}}(\beta_j)$ in the following way.

For $l \in \{j, k + 1\}$ we set

$$w_{\text{actual},P}(\beta_l) = \prod_{v \in \beta_l} w_{\text{actual},P}(v) \tag{25}$$

where $w_{\text{actual},P}(v)$ are *per-vertex* actual weights. If $l = k+1$, the per-vertex actual weights are defined in 1 and 2 above. Note that this ensures $w_{\text{actual},P}(\beta_{k+1}) \le \sqrt{d}\sqrt[4]{n}$ and that $w_{\text{actual},P}(u_w) \ge w_{\text{ideal},P}(u_w)$. If $l = j$, the per-vertex actual weights are defined below according to the cases of β_j.

- **Case $\beta_j = \beta_0 = \alpha_{A^*,3}$:** Define $w_{\text{actual},P}(u) = \sqrt[4]{\frac{n}{d}} \cdot w_{\text{ideal},P}(u) = \sqrt[4]{n}$. Here, $w_{\text{ideal},P}(u) = \sqrt[4]{d}$ follows from the fact that $u \in U_{\beta_0} = V_{\beta_0}$, so it is not a middle vertex and thus cannot be in $\text{Iso}(\alpha_P)$.
- **Case $\beta_j = \alpha_1$:** By symmetry, it suffices to consider the case where the edge $\{u_w, x_{\text{extra}}\}_2$ of α_w is identified with the edge $\{u, x_1\}$ of α_1. We now define $w_{\text{actual},P}(u) = 1$, $w_{\text{actual},P}(x_1) = \sqrt[4]{n}$, $w_{\text{actual},P}(x_2) = \sqrt{d}$, $w_{\text{actual},P}(v) = \sqrt{n}$.
- **Case $\beta_j = \alpha_{2a}$:** We divide the definition for this case into two sub-cases, based on whether or not the edge $\{u, x\}_2$ in α_{2a} vanishes.
  - **Sub-case $\{u, x\}_2$ does not vanish:** We define $w_{\text{actual},P}(u) = n^{1/4}$, $w_{\text{actual},P}(x) = n^{1/4}$, $w_{\text{actual},P}(v) = n^{1/2}$.
  - **Sub-case $\{u, x\}_2$ vanishes:** We define $w_{\text{actual},P}(u) = 1$, $w_{\text{actual},P}(x) = d$, $w_{\text{actual},P}(v) = \sqrt{n}$.
- **Case $\beta_j = \alpha_{3a}$:** We define $w_{\text{actual},P}(x_1) = \sqrt{d}\sqrt[4]{n}$ and $w_{\text{actual},P}(x_2) = \sqrt{d}\sqrt[4]{n}$.
- **Case $\beta_j = \alpha_{3b}$ or α_4:** We define $w_{\text{actual},P}(x) = \sqrt{d}\sqrt[4]{n}$.

See Table 3 for an example of how w_{actual} is computed for a particular shape and identification pattern arising in $\mathcal{A}^*(\Delta w)$. This example also demonstrates a shape and identification pattern for which w_{local} is too conservative and overestimates w_{ideal} (which corresponds to the "correct" norm bound), yet w_{actual} corrects this issue.

	Shapes							
	$\alpha_{A^*,1}$			α_{2a}			α_w	
	u	v	x	u	x	v	u	x
Ideal	$\sqrt{d}$	$\sqrt{d}$	$\sqrt[4]{n}$	$\sqrt[4]{n}$	$\sqrt[4]{d}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt[4]{d}$
Local	$\sqrt{d}$	$\sqrt{d}$	$\sqrt[4]{n}$	$\sqrt[4]{n}$	$\sqrt[4]{d}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{d}$
Actual	$\sqrt{d}$	$\sqrt{d}$	$\sqrt[4]{n}$	$\sqrt[4]{n}$	$\sqrt[4]{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{d}/\sqrt[4]{n}$

Table 3: Comparison of the three different weighting schemes applied to the shape and identification pattern from Figures 16 and 17. Each column indicates the weight contributions of a vertex to the total weight of the shape that contains it, for each of the three weighting schemes. So, the first row corresponds to weights under w_{ideal}, but which are "split" if a vertex is identified with other vertices. The second row corresponds to the values $\bar{b}_j(\cdot)$ from 19. The third row is the same as the second, but adjusted according to the definition of w_{actual} in Section B.8. The red entry indicates that w_{local} assigns more weight than the "true" weight as in w_{ideal}; this leads to an over-estimate of the true norm bound by a $\sqrt[4]{d}$ factor. The green entries indicate the weights that are modified so that w_{actual} gives the correct norm bound.

B.9. Paying for the extra square: completing the proofs of Lemmas 14 and 16

To prove Lemmas 14 and 16, we follow the proof of Lemma 12, but use w_{actual} in place of w_{local} and the following crucial lemma:

Lemma 37 *Suppose that the right-critical edge e in α_w vanishes. Let $0 \leq j < k+1$ be such that β_j has no vanishing right-critical edge and has a left-critical edge e' whose square endpoint is identified with the square endpoint of e (whose existence is guaranteed by Lemma 36). Then,*

$$w_{\textit{actual},P}(\beta_j) w_{\textit{actual},P}(\beta_{k+1}) \geq w_{\textit{ideal},P}(\beta_j) w_{\textit{ideal},P}(\beta_{k+1}). \tag{26}$$

Moreover, it holds that $c(\beta_j) w_{\textit{actual},P}(\beta_j) \leq O(d^{3/2}\sqrt[4]{n})$ if $j > 0$, and $c(\beta_j) w_{\textit{actual},P}(\beta_j) \leq O(n^{3/4}/\sqrt{d})$ if $j = 0$.

Let $j < k+1$ be the special index as in Lemma 37. Then, as in the proof of Lemma 12, the proof of Lemma 14 will be complete provided we can show the following:

1. If β_0 is any of $\alpha_{A^*,1}, \alpha_{A^*,2}, \alpha_{A^*,3}$ and $\beta_{k+1} = \alpha_w$, then $w_{\text{actual},P}(\beta_0) w_{\text{actual},P}(\beta_{k+1}) \leq \max(\sqrt{d} n^{3/4}, n)$.
2. For all $l \in [k] \setminus \{j\}$, $|w_{\text{actual},P}(\beta_l) c(\beta_l)| \leq d\sqrt{n}$.

The second condition above follows in exactly the same way as in the proof of Lemma 12, since $w_{\text{actual},P}(\beta_l) = w_{\text{local},P}(\beta_l)$ for $l \in [k] \setminus \{j\}$. To verify the first condition, we enumerate two cases:

- **Case $j \neq 0$:** If $j \neq 0$, then $w_{\text{actual},P}(\beta_0) = w_{\text{local},P}(\beta_0) \leq \sqrt{n}$ (from the proof of Lemma 12) and $w_{\text{actual},P}(\beta_{k+1}) \leq \sqrt{d}\sqrt[4]{n}$ by definition. So, we immediately have $w_{\text{actual},P}(\beta_0) w_{\text{actual},P}(\beta_{k+1}) \leq \sqrt{d} n^{3/4}$.
- **Case $j = 0$:** If $j = 0$, then $w_{\text{actual},P}(\beta_0) w_{\text{actual},P}(\beta_{k+1}) \leq (n^{3/4}/\sqrt{d}) \cdot \sqrt{d}\sqrt[4]{n} = n$, by Lemma 37 and definition of $w_{\text{actual},P}(\beta_{k+1}) = w_{\text{actual},P}(\alpha_w)$.

Given Lemma 37, the proof of Lemma 16 follows in a similar manner, but taking $\beta_0 = \alpha_{1_n^T}$ instead and noting that $w_{\textit{actual},P}(\beta_0) = w_{\textit{actual},P}(\alpha_{1_n^T}) \leq \sqrt{n}$. Also, note that if the edge $\{u_w, x_{\text{extra}}\}_2$ of α_w does not vanish, we use the local weight scheme of Lemma 12 to directly obtain a bound of $(\log n)^{O(k)}(d\sqrt{n})^{k+1}$ for Lemma 14 and $(\log n)^{O(k)}\sqrt{d} n^{3/4}(d\sqrt{n})^k$ for Lemma 16.

Thus we complete the proofs of Lemmas 14 and 16 by proving Lemma 37 below.

Proof [Proof of Lemma 37] We enumerate the possible cases for the shape β_j. Because β_j contains a left-critical edge and no vanishing right-critical edge, we have that $\beta_j \in \{\alpha_{A^*,3}, \alpha_1, \alpha_{2a}, \alpha_{3a}, \alpha_{3b}, \alpha_4\}$. In each case, we refer to the definition of $w_{\text{actual},P}(\beta_j)$ in Section B.8 to verify that (26) holds. Note that because $w_{\text{actual}}(u_w) \geq w_{\text{ideal}}(u_w)$ by definition of w_{actual}, it suffices to show

$$w_{\text{actual},P}(\beta_j) w_{\text{actual},P}(x_{\text{extra}}) \geq w_{\text{ideal},P}(\beta_j) w_{\text{ideal},P}(x_{\text{extra}})$$

in order to conclude that

$$w_{\text{actual},P}(\beta_j) w_{\text{actual},P}(\beta_{k+1}) \geq w_{\text{ideal},P}(\beta_j) w_{\text{ideal},P}(\beta_{k+1}).$$

- **Case** $\beta_j = \beta_0 = \alpha_{A^*,3}$**:** $w_{\text{ideal},P}(\beta_0) = \frac{1}{\sqrt{d}}\sqrt[4]{d}\sqrt{n} = \sqrt{n}/\sqrt[4]{d}$ and $w_{\text{ideal},P}(x_{\text{extra}}) = \sqrt[4]{d}$ since we know the square in β_0 is not isolated. On the other hand, $w_{\text{actual},P}(x_{\text{extra}}) = \sqrt{d}\sqrt[4]{n}$ and $w_{\text{actual},P}(\beta_j) = n^{3/4}/\sqrt{d}$. We immediately observe that

$$w_{\text{actual},P}(\beta_j)w_{\text{actual},P}(x_{\text{extra}}) \geq \sqrt{n} = w_{\text{ideal},P}(\beta_j)w_{\text{ideal},P}(x_{\text{extra}}).$$

- **Case** $\beta_j = \alpha_1$**:** By symmetry, it suffices to consider the case where the edge $\{u_w, x_{\text{extra}}\}_2$ of α_w is identified with the edge $\{u, x_1\}$ of α_1. Note that this means $\tau(u) = LR$, $\tau(x_1) = LR$ and $\tau(x_{\text{extra}}) = L$, so $w_{\text{ideal},P}(u) = w_{\text{ideal},P}(x_1) = 1$, $w_{\text{ideal},P}(x_2) \leq \sqrt{d}$, $w_{\text{ideal},P}(v) \leq \sqrt{n}$ and $w_{\text{ideal},P}(x_{\text{extra}}) \leq \sqrt{d}$. Assembling this information, we see:

$$w_{\text{actual},P}(\beta_j)w_{\text{actual},P}(x_{\text{extra}}) \geq n^{3/4}\sqrt{d} \cdot (\sqrt{d}/\sqrt[4]{n}) = d\sqrt{n} \geq w_{\text{ideal},P}(\beta_j)w_{\text{ideal},P}(x_{\text{extra}}),$$

  and $c(\beta_j)w_{\text{actual},P}(\beta_j) = O(n^{3/4}\sqrt{d})$.

- **Case** $\beta_j = \alpha_{2a}$**:** We divide the argument for this case into two sub-cases, based on whether or not the edge $\{u, x\}_2$ in α_{2a} vanishes.

  - **Sub-case** $\{u, x\}_2$ **does not vanish:** Note that $\tau(u) \in \{L, LR\}$, $\tau(x) \in \{R, LR\}$, $\tau(v) \in \{R, LR\}$, $\tau(x_{\text{extra}}) = L$ and u, x, x_{extra} are not isolated, so $w_{\text{ideal},P}(u) \leq n^{1/4}, w_{\text{ideal},P}(x) \leq d^{1/4}, w_{\text{ideal},P}(v) \leq n^{1/2}$ and $w_{\text{ideal},P}(x_{\text{extra}}) \leq d^{1/4}$. Assembling this information, we see:

$$w_{\text{actual},P}(\beta_j)w_{\text{actual},P}(x_{\text{extra}}) \geq n^{3/4}\sqrt{d} \geq w_{\text{ideal},P}(\beta_j)w_{\text{ideal},P}(x_{\text{extra}})$$

    and $c(\beta_j)w_{\text{actual},P}(\beta_j) = O(n)$.

  - **Sub-case** $\{u, x\}_2$ **vanishes:** If $\{u, x\}_2$ vanishes, it cannot be right-critical, so either $\tau(u) = LR$ or $\tau(v) = LR$. By symmetry, it suffices to consider the case that $\tau(u) = LR$. Note that $\tau(x) \in \{R, LR\}, \tau(v) \in \{R, LR\}$, and $\tau(x_{\text{extra}}) = L$, so $w_{\text{ideal},P}(u) = 1, w_{\text{ideal},P}(x) = \sqrt{d}, w_{\text{ideal},P}(v) \leq \sqrt{n}$ and $w_{\text{ideal},P}(x_{\text{extra}}) \leq \sqrt{d}$. Assembling this information, we see:

$$w_{\text{actual},P}(\beta_j)w_{\text{actual},P}(x_{\text{extra}}) \geq d^{3/2}n^{1/4} \geq d\sqrt{n} \geq w_{\text{ideal},P}(\beta_j)w_{\text{ideal},P}(x_{\text{extra}})$$

    and $c(\beta_j)w_{\text{actual},P}(\beta_j) = O(d\sqrt{n})$.

- **Case** $\beta_j = \alpha_{3a}$**:** Note that $w_{\text{ideal},P}(\beta_j) \leq d$. We may immediately conclude

$$w_{\text{actual},P}(\beta_j)w_{\text{actual},P}(x_{\text{extra}}) \geq d\sqrt{n} \cdot (\sqrt{d}/\sqrt[4]{n}) \geq d^{3/2} \geq w_{\text{ideal},P}(\beta_j)w_{\text{ideal},P}(x_{\text{extra}})$$

  and $c(\beta_j)w_{\text{actual},P}(\beta_j) = O(d\sqrt{n})$.

- **Case** $\beta_j = \alpha_{3b}$ **or** α_4**:** Note that $w_{\text{ideal},P}(\beta_j) \leq \sqrt{d}$. So, we have

$$w_{\text{actual},P}(\beta_j)w_{\text{actual},P}(x_{\text{extra}}) \geq d \geq w_{\text{ideal},P}(\beta_j)w_{\text{ideal},P}(x_{\text{extra}})$$

  and $c(\beta_j)w_{\text{actual},P}(\beta_j) \leq O(d^{3/2}\sqrt[4]{n})$.

■

Appendix C. Connection to an average-case discrepancy problem

Recently, Aubin, Perkins, and Zdeborová Aubin et al. (2019) and Turner, Meka and Rigollet Turner et al. (2020) studied the discrepancy of random matrices. Formally, they showed that if A is an $m \times n$ matrix with i.i.d. standard Gaussian entries and $m = \Theta(n)$, then $\mathsf{disc}(A) = \Theta(\sqrt{n})$ with high probability, where the discrepancy of A is defined to be $\mathsf{disc}(A) = \min_{\sigma \in \{\pm 1\}^n} \|A\sigma\|_\infty$. Since the proof of the lower bound in this result is via a union bound over $\sigma \in \{\pm 1\}^n$, we pose the following question: is there a *computationally efficient* algorithm for certifying a lower bound on $\mathsf{disc}(A)$ for random A? By certification algorithm, we mean an algorithm that on input A always outputs a value that lower bounds $\mathsf{disc}(A)$, but for random A, the value is close to the true value $\Theta(\sqrt{n})$ with high probability. This question is inspired by a long line of work on certifying unsatisfiability of random constraint satisfaction problems (see e.g. Raghavendra et al. (2017) and references therein), but also has an application to the detection problem in the negatively-spiked Wishart model defined below.

Consider the problem of distinguishing which of the following two distributions a matrix $A \in \mathbb{R}^{m \times n}$ is generated from:

- **Null:** $A_{ij} \sim \mathcal{N}(0, 1)$, for all $i \in [m], j \in [n]$ independently.
- **Planted:** The rows A_i are independently sampled from $\mathcal{N}(0, I_n - \frac{1}{n} vv^T)$, where $v \sim \textsc{Unif}\{\pm 1\}^n$.

As mentioned, under the null model, $\mathsf{disc}(A) = \Theta(\sqrt{n})$ with high probability Turner et al. (2020). On the other hand, it is straightforward to verify that $\mathsf{disc}(A) = 0$ under the planted model. Hence, any algorithm that can certify non-trivial lower bounds on the discrepancy of a Gaussian matrix A can also solve the above detection problem. Bandeira, Kunisky, and Wein Bandeira et al. (2020) show that in the regime $m = \alpha n$, for $\alpha > 0$ a constant, distinguishing the above two distributions is hard for the class of low-degree polynomial distinguishers when $\alpha < 1$ and easy when $\alpha > 1$. While the class of low-degree polynomial algorithms is conjectured to match the performance of all polynomial-time algorithms for a wide variety of average-case problems Hopkins (2018); Kunisky et al. (2022), the above result does not have any formal implication for the powerful class of SDP-based algorithms.

Define the following SDP relaxation (also known as *vector discrepancy* Nikolov (2013)) of discrepancy:

$$\begin{aligned} \mathsf{SDP}(A) := \min_{X \in \mathbb{R}^{n \times n}} & \max_{i \in [m]} A_i^T X A_i \\ \text{s.t. } & X \succeq 0 \\ & \mathsf{diag}(X) = 1_n. \end{aligned}$$

It can be verified that for all A, it holds that $\mathsf{SDP}(A) \leq \mathsf{disc}(A)^2$. We now state a formal connection, implicit in the work of Saunderson et al. Saunderson et al. (2012), between the ability of the SDP to certify a non-trivial lower bound on the discrepancy and the ellipsoid fitting problem.

Theorem 38 (Saunderson et al. (2012)) *Let $A \in \mathbb{R}^{m \times n}$ have i.i.d. standard Gaussian entries and $m \leq n$. Then*

$$\mathbb{P}(\mathsf{SDP}(A) = 0) = \mathbb{P}(v_1, \ldots, v_n \textit{ have the ellipsoid fitting property}),$$

where $v_1, \ldots, v_n$ are independent samples from $\mathcal{N}(0, I_d)$ and $d = n - m$.

Combining Theorems 38 and 2, we conclude that the SDP fails to solve the detection problem in the negatively-spiked Wishart model when $m < n - \sqrt{n}\,\mathrm{polylog}(n)$. In particular, the inability of the SDP to distinguish between instances with discrepancy 0 and $\Theta(\sqrt{n})$ matches the *worst-case* hardnes of approximation result due to Charikar, Newman and Nikolov Charikar et al. (2011). Further, if Conjecture 1 is true, the threshold for success of the SDP is exactly $m = n - 2\sqrt{n}$. These results complement those of Mao and Wein Mao and Wein (2021) by confirming the phase transition for the SDP takes place at the same finite-scale corrected value $m = n - \sqrt{n}\,\mathrm{polylog}(n)$ as for low-degree polynomials.

The proof of Theorem 38, which we provide for the sake of completeness, makes use of the following lemma.

Lemma 39 (Lemma 2.4 and Proposition 3.1 of Saunderson et al. (2012)) *Let $\mathcal{U} \subseteq \mathbb{R}^n$ be a subspace. There exists $X \in \mathbb{R}^{n\times n}$ with $X \succeq 0$ and $\mathsf{diag}(X) = 1_n$ such that $\mathcal{U}$ is contained in the kernel of X if and only if there is a matrix V whose row span is the orthogonal complement of $\mathcal{U}$ and whose columns have the ellipsoid fitting property.*

Proof [Proof of Theorem 38] We begin the proof with a definition from Saunderson et al. (2012): a subspace $\mathcal{U}$ has the ellipsoid fitting property if there exists a matrix whose row span is $\mathcal{U}$ and whose columns satisfy the ellipsoid fitting property. By definition, $\mathsf{SDP}(A) = 0$ means that there exists $X \in \mathbb{R}^{n\times n}$ with $X \succeq 0$ and $\mathsf{diag}(X) = 1_n$ satisfying $A_i^T X A_i = 0$ for $i = 1, \dots, m$. Equivalently, the subspace $\mathcal{U} = \mathsf{span}\{A_1, \dots, A_m\}$ is contained in the kernel of X. Defining $\mathcal{U}^\perp$ to be the orthogonal complement of $\mathcal{U}$, Lemma 39 tells us that $\mathsf{SDP}(A) = 0$ is equivalent to $\mathcal{U}^\perp$ having the ellipsoid fitting property. However, note that $\mathcal{U}^\perp$ has the same distribution as the span of $v_1, \dots, v_n$, independent samples from $\mathcal{N}(0, I_d)$ with $d = n - m$. Altogether, we have:

$$\begin{aligned}\mathbb{P}(\mathsf{SDP}(A) = 0) &= \mathbb{P}(\mathcal{U}^\perp \text{ has the ellipsoid fitting property}) \\ &= \mathbb{P}(v_1, \dots, v_n \text{ have the ellipsoid fitting property}).\end{aligned}$$

■

Appendix D. Invertibility lemma

Lemma 40 *If $n > d(d+1)/2$, then $\mathcal{A}\mathcal{A}^*$ is not invertible for any $v_1, \dots, v_n$. If $n \le d(d+1)/2$, then $\mathcal{A}\mathcal{A}^*$ is invertible with probability 1.*

Proof Consider the vectors $v_1 v_1^T, \dots, v_n v_n^T$ in the $d(d+1)/2$-dimensional vector space $\mathbb{S}^{d\times d}$ of symmetric $d \times d$ matrices. Since $\mathcal{A}\mathcal{A}^*$ is the Gram matrix of the vectors $v_1 v_1^T, \dots, v_n v_n^T$, it $\mathcal{A}\mathcal{A}^*$ is invertible iff $v_1 v_1^T, \dots, v_n v_n^T$ are linearly independent. If $n > d(d+1)/2$, then clearly there must be a linear dependency since the vector space $\mathbb{S}^{d\times d}$ has dimension $d(d+1)/2$.

We now show linear independence with probability 1 provided $n \le d(d+1)/2$. Defining $\Pi : \mathbb{S}^{d\times d} \to \mathbb{S}^{d\times d}$ to be the projector onto the orthogonal complement of $\mathsf{span}(v_2 v_2^T, \dots, v_n v_n^T)$, the proof will be complete by showing $\mathbb{P}(\Pi(v_1 v_1^T) = 0) = 0$. Observe that since Π is a projector, all of its eigenvalues are 0 or 1 and since $n \le d(d+1)/2$, Π has rank at least 1, so it has an eigenvector $M \in \mathbb{S}^{d\times d}$ with eigenvalue 1. As a consequence, we have that $\mathbb{P}(\Pi(v_1 v_1^T) = 0) \le \mathbb{P}(\langle M, v_1 v_1^T \rangle =$

0). Since $M \in \mathbb{S}^{d\times d}$, we may write its eigendecomposition as $M = \sum_{i=1}^{d} \lambda_i w_i w_i^T$, where $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$ and $\{w_1, \ldots, w_d\}$ form an orthonormal basis of $\mathbb{R}^d$. Now, we have that

$$\left\langle M, v_1 v_1^T \right\rangle = \sum_{i=1}^{d} \lambda_i \left\langle w_i, v_1 \right\rangle^2 .$$

We see that $\left\langle M, v_1 v_1^T \right\rangle$ is a non-zero linear combination (because $M \neq 0$) of d independent and identically distributed random variables $\left\langle w_1, v_1 \right\rangle^2, \ldots, \left\langle w_d, v_1 \right\rangle^2$ with distribution χ_1^2 each. Hence, we conclude that $\mathbb{P}(\left\langle M, v_1 v_1^T \right\rangle = 0) = 0$, completing the proof. ■

Appendix E. Details of experiments

In this section, we elaborate on how the plots in Figure 1 were generated. The plots corresponding to the SDP and the least-squares construction appeared in Saunderson et al. (2013), but for different ranges of (n, d). To generate the left plot in Figure 1, we used the the CVXPY package to test feasibility of the original ellipsoid fitting SDP. We implemented two shortcuts to reduce the computation time. First, for we performed the simulation only for $n \leq d(d+1)/2$ since the linear system will be infeasible with probability 1 for $n > d(d+1)/2$. For all $n > d(d+1)/2$, we filled in the cell corresponding to (n, d) with black without actually performing the simulation, since by Lemma 40 the ellipsoid fitting property will fail with probability 1 in this regime. Second, for all of plots in Figure 1, we performed the simulations starting from $d = 1$ until encountering 5 consecutive values of d for which all 10 of the trials were successful and then filled in all remaining cells corresponding to larger values of d with white. We believe that this shortcut does not affect the final appearance of the plot in any noticeable way.

We remark that for the least-squares construction, the $n = cd^2$ scaling of the phase transition is only apparent for larger values of n, d in the middle plot of Figure 1 and that the transition from infeasibility to feasibility is much coarser than in the left and right plots of Figure 1. To accentuate these effects, we reproduce the plots in Figure 1 on a $\log_2$ scale in Figure 18 so that the parabola $n = cd^2$ becomes a line with slope 2.

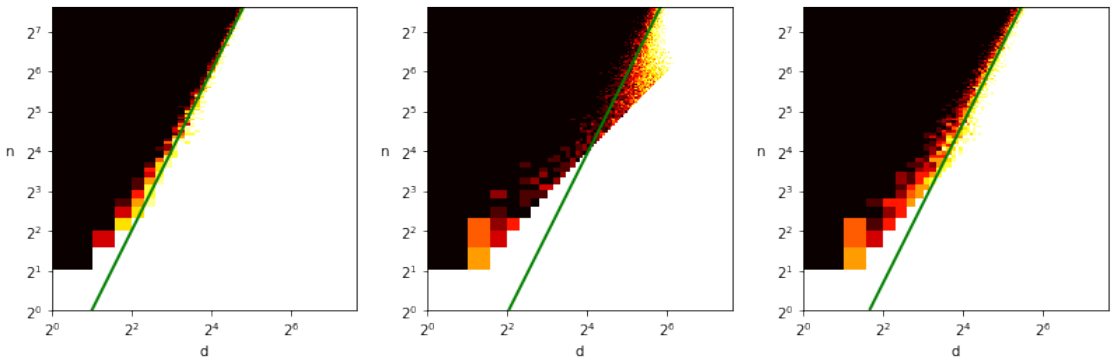

Figure 18: Plots from Figure 1 on a $\log_2$ scale. **(Left)** Ellipsoid fitting SDP, **(Middle)**: Least-squares, **(Right)**: Identity perturbation

Appendix F. Probabilistic norm bounds

We restate Theorem 22 below for convenience.

Theorem 41 (Theorem 22) *Given $D_V, D_E \in \mathbb{N}$ such that $D_E \geq D_V \geq 2$ and $\epsilon > 0$, with probability at least $1-\epsilon$, for all shapes α on square and circle vertices such that $|V(\alpha)| \leq D_V$ and $|E_\alpha| \leq D_E$, $|U_\alpha| \leq 1$, and $|V_\alpha| \leq 1$,*

$$\|M_\alpha\| \leq \left((2D_E+2)\ln(D_V) + \ln(11n) + \ln\left(\frac{1}{\epsilon}\right)\right)^{|V(\alpha)|+|E(\alpha)|} n^{\frac{\varphi(V(\alpha))-\varphi(S_{min})+\varphi(\mathrm{Iso}(\alpha))}{2}}$$

where S_{min} is a minimum vertex separator of α.

Proof Corollary 8.16 of Ahn et al. (2016) says that for all $\epsilon' > 0$ and all shapes α with square and circle vertices and no isolated vertices outside of $U_\alpha \cup V_\alpha$, with probability at least $1-\epsilon'$,

$$\|M_\alpha\| \leq 2|V_\circ(\alpha)|^{|V_\circ(\alpha)|}|V_\square(\alpha)|^{|V_\square(\alpha)|} n^{\frac{\varphi(V(\alpha))-\varphi(S_{min})+\varphi(\mathrm{Iso}(\alpha))}{2}} \cdot \left(6e\left\lceil \frac{\ln(\frac{n^{\varphi(S_{min})}}{\epsilon'})}{6(|V(\alpha)\setminus(U_\alpha\cap V_\alpha)|+|E(\alpha)|)}\right\rceil\right)^{|E(\alpha)|+|V(\alpha)\setminus(U_\alpha\cap V_\alpha)|}.$$

The result is trivial if $|V(\alpha)| \leq 1$ or $E(\alpha) = \emptyset$ so we can assume that $|V(\alpha)| \geq 1$ and $E(\alpha) \geq 1$. For the shapes α we are considering, $\varphi(S_{min}) \leq 1$ so for each such shape α, for all $\epsilon' > 0$,

$$\|M_\alpha\| \leq |V(\alpha)|^{|V(\alpha)|+|E(\alpha)|} n^{\frac{\varphi(V(\alpha))-\varphi(S_{min})+\varphi(\mathrm{Iso}(\alpha))}{2}} \left(6e\left\lceil \frac{\ln\left(\frac{n}{\epsilon'}\right)}{6|V(\alpha)|}\right\rceil\right)^{|V(\alpha)|+|E(\alpha)|}.$$

We will now apply this to all such shapes α with $\epsilon' = \frac{\epsilon}{11{D_V}^{(2D_E+2)}}$ and take a union bound. Since $D_E \geq D_V \geq |V(\alpha)|$, we have that $\ln\left(\frac{n}{\epsilon'}\right) \geq (2D_E+2)\ln(D_V) \geq 2D_V \geq 2|V(\alpha)|$, so

$$\left\lceil \frac{\ln\left(\frac{n}{\epsilon'}\right)}{6|V(\alpha)|}\right\rceil \leq \frac{\ln\left(\frac{n}{\epsilon'}\right)}{2|V(\alpha)|}.$$

Thus, for each such shape α, with probability at least $1 - \frac{\epsilon}{11{D_V}^{(2D_E+2)}}$,

$$\|M_\alpha\| \leq \left((2D_E+2)\ln(D_V) + \ln(11n) + \ln\left(\frac{1}{\epsilon}\right)\right)^{|V(\alpha)|+|E(\alpha)|} n^{\frac{\varphi(V(\alpha))-\varphi(S_{min})+\varphi(\mathrm{Iso}(\alpha))}{2}}.$$

Using the following proposition and taking a union bound, we have that with probability at least $1-\epsilon$, the above bound holds for all such shapes α, as needed. ■

Proposition 42 *If $D_V, D_E \in \mathbb{N}$ and $D_V \geq 2$ then there are at most $4{D_V}^{(2D_E+2)}$ shapes α on square and circle vertices such that $|V(\alpha)| \leq D_V$, $|E_\alpha| \leq D_E$, $|U_\alpha| \leq 1$, and $|V_\alpha| \leq 1$.*

Proof We can specify a non-empty shape α as follows:

1. Specify whether U_α and V_α have a circle vertex, a square vertex, or are empty. If U_α and V_α both have circle vertices or both have square vertices, specify whether they are the same vertex. There are a total of 11 choices for this.

2. Specify the number of circle vertices and square vertices in $V(\alpha) \setminus (U_\alpha \cup V_\alpha)$. There are at most D_V^2 choices for this.

3. For each of the D_E possible edges, either specify its two endpoints or $\emptyset$ if it does not exist. There are at most $\binom{D_V}{2} + 1 \leq D_V^2$ choices for each possible edge.

■

Appendix G. Proof of Lemma 4

For convenience we restate the lemma below.

Lemma 43 (Lemma 4) *Let $B = \Gamma + \alpha I_n$. We have*

$$(\mathcal{A}\mathcal{A}^*)^{-1} 1_n = \frac{1}{s^2 - ru} \cdot \left((1 + 1_n^T B^{-1} w) B^{-1} 1_n - (1_n^T B^{-1} 1_n) B^{-1} w \right) \tag{27}$$

where r, s, u are defined as

$$\begin{pmatrix} r & s \\ s & u \end{pmatrix} := \begin{pmatrix} 1_n^T B^{-1} 1_n & 1 + 1_n^T B^{-1} w \\ 1 + 1_n^T B^{-1} w & -d + w^T B^{-1} w \end{pmatrix}.$$

Proof

The Woodbury formula Woodbury (1950) states that

$$(B + UCV)^{-1} 1_n = B^{-1} 1_n - B^{-1} U (C^{-1} + V B^{-1} U)^{-1} V B^{-1} 1_n. \tag{28}$$

We set B as above. Let $U \in \mathbb{R}^{n \times 2}$ be defined by

$$U_{ij} = \begin{cases} 1 & \text{if } j = 1, \text{ and} \\ w_i = \|v_i\|_2^2 - d & \text{if } j = 2. \end{cases}$$

So the columns of U are 1_n and w. Let $V = U^T$, and set

$$C = \begin{pmatrix} d & 1 \\ 1 & 0 \end{pmatrix}.$$

Observe that $UCV = W$. Next,

$$\begin{aligned} C^{-1} + V B^{-1} U &= \begin{pmatrix} 0 & 1 \\ 1 & -d \end{pmatrix} + \begin{pmatrix} 1_n^T B^{-1} 1_n & 1_n^T B^{-1} w \\ 1_n^T B^{-1} w & w^T B^{-1} w \end{pmatrix} \\ &= \begin{pmatrix} 1_n^T B^{-1} 1_n & 1 + 1_n^T B^{-1} w \\ 1 + 1_n^T B^{-1} w & -d + w^T B^{-1} w \end{pmatrix} =: \begin{pmatrix} r & s \\ s & u \end{pmatrix}. \end{aligned}$$

Thus

$$(C^{-1} + VB^{-1}U)^{-1} = \frac{1}{ru - s^2}\begin{pmatrix} u & -s \\ -s & r \end{pmatrix},$$

and

$$VB^{-1}1_n = \begin{pmatrix} 1_n^T B^{-1} 1_n \\ w^T B^{-1} 1_n \end{pmatrix} = \begin{pmatrix} r \\ s-1 \end{pmatrix}.$$

Hence

$$(C^{-1} + VB^{-1}U)^{-1} VB^{-1}1_n = \frac{1}{ru - s^2} \cdot \begin{pmatrix} ru - s(s-1) \\ -sr + r(s-1) \end{pmatrix} = \frac{1}{ru - s^2} \cdot \begin{pmatrix} ru - s^2 + s \\ -r \end{pmatrix}.$$

Next, since U has first column 1_n and second column w,

$$\begin{aligned}
(AA^*)^{-1}1_n &= B^{-1}1_n - B^{-1}U(C^{-1} + VB^{-1}U)^{-1}VB^{-1}1_n \\
&= B^{-1}1_n - \frac{1}{ru - s^2}B^{-1}U \cdot \begin{pmatrix} ru - s^2 + s \\ -r \end{pmatrix} \\
&= \big(1 - \frac{ru - s^2 + s}{ru - s^2}\big)B^{-1}1_n + \frac{r}{ru - s^2}B^{-1}w \\
&= \frac{1}{ru - s^2} \cdot \big(- sB^{-1}1_n + rB^{-1}w\big) \\
&= \frac{1}{s^2 - ru} \cdot \Big((1 + 1_n^T B^{-1} w)B^{-1}1_n - (1_n^T B^{-1} 1_n)B^{-1}w\Big).
\end{aligned}$$

■

Appendix H. Hermite polynomials

Here, we provide some technical results regarding Hermite polynomials that are useful when applying Proposition 20. Throughout, we use the convention $\mathbb{N} = \{1, 2, 3 \ldots\}$ (with 0 not included).

Lemma 44 *For all $j \in \mathbb{N}$,*

$$h_1(x)h_j(x) = xh_j(x) = \sqrt{j+1}h_{j+1}(x) + \sqrt{j}h_{j-1}(x).$$

Proof Since the normalized Hermite polynomials $\{h_j : j \in \mathbb{N} \cup \{0\}\}$ are orthonormal with respect to the inner product

$$\langle h_i, h_j \rangle := \mathop{\mathbb{E}}_{x \sim N(0,1)}[h_i(x)h_j(x)],$$

we have that

$$xh_j(x) = \sum_{k=0}^{\infty} \mathop{\mathbb{E}}_{y \sim N(0,1)}[yh_j(y)h_k(y)]h_k(x).$$

We now make the following observations:

1. If $k < j-1$ then $\mathbb{E}_{y\sim N(0,1)}[yh_j(y)h_k(y)] = 0$ because $yh_k(y)$ is a degree $k+1$ polynomial and $h_j(y)$ is orthogonal to all polynomials of degree less than j.

2. If $k > j+1$ then $\mathbb{E}_{y\sim N(0,1)}[yh_j(y)h_k(y)] = 0$ because $yh_j(y)$ is a degree $j+1$ polynomial and $h_k(y)$ is orthogonal to all polynomials of degree less than k.

3. If $k = j$ then $\mathbb{E}_{y\sim N(0,1)}[yh_j(y)h_k(y)] = 0$ because $yh_j(y)h_k(y)$ is an odd polynomial.

4. If $k = j-1$ then the leading term of $yh_k(y) = \frac{x^j}{\sqrt{(j-1)!}}$ so we can write $yh_k(y) = \sqrt{j}h_j(y) + p$ where p has degree at most $j-1$. This implies that $\mathbb{E}_{y\sim N(0,1)}[yh_j(y)h_k(y)] = \mathbb{E}_{y\sim N(0,1)}[\sqrt{j}(h_j(y))^2] = \sqrt{j}$.

5. If $k = j+1$ then the leading term of $yh_j(y) = \frac{x^k}{\sqrt{j!}}$ so we can write $yh_j(y) = \sqrt{j+1}h_k(y)+p$ where p has degree at most $k-1$. This implies that

$$\mathop{\mathbb{E}}_{y\sim N(0,1)}[yh_j(y)h_k(y)] = \mathop{\mathbb{E}}_{y\sim N(0,1)}[\sqrt{j+1}(h_k(y))^2] = \sqrt{j+1}.$$

■

Corollary 45 *For all $j \in \mathbb{N}$ with $j \geq 2$,*

$$x^2h_j(x) = \sqrt{(j+1)(j+2)}h_{j+2}(x) + (2j+1)h_j(x) + \sqrt{j(j-1)}h_{j-2}(x).$$

Proof By Lemma 44,

$$\begin{aligned} x^2h_j(x) &= x\sqrt{j+1}h_{j+1}(x) + x\sqrt{j}h_{j-1}(x) \\ &= \sqrt{j+1}(\sqrt{j+2}h_{j+2}(x) + \sqrt{j+1}h_j(x)) + \sqrt{j}(\sqrt{j}h_j(x) + \sqrt{j-1}h_{j-2}(x)) \\ &= \sqrt{(j+1)(j+2)}h_{j+2}(x) + (2j+1)h_j(x) + \sqrt{j(j-1)}h_{j-2}(x). \end{aligned}$$

■

Corollary 46 *For all $j \in \mathbb{N}$,*

$$h_2(x)h_j(x) = \frac{x^2-1}{\sqrt{2}}h_j(x) = \sqrt{\frac{(j+1)(j+2)}{2}}h_{j+2}(x) + \sqrt{2}jh_j(x) + \sqrt{\frac{j(j-1)}{2}}h_{j-2}(x).$$

Remark 47 *Note that for the case $j = 1$, $\sqrt{j(j-1)} = 0$. Thus, for $j = 1$, even though $h_{j-2}(x) = h_{-1}(x)$ is undefined it does not matter as $\sqrt{j(j-1)}h_{j-2}(x) = 0$ regardless of what $h_{-1}(x)$ is.*

Appendix I. Analysis of the Pseudo-Calibration Construction

In this section, we describe and analyze the construction of M from Ghosh et al. (2020). This construction is obtained by using pseudo-calibration (for background on pseudo-calibration, see Barak et al. (2019)) on the following distributions.

Random: Sample n vectors $v_1, \ldots, v_n$ from $\mathcal{N}(0, I_d)$.

Planted: First sample a hidden direction u from $\{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$ and n random ± 1 variables $b_1, \ldots, b_n$. Then sample n vectors $v'_1, \ldots, v'_n$ from $\mathcal{N}(0, I_d)$ and replace each vector v'_i with $v_i = v'_i - \langle v'_i, u\rangle u + b_i u$.

For the planted distribution, the rank one matrix $M = uu^T$ satisfies $v_i^T M v_i = 1$ for all $i \in [n]$. For the random distribution, there is no hidden direction u but with high probability, pseudo-calibration will still give us a matrix M such that $v_i^T M v_i = 1$ for all $i \in [n]$. In order to describe this matrix M, we need a few definitions.

Definition 48 *Given values $\{\alpha_{i,a} : i \in [n], a \in [d]\}$, we make the following definitions:*

1. *Define $|\alpha|$ to be $|\alpha| = \sum_{i=1}^n \sum_{a=1}^d \alpha_{i,a}$*
2. *Define α_i to be $\alpha_i = \sum_{a=1}^d \alpha_{i,a}$*
3. *Define α_a^T to be $\alpha_a^T = \sum_{i=1}^n \alpha_{i,a}$*
4. *Define $\alpha!$ to be $\alpha! = \prod_{i=1}^n \prod_{a=1}^d \alpha_{i,a}!$*
5. *Define $h_\alpha(v_1, \ldots, v_n)$ to be $h_\alpha(v_1, \ldots, v_n) = \prod_{i=1}^n \prod_{a=1}^d h_{\alpha_{i,a}}((v_i)_a)$*

Let $T = \Omega(\log n)$ be a truncation parameter. By Lemma 4.4 of Ghosh et al. (2020), the construction given by pseudo-calibration with truncation parameter T is as follows.

Definition 49 *Define $\tilde{E}[1]$ to be*

$$\tilde{E}[1] = \sum_{\alpha : |\alpha| \le T, \text{ For all } i \in [n], a \in [d], \alpha_i \text{ and } \alpha_a^T \text{ are even}} \frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!} h_{\alpha_i}(1)\right)}{\sqrt{\alpha!} d^{\frac{|\alpha|}{2}}} h_\alpha(v_1, \ldots, v_n)$$

Definition 50 *For all $a \in [d]$, we define $\tilde{E}[x_a^2]$ to be $\tilde{E}[x_a^2] = \frac{1}{d}\tilde{E}[1]$. For all distinct $a, b \in [d]$, define $\tilde{E}[x_a x_b]$ to be*

$$\tilde{E}[x_a x_b] = \sum_{\substack{\alpha : |\alpha| \le T, \text{ For all } i \in [n], c \in [d] \setminus \{a,b\}, \alpha_i \text{ and } \alpha_c^T \text{ are even}, \\ \alpha_a^T \text{ and } \alpha_b^T \text{ are odd.}}} \frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!} h_{\alpha_i}(1)\right)}{\sqrt{\alpha!} d^{\frac{|\alpha|}{2}+1}} h_\alpha(v_1, \ldots, v_n)$$

Remark 51 *These equations have different coefficients than Lemma 4.4 of Ghosh et al. (2020) because we are using the normalized Hermite polynomials.*

Definition 52 *For all distinct $a, b \in [d]$, we take $M_{ab} = \frac{\tilde{E}[x_a x_b]}{\tilde{E}[1]}$. For all $a \in [d]$, we take $M_{aa} = \frac{\tilde{E}[x_a^2]}{\tilde{E}[1]} = \frac{1}{d}$.*

I.1. Verifying M is PSD

While the pseudo-calibration construction is more complicated than the least squares and identity perturbation constructions, it is actually easier to give a rough analysis for it. The reason is that $\tilde{E}[1]M$ can be directly decomposed into shapes. Moreover, all of the shapes α appearing in $\tilde{E}[1]M$ have the following properties

1. M_α appears in $\tilde{E}[1]M$ with coefficient $\lambda_\alpha = O(d^{-(\frac{|E(\alpha)|}{2}+1)})$ where we take $|E(\alpha)|$ to be the sum of the labels of the edges in $E(\alpha)$.
2. Let $U_\alpha = (u)$ and $V_\alpha = (v)$, every square vertex has even degree and has degree at least 2. If $u \neq v$ then u and v have odd degree. If $u = v$ then u has even degree (which may be 0).

   Note that we take the degree of a vertex to be the sum of the labels of the edges incident to that vertex.
3. Every circle vertex has even degree and has degree at least 4.

Using the same logic that we used to prove Lemma 28, each such shape α contains a path from U_α to V_α so the minimum weight vertex separator of α is a single square. By Theorem 22, with high probability, $||M_\alpha||$ is $\tilde{O}(n^{\frac{|\mathcal{V}_\circ(\alpha)|}{2}} d^{\frac{|\mathcal{V}_\square(\alpha)|-1}{2}})$. We now make the following observations:

1. Since every square vertex except u, v has degree at least 2, $|E(\alpha)| = \sum_{w\in\mathcal{V}_\square(\alpha)} deg(w) \geq 2|\mathcal{V}_\square(\alpha)| - 2$ which implies that $|\mathcal{V}_\square(\alpha)| \leq \frac{|E(\alpha)|}{2} + 1$
2. Since every circle vertex has degree at least 4, $|E(\alpha)| = \sum_{w\in\mathcal{V}_\circ(\alpha)} deg(w) \geq 4|\mathcal{V}_\circ(\alpha)|$ which implies that $|\mathcal{V}_\circ(\alpha)| \leq \frac{|E(\alpha)|}{4}$

Putting these observations together, with high probability, $\lambda_\alpha||M_\alpha||$ is

$$\tilde{O}(n^{\frac{|\mathcal{V}_\circ(\alpha)|}{2}} d^{\frac{|\mathcal{V}_\square(\alpha)|-1-|E(\alpha)|}{2}-1}) \leq \tilde{O}\left(\frac{1}{d}\left(\frac{\sqrt[8]{n}}{\sqrt[4]{d}}\right)^{|E(\alpha)|}\right)$$

This implies that the dominant term is $\frac{1}{d}I_d$ as it is the only term that appears which has no edges. Thus, with high probability, $\tilde{E}[1]M$ is PSD. As noted in Remark 5.9 of Ghosh et al. (2020), with high probability, $\tilde{E}[1]$ is $1 \pm o(1)$ so this implies that with high probability, M is PSD, as needed.

Remark 53 *This analysis is very similar to the analysis on p.21 of Ghosh et al. (2020) for attempt 1 where each edge splits its factor of $\frac{1}{\sqrt{d}}$ between its two endpoints. While this attempt fails for the higher degree setting of Ghosh et al. (2020), it succeeds for degree 2, which is what we are analyzing here.*

I.2. Verifying that $v_i^T M v_i \approx 1$

As discussed in Section 7 of Ghosh et al. (2020), pseudo-calibration guarantees that the constraints $v_i^T M v_i = 1$ are satisfied up to a very small truncation error which can be easily repaired. However, looking at the entries of M directly, it is not at all easy to see why $v_i^T M v_i \approx 1$. In this subsection, we show how to directly verify that $v_i^T M v_i \approx 1$. In particular, we give a direct proof that for each

α such that $|\alpha| \leq T-2$ (where $T = \Omega(\log n)$ is the truncation parameter), the coefficient of h_α in $\sum_{a=1}^d \sum_{b=1}^d \tilde{E}[x_a x_b](v_i)_a(v_i)_b$ matches the coefficient of h_α in $\tilde{E}[1]$. This analysis is a special case of the analysis on p.42-45 of Ghosh et al. (2020).

There are several ways that h_α can appear in $\sum_{a=1}^d \sum_{b=1}^d \tilde{E}[x_a x_b](v_i)_a(v_i)_b$. These ways are as follows:

1. For some $a \in [d]$ and $b \in [d] \setminus \{a\}$, $h_{\alpha_{i,a}-1}((v_i)_a)$ is multiplied by $(v_i)_a$ and $h_{\alpha_{i,b}-1}((v_i)_b)$ is multiplied by $(v_i)_b$, giving $\sqrt{\alpha_{i,a}\alpha_{i,b}}h_{\alpha_{i,a}}((v_i)_a)h_{\alpha_{i,b}}((v_i)_b)$.

   Letting α' be α where $\alpha_{i,a}$ and $\alpha_{i,b}$ are decreased by 1, the coefficient of $h_{\alpha'}$ in $\tilde{E}[x_a x_b]$ is

   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}} \cdot \frac{\sqrt{\alpha_{i,a}\alpha_{i,b}}h_{\alpha_i-2}(1)}{\sqrt{\alpha_i(\alpha_i-1)}h_{\alpha_i}(1)}$$

   so the total contribution from these terms is

   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}} \sum_{a=1}^d \sum_{b\in[d]\setminus\{a\}} \frac{\alpha_{i,a}\alpha_{i,b}h_{\alpha_i-2}(1)}{\sqrt{\alpha_i(\alpha_i-1)}h_{\alpha_i}(1)}$$

2. For some $a \in [d]$, $h_{\alpha_{i,a}-2}((v_i)_a)$ is multiplied by $(v_i)_a^2$, giving $\sqrt{\alpha_{i,a}(\alpha_{i,a}-1)}h_{\alpha_{i,a}}((v_i)_a)$.

   Letting α' be α where $\alpha_{i,a}$ is decreased by 2, the coefficient of $h_{\alpha'}$ in $\tilde{E}[x_a^2]$ is

   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}} \cdot \frac{\sqrt{\alpha_{i,a}(\alpha_{i,a}-1)}h_{\alpha_i-2}(1)}{\sqrt{\alpha_i(\alpha_i-1)}h_{\alpha_i}(1)}$$

   so the total contribution from these terms is

   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}} \sum_{a=1}^d \frac{\alpha_{i,a}(\alpha_{i,a}-1)h_{\alpha_i-2}(1)}{\sqrt{\alpha_i(\alpha_i-1)}h_{\alpha_i}(1)}$$

   Together, the terms in cases 1 and 2 give a total contribution of

   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}} \cdot \frac{\sqrt{\alpha_i(\alpha_i-1)}h_{\alpha_i-2}(1)}{h_{\alpha_i}(1)}$$

3. For some $a \in [d]$ and $b \in [d] \setminus \{a\}$, $h_{\alpha_{i,a}-1}((v_i)_a)$ is multiplied by $(v_i)_a$ and $h_{\alpha_{i,b}+1}((v_i)_b)$ is multiplied by $(v_i)_b$, giving $\sqrt{\alpha_{i,a}(\alpha_{i,b}+1)}h_{\alpha_{i,a}}((v_i)_a)h_{\alpha_{i,b}}((v_i)_b)$.

   Letting α' be α where $\alpha_{i,a}$ is decreased by 1 and $\alpha_{i,b}$ is increased by 1, the coefficient of $h_{\alpha'}$ in $\tilde{E}[x_a x_b]$ is

   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}} \cdot \frac{\sqrt{\alpha_{i,a}}}{d\sqrt{(\alpha_{i,b}+1)}}$$

   so the total contribution from these terms is $\frac{(d-1)\alpha_i}{d}\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}$ By symmetry, we have the same contribution from the terms where $h_{\alpha_{i,a}+1}((v_i)_a)$ is multiplied by $(v_i)_a$ and $h_{\alpha_{i,b}-1}((v_i)_b)$ is multiplied by $(v_i)_b$ so the total contribution from all of these terms is $\frac{2(d-1)\alpha_i}{d}\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}$

4. For some $a \in [d]$, $h_{\alpha_{i,a}}((v_i)_a)$ is multiplied by $(v_i)_a^2$, giving $(2\alpha_{i,a}+1)h_{\alpha_{i,a}}((v_i)_a)$.
   The coefficient of h_α in $\tilde{E}[x_a^2]$ is
   $$\frac{1}{d}\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}$$
   so the total contribution from these terms is
   $$\left(\frac{2\alpha_i}{d}+1\right)\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}$$
   Together, the terms in cases 3 and 4 give a total contribution of
   $$(2\alpha_i+1)\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}$$

5. For some $a \in [d]$ and $b \in [d]\setminus\{a\}$, $h_{\alpha_{i,a}+1}((v_i)_a)$ is multiplied by $(v_i)_a$ and $h_{\alpha_{i,b}+1}((v_i)_b)$ is multiplied by $(v_i)_b$, giving $\sqrt{(\alpha_{i,a}+1)(\alpha_{i,b}+1)}h_{\alpha_{i,a}}((v_i)_a)h_{\alpha_{i,b}}((v_i)_b)$.
   Letting α' be α where $\alpha_{i,a}$ and $\alpha_{i,b}$ are increased by 1, the coefficient of $h_{\alpha'}$ in $\tilde{E}[x_a x_b]$ is
   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}\cdot\frac{\sqrt{(\alpha_i+1)(\alpha_i+2)}h_{\alpha_i+2}(1)}{d^2\sqrt{(\alpha_{i,a}+1)(\alpha_{i,b}+1)}h_{\alpha_i}(1)}$$
   so the total contribution from these terms is
   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}\cdot\frac{d(d-1)\sqrt{(\alpha_i+1)(\alpha_i+2)}h_{\alpha_i+2}(1)}{d^2h_{\alpha_i}(1)}$$

6. For some $a \in [d]$, $h_{\alpha_{i,a}+2}((v_i)_a)$ is multiplied by $(v_i)_a^2$, giving $\sqrt{(\alpha_{i,a}+2)(\alpha_{i,a}+1)}h_{\alpha_{i,a}}((v_i)_a^2)$.
   Letting α' be α where $\alpha_{i,a}$ is increased by 2, the coefficient of $h_{\alpha'}$ in $\tilde{E}[x_a^2]$ is
   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}\cdot\frac{\sqrt{(\alpha_i+1)(\alpha_i+2)}h_{\alpha_i+2}(1)}{d^2\sqrt{(\alpha_{i,a}+2)(\alpha_{i,a}+1)}h_{\alpha_i}(1)}$$
   so the total contribution from these terms is
   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}\cdot\frac{d\sqrt{(\alpha_i+1)(\alpha_i+2)}h_{\alpha_i+2}(1)}{d^2h_{\alpha_i}(1)}$$
   Together, the terms in cases 5 and 6 give a total contribution of
   $$\frac{\left(\prod_{i=1}^n \sqrt{\alpha_i!}h_{\alpha_i}(1)\right)}{\sqrt{\alpha!}d^{\frac{|\alpha|}{2}}}\cdot\frac{\sqrt{(\alpha_i+1)(\alpha_i+2)}h_{\alpha_i+2}(1)}{h_{\alpha_i}(1)}$$

Putting everything together, it is sufficient to show that

$$\sqrt{\alpha_i(\alpha_i-1)}h_{\alpha_i-2}(1)+(2\alpha_i+1)h_{\alpha_i}(1)+\sqrt{(\alpha_i+1)(\alpha_i+2)}h_{\alpha_i+2}(1)=h_{\alpha_i}(1)$$

To show this, recall that by Corollary 45, for all $j \in \mathbb{N}\cup\{0\}$ and all $x \in \mathbb{R}$,

$$x^2h_j(x)=\sqrt{(j+1)(j+2)}h_{j+2}(x)+(2j+1)h_j(x)+\sqrt{j(j-1)}h_{j-2}(x).$$

Plugging in $x=1$ and $j=\alpha_i$, the result follows.

Appendix J. Analysis of the Identity Perturbation Construction

In this section, we provide an analysis of the identity perturbation construction and show that it is PSD provided that $n \leq d^2/\mathrm{polylog}(d)$. Again without loss of generality, we assume that $n \geq d$. Recall that

$$X := X_{\mathrm{IP}} = \frac{1}{d} I_d + \mathcal{A}^*(c)$$

where c is chosen such that $\mathcal{A}(X) = 1_n$. By direct calculation and the invertibility of $(\mathcal{A}\mathcal{A}^*)^{-1}$, there is a unique vector c satisfying the constraint $\mathcal{A}(X) = 1_n$, and it is given by $c = -\frac{1}{d}(\mathcal{A}\mathcal{A}^*)^{-1}w$, where recall $w_i = \|v_i\|^2 - d$. Again by the Woodbury formula Woodbury (1950), it holds that

$$(\mathcal{A}\mathcal{A}^*)^{-1}w = (B + UCV)^{-1}w = B^{-1}w - B^{-1}U(C^{-1} + VB^{-1}U)^{-1}VB^{-1}w \tag{29}$$

where B, U, C, and V are defined in Section G. Using a similar calculation as in Section G and recalling also the definitions of r, s, and u given there, we obtain

$$\begin{aligned}
(\mathcal{A}\mathcal{A}^*)^{-1}w &= B^{-1}w - B^{-1}U(C^{-1} + VB^{-1}U)^{-1}VB^{-1}w \\
&= B^{-1}w - \frac{1}{ru - s^2} B^{-1}U \begin{pmatrix} u & -s \\ -s & r \end{pmatrix} \begin{pmatrix} s-1 \\ u+d \end{pmatrix} \\
&= (\frac{u + sd}{ru - s^2})B^{-1}1_n - (\frac{s + rd}{ru - s^2})B^{-1}w
\end{aligned}$$

Hence,

$$X = \frac{1}{d} I_d + \frac{1}{d}(\frac{u+d}{s^2 - ru})\mathcal{A}^*B^{-1}1_n + \frac{1}{d}(\frac{d(s-1)}{s^2 - ru})\mathcal{A}^*B^{-1}1_n - \frac{1}{d}(\frac{s + rd}{s^2 - ru})\mathcal{A}^*B^{-1}w. \tag{30}$$

By (9), it holds that with high probability

$$u + d = w^T B^{-1} w = \Theta(\frac{1}{d^2})\|w\|_2^2,$$

which implies in particular that $u + d = w^T B^{-1} w > 0$. Moreover, $s^2 - ru = \Omega(n/d)$ with high probability by (8), so also $s^2 - ru \geq 0$. It follow from Lemma 8 that the second term of (30) is PSD with high probability.

Next we show that the third term of (30) satisfies

$$\|\frac{1}{d}(\frac{d(s-1)}{s^2 - ru})\mathcal{A}^*B^{-1}1_n\|_{op} = o(1/d) \tag{31}$$

with high probability. In the proof of Lemma 7, we in fact showed that $|s - 1| = |1_n^T B^{-1} w| = \tilde{o}(n/d^2)$. Moreover, the proof of Lemma 8 implies also that $\|\mathcal{A}^*B^{-1}1_n\|_{op} = O(n/d^2)$. Thus (31) follows from $s^2 - ru = \Omega(n/d)$ (see (8)) assuming that $n \leq d^2/\mathrm{polylog}(d)$.

Moreover, by Lemmas 5, 7, and 9 as well as (8), we obtain that the last term of (30) is $o(1/d)$ in operator norm assuming that $n \geq d$ and $n \leq d^2/\mathrm{polylog}(d)$.

Combining the results for the last three terms of (30), we conclude that $X = X_{\mathrm{IP}} \succeq 0$ with high probability, as desired. □

Appendix K. Notes on a previous approach

In this section, we discuss the mistake appearing in a previous version of this paper, sketch how this mistake can be repaired, and explain why we instead use the Woodbury matrix identity in the current paper.

The approach used in the previous version of this paper was as follows. Taking $M = (d^2 + d)I + d1_n1_n^T$ and $\Delta = M - AA^*$,[7] we have that

$$X_{\text{LS}} = \mathcal{A}^*((\mathcal{A}\mathcal{A}^*)^{-1}1_n) = \mathcal{A}^*((I_n - M^{-1}\Delta)^{-1}M^{-1}1_n).$$

Expanding $(I_n - M^{-1}\Delta)^{-1}$ as a Neumann series:

$$(I_n - M^{-1}\Delta)^{-1} = I_n + M^{-1}\Delta + \sum_{j=2}^{\infty}(M^{-1}\Delta)^j$$

and observing that $M^{-1}1_n = \frac{1}{d^2+d+dn}1_n$, we have that

$$(d^2 + d + dn)X_{\text{LS}} = \mathcal{A}^*(1_n) + \mathcal{A}^*((M^{-1}\Delta)1_n) + \mathcal{A}^*\left(\left(\sum_{j=2}^{\infty}(M^{-1}\Delta)^j\right)1_n\right).$$

It is a standard fact that when $n = \Omega(d)$, with high probability $\lambda_{min}(\mathcal{A}^*(1_n)) = \lambda_{min}(\sum_{i=1}^n v_i v_i^T)$ is $\Omega(n)$. In order to show that $X_{\text{LS}} \succeq 0$ with high probability, it is sufficient to show that with high probability:

1. $\|M^{-1}\Delta\|_{op} < 1$,
2. $\|\mathcal{A}^*(M^{-1}\Delta 1_n)\|_{op} = o(n)$,
3. For all $j \geq 2$, $\|\mathcal{A}^*((M^{-1}\Delta)^j 1_n)\|_{op} = o(n)$.

Proposition 5.2 of the previous version of this paper claimed that with high probability, $\|\Delta 1_n\|_\infty = \tilde{O}(d\sqrt{n})$ which implies that $\|M^{-1}\Delta 1_n\|_\infty = \tilde{O}(\frac{d\sqrt{n}}{d^2}) = o(1)$. In turn, this implies that $\|M^{-1}\Delta 1_n\|_2 = o(\sqrt{n})$. By Lemma 3 of Saunderson (2011), with high probability $\|A^*\|_{2\to op} = \Theta(d + \sqrt{n})$ so this implies that for all $j \geq 1$, $\|A^*(M^{-1}\Delta)^j 1_n\| = o(n)$.

Unfortunately, this proposition is incorrect. As we discuss below, the correct bound on $\|\Delta 1_n\|_\infty$ is $\tilde{O}(n\sqrt{d})$. This gives $\|M^{-1}\Delta 1_n\|_\infty = \tilde{O}\left(\frac{n\sqrt{d}}{d^2}\right) = \tilde{O}\left(n/d^{3/2}\right)$. This is sufficient when $n \ll d^{3/2}$ but is not sufficient when $n \gg d^{3/2}$. This means that in order to prove our result using this approach, we cannot consider $\mathcal{A}^*$ and $(M^{-1}\Delta)^j 1_n$ separately. Instead, we must analyze their product $\mathcal{A}^*((M^{-1}\Delta)^j 1_n)$.

Remark 54 *If we showed the stronger statement $\|(M^{-1}\Delta)^j 1_n\|_\infty = o(1)$ for all j then we would have that every coordinate of $(I_n - M^{-1}\Delta)^{-1}1_n = 1_n + M^{-1}\Delta 1_n + \sum_{j=2}^{\infty}(M^{-1}\Delta)^j 1_n$ is positive. This implies that $X_{LS} \succeq 0$. We found experimentally that this is true when $n \ll d^{3/2}$. However, when $n \gg d^{3/2}$, $(I_n - M^{-1}\Delta)^{-1}1_n$ has negative coordinates, so the interaction between $\mathcal{A}^*$ and $(I_n - M^{-1}\Delta)^{-1}1_n$ is crucial.*

7. Note that this Δ is different from the one used in the rest of the current paper.

K.1. Computing Δ and $M^{-1}\Delta$

In order to discuss why Proposition 5.2 of the previous version of this paper is incorrect and how to actually carry out this approach, it is helpful to express Δ and $M^{-1}\Delta$ in terms of graph matrices. Recall that:

$$\mathcal{A}\mathcal{A}^* = M_{\alpha_1} + 2M_{\alpha_{2a}} + \sqrt{2}M_{\alpha_{2b}} + \sqrt{2}M_{\alpha_{2c}} + dM_{\alpha_{2d}} + 2M_{\alpha_{3a}} + 2\sqrt{2}(d-1)M_{\alpha_{3b}} + (d^2-d)M_{\alpha_{3c}} + \sqrt{24}M_{\alpha_4} + 6\sqrt{2}M_{\alpha_{3b}} + 3dM_{\alpha_{3c}}.$$

where α_1, α_{2a}, α_{2b}, α_{2c}, α_{2d}, α_{3a}, α_{3b}, α_{3c}, and α_4 are the following proper shapes:

1. α_1 is the same as in Section B.2.
2. $U_{\alpha_{2a}} = (u)$ and $V_{\alpha_{2a}} = (v)$ where u, v are circle vertices, $W_{\alpha_{2a}} = \{w\}$ where w is a square vertex, and $E(\alpha_{2a}) = \{\{u,w\}_2, \{w,v\}_2\}$.
3. $U_{\alpha_{2b}} = (u)$ and $V_{\alpha_{2b}} = (v)$ where u, v are circle vertices, $W_{\alpha_{2b}} = \{w\}$ where w is a square vertex, and $E(\alpha_{2b}) = \{\{u,w\}_2\}$.
4. $U_{\alpha_{2c}} = (u)$ and $V_{\alpha_{2c}} = (v)$ where u, v are circle vertices, $W_{\alpha_{2c}} = \{w\}$ where w is a square vertex, and $E(\alpha_{2c}) = \{\{w,v\}_2\}$.
5. $U_{\alpha_{2d}} = (u)$ and $V_{\alpha_{2d}} = (v)$ where u, v are circle vertices, $W_{\alpha_{2d}} = \{\}$, and $E(\alpha_{2d}) = \{\}$.
6. $U_{\alpha_{3a}} = V_{\alpha_{3a}} = (u)$ where u is a circle vertex, $W_{\alpha_{3a}} = \{w_1, w_2\}$ where w_1, w_2 are square vertices, and $E(\alpha_{3a}) = \{\{u,w_1\}_2, \{u,w_2\}_2\}$.
7. $U_{\alpha_{3b}} = V_{\alpha_{3a}} = (u)$ where u is a circle vertex, $W_{\alpha_{3b}} = \{w\}$ where w is a square vertex, and $E(\alpha_{3b}) = \{\{u,w\}_2\}$.
8. $U_{\alpha_{3c}} = V_{\alpha_{3c}} = (u)$ where u is a circle vertex, $W_{\alpha_{3c}} = \{\}$, and $E(\alpha_{3c}) = \{\}$.
9. $U_{\alpha_4} = V_{\alpha_4} = (u)$ where u is a circle vertex, $W_{\alpha_4} = \{w\}$ where w is a square vertex, and $E(\alpha_4) = \{\{u,w\}_4\}$.

Since $M_{\alpha_{2d}} = 1_n 1_n^T - I_n$ and $M_{\alpha_{3c}} = I_n$, we have that:

$$\mathcal{A}\mathcal{A}^* = (d^2+d)Id_n + d1_n1_n^T + M_{\alpha_1} + 2M_{\alpha_{2a}} + \sqrt{2}M_{\alpha_{2b}} + \sqrt{2}M_{\alpha_{2c}} + 2M_{\alpha_{3a}} + (2\sqrt{2}d + 4\sqrt{2})M_{\alpha_{3b}} + \sqrt{24}M_{\alpha_4}.$$

Since $M = (d^2+d)I_n + d1_n1_n^T$,

$$\Delta = M - \mathcal{A}\mathcal{A}^* = -M_{\alpha_1} - 2M_{\alpha_{2a}} - \sqrt{2}M_{\alpha_{2b}} - \sqrt{2}M_{\alpha_{2c}} - 2M_{\alpha_{3a}} - (2\sqrt{2}d + 4\sqrt{2})M_{\alpha_{3b}} - \sqrt{24}M_{\alpha_4}.$$

We now compute $M^{-1}\Delta$. Since $M = (d^2+d)I + d1_n1_n^T$, we have that $M^{-1} = \frac{1}{d^2+d}\left(I - \frac{1}{n+d+1}1_n1_n^T\right)$.

Definition 55 *Define α_J to be the shape with $U_{\alpha_J} = (u)$, $V_{\alpha_J} = (v)$, $W_{\alpha_J} = \emptyset$, and $E(\alpha_J) = \emptyset$.*

Since $1_n1_n^T = I + M_{\alpha_J}$, we have

$$M^{-1} = \frac{1}{d^2+d}\left(\frac{n+d}{n+d+1}I - \frac{1}{n+d+1}M_{\alpha_J}\right).$$

We now compute the product of M_{α_J} with each of the graph matrices appearing in Δ.

1. $M_{\alpha_J} M_{\alpha_1} = M_{\alpha_5} + M_{\alpha_1}$ where M_{α_5} is the shape such that $U_{\alpha_5} = (u)$ and $V_{\alpha_5} = (v)$ where u, v are circle vertices, $W_{\alpha_5} = \{w_\circ, w_1, w_2\}$ where $w_\circ$ is a circle vertex and w_1, w_2 are square vertices, and $E(\alpha_1) = \{\{w_\circ, w_1\}, \{w_\circ, w_2\}, \{w_1, v\}, \{w_2, v\}\}$.

2. $M_{\alpha_J} M_{\alpha_{2a}} = M_{\alpha_{6a}} + M_{\alpha_{2a}}$ where $M_{\alpha_{6a}}$ is the shape such that $U_{\alpha_{6a}} = (u)$ and $V_{\alpha_{6a}} = (v)$ where u, v are circle vertices, $W_{\alpha_{6a}} = \{w_\circ, w\}$ where $w_\circ$ is a circle vertex and w is a square vertex, and $E(\alpha_{6a}) = \{\{w_\circ, w\}_2, \{w, v\}_2\}$.

3. $M_{\alpha_J} M_{\alpha_{2b}} = M_{\alpha_{6b}} + M_{\alpha_{2c}}$ where $M_{\alpha_{6b}}$ is the shape such that $U_{\alpha_{6b}} = (u)$ and $V_{\alpha_{6b}} = (v)$ where u, v are circle vertices, $W_{\alpha_{6b}} = \{w_\circ, w\}$ where $w_\circ$ is a circle vertex and w is a square vertex, and $E(\alpha_{6b}) = \{\{w_\circ, w\}_2\}$.

4. $M_{\alpha_J} M_{\alpha_{2c}} = (n-2) M_{\alpha_{2c}} + M_{\alpha_{2b}}$.

5. $M_{\alpha_J} M_{\alpha_{3a}} = M_{\alpha_{7a}}$ where $M_{\alpha_{7a}}$ is the shape such that $U_{\alpha_{7a}} = (u)$ and $V_{\alpha_{7a}} = (v)$ where u, v are circle vertices, $W_{\alpha_{7a}} = \{w_1, w_2\}$ where w_1, w_2 are square vertices, and $E(\alpha_{7a}) = \{\{v, w_1\}_2, \{v, w_2\}_2\}$.

6. $M_{\alpha_J} M_{\alpha_{3b}} = M_{\alpha_{7b}}$ where $M_{\alpha_{7b}}$ is the shape such that $U_{\alpha_{7b}} = (u)$ and $V_{\alpha_{7b}} = (v)$ where u, v are circle vertices, $W_{\alpha_{7b}} = \{w\}$ where w is a square vertex, and $E(\alpha_{7b}) = \{\{v, w\}_2\}$.

7. $M_{\alpha_J} M_{\alpha_4} = M_{\alpha_8}$ where M_{α_8} is the shape such that $U_{\alpha_8} = (u)$ and $V_{\alpha_8} = (v)$ where u, v are circle vertices, $W_{\alpha_8} = \{w\}$ where w is a square vertex, and $E(\alpha_8) = \{\{v, w\}_4\}$.

Putting everything together, we have that

$$\begin{aligned}
M^{-1}\Delta =& \frac{1}{(d^2+d)(n+d+1)}\Big(-(n+d)M_{\alpha_1} + (M_{\alpha_5} + M_{\alpha_1}) - 2(n+d)M_{\alpha_{2a}} + 2\left(M_{\alpha_{6a}} + M_{\alpha_{2a}}\right) \\
& - \sqrt{2}(n+d)M_{\alpha_{2b}} + \sqrt{2}\left(M_{\alpha_{6b}} + M_{\alpha_{2c}}\right) - \sqrt{2}(n+d)M_{\alpha_{2c}} + \sqrt{2}\left((n-2)M_{\alpha_{2c}} + M_{\alpha_{2b}}\right) \\
& - 2(n+d)M_{\alpha_{3a}} + 2M_{\alpha_{7a}} - (2\sqrt{2}d + 4\sqrt{2})(n+d)M_{\alpha_{3b}} + (2\sqrt{2}d + 4\sqrt{2})M_{\alpha_{7b}} \\
& - \sqrt{24}(n+d)M_{\alpha_4} + \sqrt{24}M_{\alpha_8}\Big) \\
=& \frac{1}{(d^2+d)(n+d+1)}\Big(-(n+d-1)M_{\alpha_1} - 2(n+d-1)M_{\alpha_{2a}} - \sqrt{2}(n+d-1)M_{\alpha_{2b}} \\
& - \sqrt{2}(d-1)M_{\alpha_{2c}} - 2(n+d)M_{\alpha_{3a}} - (2\sqrt{2}d + 4\sqrt{2})(n+d)M_{\alpha_{3b}} - \sqrt{24}(n+d)M_{\alpha_4} \\
& + M_{\alpha_5} + 2M_{\alpha_{6a}} + \sqrt{2}M_{\alpha_{6b}} + 2M_{\alpha_{7a}} + (2\sqrt{2}d + 4\sqrt{2})M_{\alpha_{7b}} + \sqrt{24}M_{\alpha_8}\Big).
\end{aligned}$$

Now that we have obtained a graph matrix decomposition of $M^{-1}\Delta$, we demonstrate why Proposition 5.2 of the previous version of this paper is incorrect. Note that $M^{-1}\Delta$ contains the term

$$-\frac{1}{(d^2+d)(n+d+1)}\sqrt{2}(n+d-1)M_{\alpha_{2b}}.$$

We now make the following observations:

1. $M_{\alpha_{2b}} 1_n = (n-1)M_{\alpha_w}$.

2. M_{α_w} is an $n \times 1$ vector, $\|M_{\alpha_w}\|_2 = \tilde{O}(\sqrt{dn})$, and $\|M_{\alpha_w}\|_\infty = \tilde{O}(\sqrt{d})$.

Putting these observations together, the contribution to $\|M^{-1}\Delta 1_n\|_\infty$ from the $M_{\alpha_{2b}}$ term of $M^{-1}\Delta$ is $\tilde{O}(\frac{n\sqrt{d}}{d^2}) = \tilde{O}(\frac{n}{d^{3/2}})$. It can be checked that the contribution to $\|M^{-1}\Delta 1_n\|_\infty$ from the other terms of $M^{-1}\Delta$ is smaller. Thus, the correct bound is $\left\|M^{-1}\Delta 1_n\right\|_\infty = O(n/d^{3/2})$.

K.2. Proof sketch for repairing the argument

While this proposition is not correct, the three statements needed for this approach to succeed are correct. For convenience, we recall these statements here.

1. $\|M^{-1}\Delta\|_{op} < 1$,
2. $\|\mathcal{A}^*(M^{-1}\Delta 1_n)\|_{op} = o(n)$,
3. For all $j \geq 2$, $\|\mathcal{A}^*((M^{-1}\Delta)^j 1_n)\|_{op} = o(n)$.

To show these statements, we can follow the proof of Lemma 28 to show that in all of the terms which appear in these expressions, the minimum vertex separator consists of one square vertex. We can then use the similar weighting schemes to bound these terms. Two notable cases for w_{actual} are as follows.

1. For α_{2b} where the square appears again to the left, but not to the right, and the edge vanishes, we assign $\sqrt{n}$ to each vertex and 1 to the square (which is an under-assignment). This means that each time α_{2b} appears, we may have a debt for this square. Fortunately, we can use the same ideas as before. In particular, we can pay off this debt by making the edge in α_{2b} a right-critical edge if it vanishes and finding the corresponding left-critical edge.
2. For α_{2c}, because of M^{-1} we have a coefficient of $O(\frac{1}{nd})$ rather than $O(\frac{1}{d^2})$. This allows us to assign weight $\sqrt{n}$ to the two circle vertices and d to the square vertex which is sufficient to handle any debt. Note that this is one of the cases which can have a left-critical edge.

While this approach can be made to work, we use the Woodbury matrix identity approach in the current paper for two reasons. First, it gives a better approximation to $\mathcal{A}\mathcal{A}^*$. Second, it requires less casework and only requires paying off debt for one square, instead of several such squares.