UNIVERSALITY AND SHARP MATRIX CONCENTRATION INEQUALITIES

TATIANA BRAILOVSKAYA AND RAMON VAN HANDEL

ABSTRACT. We show that, under mild assumptions, the spectrum of a sum of independent random matrices is close to that of the Gaussian random matrix whose entries have the same mean and covariance. This nonasymptotic universality principle yields sharp matrix concentration inequalities for general sums of independent random matrices when combined with the Gaussian theory of Bandeira, Boedihardjo, and Van Handel. A key feature of the resulting theory is that it is applicable to a remarkably broad class of random matrix models that may have highly nonhomogeneous and dependent entries, which can be far outside the mean-field situation considered in classical random matrix theory. We illustrate the theory in applications to random graphs, matrix concentration inequalities for smallest singular values, sample covariance matrices, strong asymptotic freeness, and phase transitions in spiked models.

1. INTRODUCTION

1.1. **Matrix concentration inequalities.**
Let $Z_1, \ldots, Z_n$ be independent $d \times d$ random matrices with zero mean, and let

$$X := \sum_{i=1}^{n} Z_i. \tag{1.1}$$

Random matrices of this form arise in numerous applications. As guiding examples, the reader may keep in mind the following very special cases:

- Any random matrix X with centered jointly Gaussian entries may be represented in this form by setting $X = \sum_{i=1}^{n} g_i A_i$ for suitable deterministic matrices A_i, where $g_1, \ldots, g_n$ are i.i.d. standard Gaussian variables.
- Any random matrix X with centered independent entries may be represented in this form as $X = \sum_{i,j=1}^{d} \eta_{ij} e_i e_j^*$, where η_{ij} are independent centered random variables and $e_1, \ldots, e_d$ denotes the standard basis of $\mathbb{C}^d$.

Many other kinds of summands Z_i arise naturally in a diverse range of pure and applied mathematical problems; cf. [67] and the references therein, and the applications that are discussed in sections 1.3 and 3 below.

Already in the special cases highlighted above, it is clear that random matrices of the form (1.1) can possess a nearly arbitrary structure: the model allows for essentially any pattern of entry variances, dependencies, and distributions. Such general models are outside the reach of classical random matrix theory, which is

---

2010 *Mathematics Subject Classification.* 60B20; 60E15; 46L53; 46L54; 15B52.
Key words and phrases. Random matrices; matrix concentration; universality; free probability.

primarily concerned with the asymptotic behavior of highly symmetric models such as matrices with i.i.d. entries or invariant ensembles [5, 65].

Rather surprisingly, one of the most fruitful ideas that has been developed in the present setting is that one can treat the model (1.1) essentially as though it is a sum of independent *scalar* random variables. This approach results in a somewhat crude but extremely versatile family of nonasymptotic *matrix concentration inequalities*. Two important examples of such inequalities are:[1]

- For a self-adjoint random matrix X with centered jointly Gaussian entries, the noncommutative Khintchine inequality of Lust-Piquard and Pisier [58, §9.8] yields

$$\mathbf{E}\|X\| \lesssim \|\mathbf{E}X^2\|^{\frac{1}{2}}\sqrt{\log d}. \tag{1.2}$$

- For a self-adjoint random matrix X of the form (1.1) with $\|Z_i\| \le R$ a.s., the matrix Bernstein inequality of Oliveira and Tropp [54, 67] yields

$$\mathbf{E}\|X\| \lesssim \|\mathbf{E}X^2\|^{\frac{1}{2}}\sqrt{\log d} + R\log d. \tag{1.3}$$

To understand the significance of these inequalities, note that $(\mathbf{E}\|X\|^2)^{\frac{1}{2}} \ge \|\mathbf{E}X^2\|^{\frac{1}{2}}$ by Jensen's inequality. The above bounds can therefore capture the norm of very general random matrices up to a logarithmic dimensional factor. The dimensional factor proves to be suboptimal, however, even for the simplest random matrix models (such as those with i.i.d. entries).

The inefficiency of classical matrix concentration inequalities stems from the fact that by mimicking the proofs of scalar concentration inequalities, these bounds ignore noncommutativity of the summands Z_i in (1.1). In the setting of Gaussian random matrices, a significant step toward addressing this inefficiency was recently made by Bandeira, Boedihardjo, and the second author [9], who developed a powerful new class of *sharp* matrix concentration inequalities that capture noncommutativity. For example, if X is a self-adjoint random matrix with centered jointly Gaussian entries, [9, Corollary 2.2] yields

$$\mathbf{E}\|X\| \le \|X_{\mathrm{free}}\| + C\|\mathbf{E}X^2\|^{\frac{1}{4}}\|\mathrm{Cov}(X)\|^{\frac{1}{4}}(\log d)^{\frac{3}{4}} \tag{1.4}$$

for a universal constant C. Here $\mathrm{Cov}(X)$ denotes the $d^2\times d^2$ covariance matrix of the entries of X, while X_{free} is a certain noncommutative model of X that arises from free probability theory. As $\|X_{\mathrm{free}}\| \le 2\|\mathbf{E}X^2\|^{\frac{1}{2}}$, the inequality (1.4) shows that the dimensional factor in (1.2) can be removed as soon as $\|\mathrm{Cov}(X)\| \ll (\log d)^{-3}\|\mathbf{E}X^2\|$, which is a mild assumption in many applications. The theory of [9] yields much more, however: both the support and the empirical distribution of the spectrum of X is close to that of X_{free}, and similar results hold for polynomials of such matrices. Such results open the door to developing a nonasymptotic random matrix theory for nearly arbitrarily structured random matrices.

In view of these developments, it is of considerable interest to extend the Gaussian theory of [9] to the much more general setting (1.1) of sums of independent random matrices. For classical matrix concentration inequalities, this extension has been achieved in two distinct ways: one may either derive both (1.2) and (1.3) by a common method of proof [54, 67], or deduce (1.3) from (1.2) by a symmetrization argument as in [61, 68]. Unfortunately, neither of these approaches appears to

[1] Here and in the sequel, $\|M\|$ denotes the operator norm (i.e., the largest singular value) of a matrix M, and $a \lesssim b$ denotes $a \le Cb$ for a universal constant C.

give rise to a satisfactory extension of the theory of [9]. The methods of [9] rely heavily on Gaussian analysis, and it is unclear how to adapt them to non-Gaussian situations. On the other hand, sharp inequalities are fundamentally inaccessible by symmetrization, as is explained in [9, §8.2.2].

1.2. **Universality.** In this paper, we take an entirely different viewpoint on such problems. To motivate the form of our main results, let us note that if the last term in the inequality (1.3) is negligible, then (1.3) has exactly the same form as the Gaussian inequality (1.2). The main theme of this paper is that this phenomenon has nothing to do with matrix concentration inequalities themselves, but is rather a consequence of a general *universality principle*:

> *If* $\max_{1\le i\le n}\|Z_i\| \ll \|\mathbf{E}X^2\|^{\frac{1}{2}}(\log d)^{-\beta}$ *(for an appropriate* $\beta>0$*), then the spectrum of a self-adjoint random matrix* $X=\sum_{i=1}^n Z_i$ *as in* (1.1) *nearly coincides with that of the Gaussian random matrix* G *whose entries have the same mean and covariance as* X*.*

This principle directly reduces the study of the spectrum of sums of independent random matrices to that of Gaussian matrices, regardless of what theory is applied to the Gaussian matrices. In particular, it simultaneously explains the phenomenon behind (1.2)—(1.3), and enables us to fully extend the sharp matrix concentration theory of [9] to the model (1.1).

The universality principle was stated above in an informal manner. A detailed formulation of our results will be given in section 2 below. In particular, we will obtain nonasymptotic inequalities that establish closeness both of the spectral distributions of X and G, and of the spectra themselves in Hausdorff distance. (These results apply in a more general setting than (1.1), where the random matrices may have an arbitrary mean.) We further formulate resulting sharp matrix concentration inequalities that arise from the theory of [9].

Universality phenomena have been widely investigated in classical random matrix theory. As in many previous works on this topic, the starting point for our analysis is the cumulant expansion of Barbour [14] and Lytova and Pastur [47], which has been primarily applied to classical random matrix models with independent entries. A rather complicated extension of the cumulant expansion to dependent models appears in [36], where it is used to study random matrices whose entries exhibit decay of correlations. A straightforward extension of Barbour's method to the dependent setting will be formulated in section 4; such an extension does not in itself require any new idea as compared to [14, 47].

The core contribution of this paper lies in the mechanism that gives rise to universality. To the best of our knowledge, prior universality results are essentially limited to the "classical random matrix regime" where the entry variances are of order $d^{-\frac{1}{2}}$ and the entries are independent or nearly independent (in the sense that they exhibit decay of correlations). In other words, these results rely on restrictive mean-field assumptions. In contrast, the independent sum models (1.1) of the present paper can lie far outside the mean-field regime: they can be highly nonhomogeneous, sparse, and exhibit strong dependence among the entries, and are not assumed to possess any special structure or symmetries. The properties of these models therefore cannot be explained by previous universality results that essentially mimic the behavior of independent entry models. The central idea of this paper is that universality arises in these models in a different manner through an operator-theoretic mechanism: a key ingredient of our approach are high-order trace inequalities (cf. section 5) that enable us to control the contributions of the terms in the cumulant expansion without imposing any correlation decay or mean-field assumptions. This operator-theoretic viewpoint on universality, together with a number of other new tools that are developed throughout this paper (such as nonstandard concentration inequalities for spectral statistics), provides access to many applications that are not captured by classical random matrix models.

Remark 1.1. As is common in probability theory (see, e.g., [29]), we use the term "universality" to denote insensitivity of various model statistics (e.g., the support and distribution of the spectrum) to the distribution of the matrix entries. This should not be confused with the much narrower notion of "Wigner universality", which posits that the local statistics of certain random matrix models behave as those of Wigner matrices. The latter is false even for the simplest nonhomogeneous models, see, e.g., [63, 25]. Given the wide variety of possible behaviors of nonhomogeneous random matrices at the local scale, Wigner universality does not appear to be a meaningful question in the general setting of this paper.

1.3. **Applications.** To illustrate the versatility of the main results of this paper, we will develop several applications that we briefly describe here (see section 3 for detailed statements). Beyond their independent interest, we emphasize that the completely general universality phenomenon described by our main results is the common mechanism underlying all these rather diverse applications.

Random graphs and expanders. The expansion properties of random regular graphs have been extensively studied for graphs of bounded degree k. In particular, such graphs are nearly Ramanujan, i.e., they have the smallest possible (by [53]) second eigenvalue $\lambda_2 = (1+o(1))2\sqrt{k-1}$ to leading order [38]. While the strong expansion properties of such graphs are expected to persist when the degree is allowed to diverge, this situation remains much more poorly understood; see, e.g., the survey [74]. The universality principles of this paper enable us to address this question both in classical and in new situations:

- The permutation model of random regular graphs with n vertices of degree k is nearly Ramanujan when $k \gg (\log n)^4$, addressing a well known question [15, §1.4]. To date, the best known bound in this setting was $\lambda_2 = O(\sqrt{k})$ [39, 34, 31].
- A classical result of Alon and Roichman [2] states that if Γ is any finite group and $k \gg \log|\Gamma|$ generators are chosen uniformly at random, the resulting Cayley graph is an expander. This result cannot be improved for abelian groups. Here we show that under mild assumptions that hold, e.g., for all nonabelian finite simple groups, the Cayley graph defined by choosing $k \gg (\log|\Gamma|)^4$ random generators is nearly Ramanujan. This appears to be the first result of its kind.
- A fundamental result of Bordenave and Collins [19] states that for any fixed base graph H, the new eigenvalues of its random n-lift are bounded as $n \to \infty$ by the spectral radius of the universal cover of H. Here we show that this conclusion remains valid for any sequence of base graphs H_n whose maximal degrees grow at least polylogarithmically in the number of vertices of their random lifts. Moreover, in this setting we uncover a new phenomenon: when the base graphs are simple, random 2-lifts already achieve the optimal bound.

Matrix concentration inequalities for smallest singular values. By their nature, classical matrix concentration inequalities can only control the largest singular value of nonhomogeneous random matrices. In contrast, the universality principles of this paper apply not only to the largest singular value but also to the entire spectrum. By combining our results with the Gaussian theory of [9], we are therefore able to obtain sharp matrix concentration inequalities for the smallest singular value of random matrices of the form (1.1) that may be viewed as a nonasymptotic, nonhomogeneous form of the classical Bai-Yin law [7]. Let us emphasize that even if one is interested in suboptimal bounds on the smallest singular value, such information is fundamentally inacceassible by the methods used to prove classical matrix concentration inequalities for general models of the form (1.1).

A direct application yields bounds for the smallest singular value of sparse nonhomogeneous bipartite Erdős-Rényi graphs that are sharp to leading order. To date, the best known bounds [35] were suboptimal for nonhomogeneous graphs.

Sample covariance matrices. Let $Y_1, \ldots, Y_n$ be independent, centered random vectors in $\mathbb{R}^d$. The $d \times d$ random matrix defined by $S := \sum_{i=1}^n Y_i Y_i^*$ is called the (nonhomogeneous) sample covariance matrix. Equivalently, $S = YY^*$, where $Y := \sum_{i=1}^n Y_i e_i^*$ is the $d \times n$ matrix whose columns are $Y_1, \ldots, Y_n$.

A central problem in this setting is to control the deviation of the sample covariance matrix from its mean $\|S - \mathbf{E}S\|$. A curious feature of this problem is that we may express S in terms of a model of the form (1.1) in two different ways: we may either consider S itself as a model of the form (1.1), or we may consider Y as a model of the form (1.1). These two representations give rise to distinct universality principles: roughly speaking, applying our universality principles to S is efficient when n is sufficiently large compared to d, while applying universality to Y is efficient when d is sufficiently large compared to n.

We will illustrate this phenomenon in the setting of nonhomogeneous Gaussian sample covariance matrices with arbitrary covariance matrices of $Y_1, \ldots, Y_n$, for which we obtain nonasymptotic bounds on $\|S - \mathbf{E}S\|$ that are sharp for a wide range of parameters. No sharp bounds appear to be known in the literature at this level of generality. We also discuss non-Gaussian sample covariance matrices, which will be developed further in forthcoming work [55].

Strong asymptotic freeness. A celebrated result of Voiculescu [73] states that the traces of polynomials of independent $N \times N$ Wigner matrices converge as $N \to \infty$ to the traces of polynomials of certain limiting objects that arise in free probability theory. In an important breakthrough, Haagerup and Thorbjørnsen [41] showed that this convergence holds not only for the trace but also for the norm:

$$\lim_{N\to\infty} \|p(X_1^N, \ldots, X_m^N)\| = \|p(s_1, \ldots, s_m)\| \quad \text{a.s.} \tag{1.5}$$

for every noncommutative polynomial p, where $X_1^N, \ldots, X_m^N$ are independent $N \times N$ complex Gaussian Wigner matrices and $s_1, \ldots, s_m$ is a free semicircular family. This property, called strong asymptotic freeness, is of fundamental importance both to random matrices and in the theory of operator algebras.

Whether (1.5) holds for more general models of random matrices X_i^N is far from clear from the original rather delicate proofs. Previously, the state-of-the-art [4] was that (1.5) holds for matrices with i.i.d. centered entries with unit variance and bounded fourth moment. Very recently, however, the sharp matrix concentration theory of [9] made it possible to establish (1.5) for an extremely general class of Gaussian random matrices, showing that this phenomenon is much more ubiquitous than was previously understood. Our universality principles extend this conclusion even further to general non-Gaussian random matrices of the form (1.1) under surprisingly mild assumptions that allow for significant sparsity and dependence. This will imply, for example, that $N \times N$ random matrices can achieve strong asymptotic freeness using only $O(N \log^5 N)$ bits of randomness.

Phase transitions in spiked models. The behavior of low-rank perturbations of random matrices (so-called "spiked" models) has attracted much attention in pure and applied random matrix theory since the work of Baik, Ben Arous and Péché [8]. The characteristic feature of such models is that they exhibit a phase transition depending on the size of the perturbation: there is a threshold above which one or more isolated eigenvalues detach from the bulk of the spectrum. Most of the literature on this topic is concerned with Wigner matrices or with unitarily invariant models; see, e.g., the survey [24]. The universality principles of this paper enable us to investigate such phenomena in much more general situations, including models that exhibit significant sparsity and dependence. Little appears to be known in this direction: previous work on a special type of low-rank perturbations of sparse random matrices appeared only very recently in [64].

As our primary aim here is to illustrate the main results of this paper, we will focus our attention on sparse and dependent models whose behavior can be reduced by universality to the classical spiked Wigner model. In this setting, we will show how our universality principles enable us to capture the number and locations of the outlier eigenvalues, as well as the overlaps of the associated eigenvectors with those of the low-rank perturbation. However, much more general situations become amenable to analysis in combination with the Gaussian theory of [9], which makes it possible to investigate analogous phase transition phenomena in nonhomogeneous models. The computations involved in the nonhomogenous setting are unrelated to universality, and will be treated in forthcoming work [10].

1.4. **Organization of this paper.** The remainder of this paper is organized as follows. In section 2, we formulate the main results of this paper. Section 3 is devoted to a detailed formulation of the applications described above.

In section 4, we provide a brief self-contained treatment of the multivariate cumulant expansion. Section 5 develops some key tools that are used in the proofs of our main results: high-order trace inequalities that provide the main mechanism for controlling the terms in the cumulant expansion, and certain nonstandard concentration of measure inequalities. The following three sections are devoted to the proofs of our main results. Section 6 proves the universality principles for spectral statistics, while section 7 proves the universality principle for the support of the spectrum. Section 8 is devoted to a truncation argument that extends our main results to models that satisfy minimal moment assumptions. Finally, section 9 is devoted to the proofs of the various applications discussed in section 3.

1.5. **Notation.** The following notation will be used throughout the paper. We write $[n] := \{1, \ldots, n\}$ for $n \in \mathbb{N}$. The algebra of $d \times d$ matrices with values in a *-algebra $\mathcal{A}$ is denoted as $\mathrm{M}_d(\mathcal{A})$, and its subspace of self-adjoint matrices is denoted as $\mathrm{M}_d(\mathcal{A})_{\mathrm{sa}}$. For a matrix or operator X, we denote by $\|X\|$ its operator norm, by $\mathrm{sp}(X)$ its spectrum, and by $|X| := (X^*X)^{\frac{1}{2}}$. The identity matrix or operator is denoted as $\mathbf{1}$. For $M \in \mathrm{M}_d(\mathbb{C})$, we denote by $\mathrm{Tr}\, M := \sum_{i=1}^d M_{ii}$ the unnormalized trace and by $\mathrm{tr}\, M := \frac{1}{d}\,\mathrm{Tr}\, M$ the normalized trace.

2. Main results

2.1. Random matrix models and matrix parameters.

2.1.1. *The general model.* The basic random matrix model of this paper is defined as follows. Fix $d \geq 2$ and $n \in \mathbb{N}$, let $Z_0 \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ be any deterministic $d \times d$ self-adjoint matrix, and let $Z_1, \ldots, Z_n$ be any $d \times d$ self-adjoint random matrices with zero mean $\mathbf{E}[Z_i] = 0$ and complex-valued entries. We define

$$X := Z_0 + \sum_{i=1}^n Z_i. \tag{2.1}$$

Note that this model is slightly more general than the model (1.1) discussed in the introduction, in that we allow for an arbitrary mean.

Remark 2.1. The assumption that X is self-adjoint is made primarily for notational convenience. Our main results extend directly to non-self-adjoint matrices as follows. For any matrix $M \in \mathrm{M}_d(\mathbb{C})$, define its dilation $\breve{M} \in \mathrm{M}_{2d}(\mathbb{C})_{\mathrm{sa}}$ as

$$\breve{M} := \begin{bmatrix} 0 & M \\ M^* & 0 \end{bmatrix}.$$

If we denote by $M = U|M|$ the polar decomposition of M, it follows that

$$\breve{M} = V \begin{bmatrix} -|M| & 0 \\ 0 & |M| \end{bmatrix} V^* \qquad \text{with} \qquad V := \frac{1}{\sqrt{2}} \begin{bmatrix} U & U \\ -\mathbf{1} & \mathbf{1} \end{bmatrix},$$

where $|M| := (M^*M)^{\frac{1}{2}}$. As V is unitary, this shows that the eigenvalues of $\breve{M}$ coincide precisely (including multiplicities) with $\{\pm\sigma_i : i \in [d]\}$, where $\sigma_1, \ldots, \sigma_d$ are the singular values of M. Consequently, by applying our results to $\breve{X}$, we can immediately extend their conclusions on the eigenvalues of self-adjoint random matrices to the singular values of non-self-adjoint random matrices. For further comments on the non-self-adjoint case, see [9, Remark 2.6] and Corollary 2.15.

Associated with the random matrix X are two models that capture its structure in an idealized manner. We introduce these models presently.

2.1.2. *The Gaussian model.* Throughout this paper, we denote by G the Gaussian model that has the same mean and covariance structure as X. More precisely, denote by $\mathrm{Cov}(X)$ the $d^2 \times d^2$ covariance matrix of the entries of X, that is,

$$\mathrm{Cov}(X)_{ij,kl} := \mathbf{E}[(X - \mathbf{E}X)_{ij}\overline{(X - \mathbf{E}X)_{kl}}]. \tag{2.2}$$

We define G to be the $d \times d$ self-adjoint random matrix such that:

1. $\{\mathrm{Re}\, G_{ij}, \mathrm{Im}\, G_{ij} : i, j \in [d]\}$ are jointly Gaussian;
2. $\mathbf{E}[G] = \mathbf{E}[X]$ and $\mathrm{Cov}(G) = \mathrm{Cov}(X)$.

Note that as G is a self-adjoint matrix $G_{lk} = \overline{G_{kl}}$, the covariance matrix of the real-valued Gaussian vector $\{\mathrm{Re}\, G_{ij}, \mathrm{Im}\, G_{ij} : i, j \in [d]\}$ is fully specified by $\mathrm{Cov}(G)$. Thus the above properties uniquely define the distribution of G.

2.1.3. *The noncommutative model.* We now introduce a noncommutative model $X_{\rm free}$ that has the same mean and covariance structure as X. To this end, we must recall some basic notions from free probability theory; we refer to [52] for precise definitions and a comprehensive treatment.

Fix a C^*-probability space $(\mathcal{A},\tau)$, that is, a unital C^*-algebra $\mathcal{A}$ endowed with a faithful trace τ. The following may be viewed as a noncommutative analogue of jointly Gaussian variables with mean μ and covariance C, cf. [52, p. 128].

Definition 2.2. A family of self-adjoint elements $s_1,\ldots,s_m\in\mathcal{A}$ is said to be a *semicircular family with mean μ and covariance C* if

$$\tau(s_k)=\mu_k,\qquad \tau((s_{k_1}-\mu_{k_1}\mathbf{1})\cdots(s_{k_p}-\mu_{k_p}\mathbf{1}))=\sum_{\pi\in\mathrm{NC}_2([p])}\prod_{\{i,j\}\in\pi}C_{k_ik_j}$$

for all $p\ge 1$ and $k,k_1,\ldots,k_p\in[m]$, where $\mathrm{NC}_2([p])$ denotes the collection of noncrossing pair partitions of $[p]$.

A $d\times d$ matrix $Y\in\mathrm{M}_d(\mathcal{A})$ with $\mathcal{A}$-valued entries is naturally identified with an element of the C^*-algebra $\mathrm{M}_d(\mathbb{C})\otimes\mathcal{A}$, which we endow with the normalized trace $\mathrm{tr}\otimes\tau$, cf. [50, Chapter 9]. Define the entry covariance matrix $\mathrm{Cov}(Y)$ as

$$\mathrm{Cov}(Y)_{ij,kl}:=\tau((Y_{ij}-\tau(Y_{ij})\mathbf{1})(Y_{kl}-\tau(Y_{kl})\mathbf{1})^*).$$

With these definitions in place, we can now define $X_{\rm free}\in\mathrm{M}_d(\mathcal{A})_{\rm sa}$ as follows:

1. $\{\mathrm{Re}\,(X_{\rm free})_{ij},\mathrm{Im}\,(X_{\rm free})_{ij}:i,j\in[d]\}$ is a semicircular family;
2. $(\mathrm{id}\otimes\tau)(X_{\rm free})=\mathbf{E}[X]$ and $\mathrm{Cov}(X_{\rm free})=\mathrm{Cov}(X)$.

Here we write $\mathrm{Re}\,a:=\frac{1}{2}(a+a^*)$ and $\mathrm{Im}\,a:=\frac{1}{2i}(a-a^*)$ for $a\in\mathcal{A}$.

Remark 2.3. As jointly Gaussian variables can always be written as linear combinations of independent standard Gaussian variables, G may be expressed as

$$G=Z_0+\sum_{i=1}^N A_ig_i$$

for some deterministic matrices $A_1,\ldots,A_N\in\mathrm{M}_d(\mathbb{C})_{\rm sa}$ and i.i.d. (real) standard Gaussians $g_1,\ldots,g_N$ (note that this representation is not unique). Given any such a representation, it is readily verified that one may express $X_{\rm free}$ as

$$X_{\rm free}=Z_0\otimes\mathbf{1}+\sum_{i=1}^N A_i\otimes s_i,$$

where $s_1,\ldots,s_N$ is a *free* semicircular family, that is, with zero mean and identity covariance matrix. Thus the present definition of $X_{\rm free}$ agrees with the one in [9].

2.1.4. *Matrix parameters.* Let X be a self-adjoint random matrix as in (2.1). The following basic parameters will appear in our main results:

$$\sigma(X):=\big\|\mathbf{E}[(X-\mathbf{E}X)^2]\big\|^{\frac{1}{2}},\tag{2.3}$$

$$\sigma_*(X):=\sup_{\|v\|=\|w\|=1}\mathbf{E}\big[|\langle v,(X-\mathbf{E}X)w\rangle|^2\big]^{\frac{1}{2}},\tag{2.4}$$

$$v(X):=\|\mathrm{Cov}(X)\|^{\frac{1}{2}},\tag{2.5}$$

$$R(X):=\Big\|\max_{1\le i\le n}\|Z_i\|\Big\|_\infty.\tag{2.6}$$

The significance of these parameters may be summarized as follows. The parameter $\sigma(X)$ roughly captures the spread of the spectrum of $X - \mathbf{E}X$, as was explained in the introduction. The parameter $\sigma_*(X)$ controls the fluctuations of the spectral statistics of X and G, see, e.g., section 5.2 below. The parameter $v(X)$ quantifies the degree to which the spectral properties of G are captured by those of X_{free}: this is the main outcome of the theory of [9]. Finally, the universality principles of this paper will show that the parameter $R(X)$ quantifies the degree to which the spectral properties of X are captured by those of G.

The parameter $R(X)$ is meaningful only when the random matrices Z_i are uniformly bounded. We will also prove versions of our main results that apply to unbounded summands under minimal moment assumptions. The formulation of these results requires the following modified matrix parameters:

$$\sigma_q(X) := \big(\operatorname{tr}\mathbf{E}[(X-\mathbf{E}X)^2]^{\frac{q}{2}}\big)^{\frac{1}{q}}, \tag{2.7}$$

$$R_q(X) := \big(\textstyle\sum_{i=1}^n \mathbf{E}[\operatorname{tr}|Z_i|^q]\big)^{\frac{1}{q}}, \tag{2.8}$$

$$\bar{R}(X) := \mathbf{E}\Big[\max_{1\le i\le n}\|Z_i\|^2\Big]^{\frac{1}{2}} \tag{2.9}$$

for $q < \infty$, and $\sigma_\infty(X) := \sigma(X)$, $R_\infty(X) := R(X)$.

Remark 2.4. Let us emphasize the following basic facts.

- All these parameters depend only on $X - \mathbf{E}X$, i.e., they do not depend on Z_0.
- $\sigma(X), \sigma_q(X), \sigma_*(X), v(X)$ only depend on the covariance of the entries of X, and therefore capture the universal behavior that is shared between X, G, and X_{free}. In contrast, $R(X), R_q(X), \bar{R}(X)$ are specific to the non-Gaussian model.
- Recall the basic inequalities $\sigma_*(X) \le \sigma(X)$ and $\sigma_*(X) \le v(X)$ [9, §2.1]. In most (but not all) applications of our theory, we have $\sigma_*(X), v(X), R(X) \ll \sigma(X)$.

2.2. **Universality.** We now provide precise formulations of the universality principle. We prove several results that capture different aspects of the spectrum.

2.2.1. *Universality of the spectrum.* In this section, we formulate the universality principle for the spectrum $\mathrm{sp}(X)$ of X. Recall that the Hausdorff distance between two subsets $A, B \subseteq \mathbb{R}$ of the real line is defined as

$$\mathrm{d_H}(A,B) := \inf\{\varepsilon > 0 : A \subseteq B + [-\varepsilon,\varepsilon] \text{ and } B \subseteq A + [-\varepsilon,\varepsilon]\}.$$

Our main result is the following.

Theorem 2.5 (Spectrum universality)**.** *For any $t \ge 0$, we have*

$$\mathbf{P}\big[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(G)) > C\varepsilon(t)\big] \le de^{-t},$$

where C is a universal constant and

$$\varepsilon(t) = \sigma_*(X)\,t^{\frac{1}{2}} + R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}t^{\frac{2}{3}} + R(X)\,t.$$

Moreover,

$$\mathbf{E}\big[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(G))\big] \lesssim \sigma_*(X)\,(\log d)^{\frac{1}{2}} + R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}(\log d)^{\frac{2}{3}} + R(X)\log d.$$

Note that while we defined the distributions of X and G in section 2.1, we did not specify their joint distribution. However, the conclusion of Theorem 2.5 is valid regardless of how X and G are defined on the same probability space due to the strong concentration properties of random matrices.

Theorem 2.5 readily yields a universality principle for the spectral edge. In the following, we denote by $\lambda_{\max}(X) := \sup \mathrm{sp}(X)$ the upper edge of the spectrum. (Inequalities for the lower edge follow readily as $\inf \mathrm{sp}(X) = -\lambda_{\max}(-X)$.)

Corollary 2.6 (Edge universality). *For any $t \geq 0$, we have*

$$\mathbf{P}\big[|\lambda_{\max}(X) - \lambda_{\max}(G)| > C\varepsilon(t)\big] \leq de^{-t},$$

as well as

$$\mathbf{P}\big[|\lambda_{\max}(X) - \mathbf{E}\lambda_{\max}(G)| > C\varepsilon(t)\big] \leq de^{-t},$$

where C is a universal constant and $\varepsilon(t)$ is as in Theorem 2.5. Moreover,

$$|\mathbf{E}\lambda_{\max}(X) - \mathbf{E}\lambda_{\max}(G)| \lesssim \sigma_*(X)\,(\log d)^{\frac{1}{2}} + R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}(\log d)^{\frac{2}{3}} + R(X)\log d.$$

The same bounds hold if $\lambda_{\max}(X), \lambda_{\max}(G)$ are replaced by $\|X\|, \|G\|$, respectively.

The proofs of Theorem 2.5 and Corollary 2.6 are given in section 7.

The above results are meaningful only when $R(X) < \infty$, which requires that the matrices Z_i are uniformly bounded. However, this restriction is almost entirely removed by the following result that is proved in section 8.

Theorem 2.7 (Spectrum universality: unbounded case). *We have*

$$\mathbf{P}\Big[\mathrm{d_H}(\mathrm{sp}(X), \mathrm{sp}(G)) > C\varepsilon_R(t),\ \max_{1\leq i\leq n}\|Z_i\| \leq R\Big] \leq de^{-t}$$

for all $t \geq 0$ and $R \geq \bar{R}(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}} + 2^{\frac{1}{2}}\bar{R}(X)$, where

$$\varepsilon_R(t) = \sigma_*(X)\,t^{\frac{1}{2}} + R^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}t^{\frac{2}{3}} + Rt$$

and C is a universal constant. Moreover,

$$\mathbf{E}\big[\mathrm{d_H}(\mathrm{sp}(X), \mathrm{sp}(G))\big] \lesssim \sigma_*(X)\,(\log d)^{\frac{1}{2}} + \bar{R}(X)^{\frac{1}{6}}\sigma(X)^{\frac{5}{6}}\log d$$

whenever $\bar{R}(X)\,(\log d)^3 \lesssim \sigma(X)$.

2.2.2. *Universality of spectral statistics.* We now complement the above results by formulating universality principles for various spectral statistics.

We begin by establishing universality of even moments.

Theorem 2.8 (Moment universality). *For any $p \in \mathbb{N}$ and $2p \leq q \leq \infty$, we have*

$$\big|\mathbf{E}[\mathrm{tr}\,X^{2p}]^{\frac{1}{2p}} - \mathbf{E}[\mathrm{tr}\,G^{2p}]^{\frac{1}{2p}}\big| \lesssim R_q(X)^{\frac{1}{3}}\sigma_q(X)^{\frac{2}{3}}p^{\frac{2}{3}} + R_q(X)\,p$$

as well as

$$\big|\mathbf{E}[\mathrm{tr}\,X^{2p}]^{\frac{1}{2p}} - \mathbf{E}[\mathrm{tr}\,G^{2p}]^{\frac{1}{2p}}\big| \lesssim R_q(X)\,p^2.$$

The first inequality has a better dependence on p, while the second inequality yields a sharper estimate when $R_q(X)\,p^2 \ll \sigma_q(X)$. Both inequalities are variations of the same proof, which is given in section 6.

Moment bounds provide limited information on the spectrum of a random matrix. Complementary information can be extracted from the resolvent.

Theorem 2.9 (Resolvent universality). *We have*

$$\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}]-\mathbf{E}[(z\mathbf{1}-G)^{-1}]\big\| \lesssim \frac{R(X)\sigma(X)^2+R(X)^3\log d}{(\operatorname{Im} z)^4}$$

for every $z\in\mathbb{C}$ *with* $\operatorname{Im} z>0$. *Consequently,*

$$\big\|\mathbf{E}[\varphi(X)]-\mathbf{E}[\varphi(G)]\big\| \lesssim \big(R(X)\sigma(X)^2+R(X)^3\log d\big)\|\varphi\|_{W^{5,1}(\mathbb{R})}$$

for every $\varphi\in W^{5,1}(\mathbb{R})$.

We can also generalize Theorem 2.9 to unbounded random matrices, at the expense of somewhat worse quantitative error bounds.

Theorem 2.10 (Resolvent universality: unbounded case). *We have*

$$\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}]-\mathbf{E}[(z\mathbf{1}-G)^{-1}]\big\| \lesssim \frac{\sigma_*(X)+\bar{R}(X)^{\frac{1}{10}}\sigma(X)^{\frac{9}{10}}}{(\operatorname{Im} z)^2}$$

for every $z\in\mathbb{C}$ *with* $\operatorname{Im} z>0$ *and*

$$\big\|\mathbf{E}[\varphi(X)]-\mathbf{E}[\varphi(G)]\big\| \lesssim \big(\sigma_*(X)+\bar{R}(X)^{\frac{1}{10}}\sigma(X)^{\frac{9}{10}}\big)\|\varphi\|_{W^{3,1}(\mathbb{R})}$$

for every $\varphi\in W^{3,1}(\mathbb{R})$, *provided that* $\bar{R}(X)(\log d)^{\frac{5}{3}}\lesssim\sigma(X)$.

The proofs of Theorems 2.9 and 2.10 are given in sections 6 and 8.

While the above universality principles suffice for many applications, our proofs can be readily adapted to the study of other spectral statistics. In particular, a universality principle for moments of the resolvent $\mathbf{E}[\operatorname{tr}|z\mathbf{1}-X|^{-2p}]$ (Theorem 6.8) will play a central role in the proof of Theorem 2.5. Let us also note that the quantitative bounds of Theorems 2.9 and 2.10 can be considerably improved if one is interested only in the trace of the resolvent, cf. Remark 6.13. However, the norm bounds given here are particularly useful as they contain information on the eigenvectors of the random matrices, as will be explained in section 3.5.

Remark 2.11. For simplicity, we formulated the results of this section only for expected spectral statistics. However, corresponding tail bounds follow by combining these bounds with concentration inequalities for the relevant spectral statistics; cf. Lemma 9.20 for moments, and Proposition 5.10 for general spectral statistics.

2.3. **Matrix concentration inequalities.** Universality principles show that a non-Gaussian random matrix X behaves as a Gaussian random matrix G, but do not explain in themselves what the spectra of these matrices look like. To apply these results to specific models, our universality principles must be combined with suitable bounds for Gaussian random matrices. We will presently show that both classical matrix concentration inequalities, and new sharp matrix concentration inequalities, arise directly from our main results.

2.3.1. *Classical matrix concentration inequalities.* We begin by briefly illustrating how two classical matrix concentration inequalities can be recovered from our main results. While direct proofs of these inequalities [67, 48] are considerably simpler (and yield better numerical constants), this provides a new explanation for the form of these inqualities and serves as the simplest illustration of our results.

Example 2.12 (Matrix Bernstein). As $\sigma_*(X) \le \sigma(X)$ and as $R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}t^{\frac{2}{3}} \lesssim \sigma(X)\,t^{\frac{1}{2}} + R(X)\,t$ by Young's inequality, Corollary 2.6 implies

$$\mathbf{E}\|X\| \lesssim \mathbf{E}\|G\| + \sigma(X)\sqrt{\log d} + R(X)\log d.$$

We may therefore view the matrix Bernstein inequality (1.3) as a direct consequence of the Gaussian bound (1.2) and Corollary 2.6. The tail bound of [67, Theorem 6.1.1] can also easily be recovered up to universal constants from Corollary 2.6.

Example 2.13 (Matrix Rosenthal). Suppose that $\mathbf{E}[X]=0$. Then the noncommutative Khintchine inequality [58, §9.8] states that for every $p \in \mathbb{N}$, we have

$$\mathbf{E}[\operatorname{tr} G^{2p}]^{\frac{1}{2p}} \lesssim \sigma_{2p}(X)\sqrt{p}$$

(the norm bound (1.2) follows directly from this estimate by choosing $p \sim \log d$). Combining the noncommutative Khintchine inequality with Theorem 2.8 yields

$$\mathbf{E}[\operatorname{tr} X^{2p}]^{\frac{1}{2p}} \lesssim \sigma_{2p}(X)\sqrt{p} + R_{2p}(X)\,p,$$

where we used that $R_{2p}(X)^{\frac{1}{3}}\sigma_{2p}(X)^{\frac{2}{3}}p^{\frac{2}{3}} \lesssim \sigma_{2p}(X)\sqrt{p} + R_{2p}(X)\,p$ by Young's inequality. This matrix Rosenthal inequality [48, Corollary 7.4] may therefore be viewed as another consequence of the universality principle.

2.3.2. *Sharp matrix concentration inequalities.* A primary motivation behind our universality principles is that they may be combined with the Gaussian theory of [9] to obtain a powerful new family of sharp matrix concentration inequalities for sums of independent random matrices. These inequalities reduce the study of a very large family of nonhomogeneous random matrices to explicit computations. Let us state a prototypical inequality of this kind for sake of illustration.

Theorem 2.14 (Sharp matrix concentration). *For any $t \ge 0$, we have*

$$\mathbf{P}\big[\mathrm{sp}(X) \subseteq \mathrm{sp}(X_{\mathrm{free}}) + C\big\{v(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}(\log d)^{\frac{3}{4}} + \varepsilon(t)\big\}[-1,1]\big] \ge 1 - 2de^{-t},$$

where C is a universal constant and $\varepsilon(t)$ is as in Theorem 2.5. In particular,

$$\mathbf{P}\big[\lambda_{\max}(X) \ge \lambda_{\max}(X_{\mathrm{free}}) + Cv(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}(\log d)^{\frac{3}{4}} + C\varepsilon(t)\big] \le 2de^{-t}$$

and

$$\mathbf{E}\lambda_{\max}(X) \le \lambda_{\max}(X_{\mathrm{free}}) + \\ C\big\{v(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}(\log d)^{\frac{3}{4}} + R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}(\log d)^{\frac{2}{3}} + R(X)\log d\big\}.$$

The same bounds hold if $\lambda_{\max}(X), \lambda_{\max}(X_{\mathrm{free}})$ are replaced by $\|X\|, \|X_{\mathrm{free}}\|$.

Proof. This follows immediately by the union bound from Theorem 2.5 and an application of [9, Theorem 2.1] to the Gaussian matrix G. □

Theorem 2.14 shows that when $v(X)$ and $R(X)$ are sufficiently small, the spectrum of X is controlled by that of its noncommutative model X_{free}. The latter admits explicit computations using tools of free probability. For example, it was shown by Lehner [46, Corollary 1.5] (cf. [9, §4.1]) that

$$\lambda_{\max}(X_{\mathrm{free}}) = \inf_{B>0} \lambda_{\max}\big(B^{-1} + \mathbf{E}X + \mathbf{E}[(X-\mathbf{E}X)B(X-\mathbf{E}X)]\big),$$

where the infimum is over positive definite $B \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ (the infimum may be further restricted to B for which the matrix in $\lambda_{\max}(\cdots)$ is a multiple of the identity). On the other hand, one may also easily deduce "user-friendly" bounds in the spirit of [67] whose statements make no reference to free probability.

Corollary 2.15 ("User-friendly" bound)**.** *Let $Y = \sum_{i=1}^n Z_i$, where $Z_1, \ldots, Z_n$ are independent (possibly non-self-adjoint) $d \times d$ random matrices with $\mathbf{E}[Z_i] = 0$. Then*

$$\mathbf{P}\big[\|Y\| \geq \|\mathbf{E}Y^*Y\|^{\frac{1}{2}} + \|\mathbf{E}YY^*\|^{\frac{1}{2}} + C\big\{v(Y)^{\frac{1}{2}}\sigma(Y)^{\frac{1}{2}}(\log d)^{\frac{3}{4}} + \sigma_*(Y)\,t^{\frac{1}{2}} + R(Y)^{\frac{1}{3}}\sigma(Y)^{\frac{2}{3}}t^{\frac{2}{3}} + R(Y)\,t\big\}\big] \leq 4de^{-t}$$

for a universal constant C and all $t \geq 0$, and

$$\mathbf{E}\|Y\| \leq \|\mathbf{E}Y^*Y\|^{\frac{1}{2}} + \|\mathbf{E}YY^*\|^{\frac{1}{2}} + C\big\{v(Y)^{\frac{1}{2}}\sigma(Y)^{\frac{1}{2}}(\log d)^{\frac{3}{4}} + R(Y)^{\frac{1}{3}}\sigma(Y)^{\frac{2}{3}}(\log d)^{\frac{2}{3}} + R(Y)\log d\big\}.$$

Here we define $\sigma(Y) := \max(\|\mathbf{E}Y^*Y\|^{\frac{1}{2}}, \|\mathbf{E}YY^*\|^{\frac{1}{2}})$ in the non-self-adjoint case, while $\sigma_*(Y), v(Y), R(Y)$ are defined as in section 2.1.4.

Proof. Combine Theorem 2.14, Remark 2.1, and [9, Lemmas 2.5 and 4.10]. □

There are many other possible combinations of the results of [9] and in section 2.2 above. For example, unbounded variants of Theorem 2.14 and Corollary 2.15 that involve the parameter $\bar{R}(X)$ instead of $R(X)$ are readily deduced from Theorem 2.7, and we obtain two-sided bounds for the moments and other spectral statistics of X in terms of those of X_{free} by combining the results of section 2.2.2 with those of [9, §2.2]. In the interest of space we do not spell out further combinations of this kind here; the appropriate results are easily applied directly in any given application.

3. Applications

In this section, we provide precise formulations of the applications that were introduced in section 1.3 above, and discuss how they arise from our main results. Some technical proofs are postponed until section 9.

3.1. **Random graphs and expanders.**

While random regular graphs, random Cayley graphs, and random n-lifts appear at first sight to be rather different models, our analysis of all these models will ultimately be based on a basic observation regarding random matrices defined by group representations. We first introduce some general facts, and then consider each of the above models in turn.

3.1.1. *Random matrices defined by group representations.* Let Γ be a finite group, let $\rho : \Gamma \to U(d)$ be a nontrivial unitary representation of dimension d, and let $g_1, \ldots, g_k$ be i.i.d. random variables drawn uniformly from Γ. Then

$$X = \sum_{i=1}^{k} (\rho(g_i) + \rho(g_i)^*) \tag{3.1}$$

defines a random matrix of the form (2.1). As $\rho(g_i)$ are unitary, it is easy to see that $\sigma(X) \asymp \sqrt{k}$ and $R(X) \asymp 1$. Therefore, our universality principles show that the spectrum of such matrices behaves as the associated Gaussian model when $k \gg (\log d)^{\beta}$ for a suitable $\beta > 0$. The random regular graph, Cayley graph, and n-lift models will all arise as variations on this theme.

To understand the behavior of such matrices, it then remains to understand the behavior of the Gaussian model associated to X. The following standard group-theoretic facts will suffice for this purpose.

Lemma 3.1. *Suppose that ρ is irreducible. Then there exists $s \in \{-1, 0, 1\}$ so that*

$$\mathbf{E}[\rho(g_i)] = 0, \qquad \mathrm{Cov}(\rho(g_i)) = \frac{1}{d}\mathbf{1}, \qquad \mathbf{E}[\rho(g_i)^2] = \frac{s}{d}\mathbf{1}.$$

Proof. The first two statements are standard facts about nontrivial irreducible representations [44, Proposition 4.3.1 and Corollary 4.3.9]. To prove the last statement, we first observe that $\mathrm{Tr}\,\mathbf{E}[\rho(g_i)^2] =: s \in \{-1, 0, 1\}$ by a theorem of Frobenius and Schur [44, Theorem 6.2.3]. On the other hand, for any $h \in \Gamma$, the random variables $g_i, h^{-1}g_i, g_i h^{-1}$ are uniformly distributed on Γ. Therefore,

$$\rho(h)\,\mathbf{E}[\rho(g_i)^2]\,\rho(h^{-1}) = \mathbf{E}[\rho(g_i h^{-1} g_i)]\,\rho(h^{-1}) = \mathbf{E}[\rho(g_i)^2]$$

for every $h \in \Gamma$, and the conclusion follows by [44, Proposition 4.3.4]. □

Remark 3.2. Let us emphasize that as $\rho(g_i)$ is not self-adjoint, the entry covariance matrix $\mathrm{Cov}(\rho(g_i))$ as defined in (2.2) does not fully determine the covariance of the real and imaginary parts of the entries of $\rho(g_i)$. In fact, by the Frobenius-Schur theorem used in the proof, it is the value s that determines whether the representation is real ($s = 1$), complex ($s = 0$), or quaternionic ($s = -1$).

3.1.2. *Random regular graphs.* In this subsection, let $\Pi_1, \ldots, \Pi_k$ be i.i.d. uniformly distributed random $d \times d$ permutation matrices. Then

$$X = \sum_{i=1}^{k} (\Pi_i + \Pi_i^*)$$

is the adjacency matrix of a (not necessarily simple) $2k$-regular graph with d vertices. This is the *permutation model* of random regular graphs. Before we proceed, let us recall a basic fact about the adjacency matrix of any regular graph.

Lemma 3.3 (Alon-Boppana)**.** *Let A be the adjacency matrix of an m-regular graph with d vertices. Then the largest and second largest eigenvalue of A satisfy*

$$\lambda_1(A) = m, \qquad \lambda_2(A) \geq \left(1 - C\,\frac{\log m}{\log d}\right) 2\sqrt{m-1}$$

for a universal constant C. Moreover, $1 \in \mathbb{R}^d$ (the vector all of whose entries are one) is an eigenvector of A with eigenvalue m.

Proof. The statement about the largest eigenvalue and eigenvector follow immediately from the Perron-Frobenius theorem and $A1 = m1$. The bound on the second eigenvalue is a classical result of Alon-Boppana (in this form, see [53]). □

In the seminal paper [38], Friedman shows that when k is fixed and $d \to \infty$, the second largest eigenvalue of the adjacency matrix X of the permutation model satisfies $\lambda_2(X) \leq (1 + o(1))2\sqrt{2k-1}$ with high probability. Thus, by Lemma 3.3, such graphs have the largest possible spectral gap to leading order, that is, they are "nearly Ramanujan". It is an old question whether this conclusion persists when both $k, d \to \infty$; see, e.g., [15, §1.4], and [74] for further questions of this kind. While some quantitative information can be extracted from proofs of Friedman's theorem (for example, a special case of [20, Theorem 1.4] shows that Friedman's result remains valid when $k \ll \frac{\log d}{(\log\log d)^2}$), all known proofs appear to break down for larger k. In the latter regime, it is known [39, 34, 31] that $\lambda_2(X) = O(\sqrt{k})$, but these results cannot recover the optimal constant.

We presently settle this question for $k \gg (\log d)^4$. This leaves only a narrow range of parameters $\frac{\log d}{(\log\log d)^2} \lesssim k \lesssim (\log d)^4$ open.

Theorem 3.4. *Denote by $X^\perp$ the restriction of X to $1^\perp$. Then for every $a > 0$, there is a constant $C > 0$ depending only on a so that*

$$\|X^\perp\| \le \left(1 + C\frac{(\log d)^{\frac{3}{4}}}{d^{\frac{1}{4}}} + C\frac{(\log d)^{\frac{2}{3}}}{k^{\frac{1}{6}}} + C\frac{\log d}{k^{\frac{1}{2}}}\right) 2\sqrt{2k}$$

with probability at least $1 - d^{-a}$. In particular, $\lambda_2(X) \le (1 + o(1))2\sqrt{2k-1}$ with probability $1 - o(1)$ whenever $d, k \to \infty$ with $(\log d)^4 = o(k)$.

Proof. The random matrix X is a special case of the model of the previous section, where we choose ρ to be the permutation representation of the symmetric group S_d. Moreover, $\rho^\perp = \rho|_{1^\perp}$ is an irreducible representation of S_d of dimension $d-1$. We can therefore compute using Lemma 3.1

$$\mathbf{E}[X^\perp] = 0, \qquad \|\mathbf{E}[(X^\perp)^2]\| = 2k\left(1 + \frac{s}{d-1}\right)$$

with $|s| \le 1$, as well as (using that $v(A + B) \le v(A) + v(B)$)

$$\sigma_*(X^\perp) \le v(X^\perp) \le 2\,v\left(\sum_{i=1}^k \rho^\perp(g_i)\right) = 2\sqrt{\frac{k}{d-1}}, \qquad R(X^\perp) \le 2.$$

The bound on $\|X^\perp\|$ follows directly by applying the tail bound of Corollary 2.15 with $t = (a+3)\log d$. As 1 is an eigenvector of X with eigenvalue $\lambda_1(X)$, we clearly have $\lambda_2(X) \le \|X^\perp\|$ and the proof is complete. □

Remark 3.5. Theorem 3.4 yields an upper bound on $\lambda_2(X)$ for the permutation model, which agrees with the Alon-Boppana lower bound that holds for *any* regular graph. Note, however, that the Alon-Boppana bound (Lemma 3.3) is only meaningful when $\log k \ll \log d$. A variant of Theorem 3.4 readily shows that for the permutation model, the lower bound $\lambda_2(X) \ge (1 + o(1))2\sqrt{2k-1}$ remains valid for *any* $k \gg (\log d)^4$, even when the Alon-Boppana bound fails (e.g., combine Corollary 2.6 with [9, Corollary 2.11]). However, when the Alon-Boppana bound fails it need not be the case that random regular graphs are optimal expanders.

3.1.3. *Random Cayley graphs.* In this subsection we let Γ be a finite group, and let $g_1, \ldots, g_k$ be i.i.d. variables drawn uniformly from Γ. We consider the Cayley graph defined by the generating set $\{g_1, \ldots, g_k, g_1^{-1}, \ldots, g_k^{-1}\}$, and denote its adjacency matrix by X. This model is a special case of (3.1) where $\rho : \Gamma \to U(\ell_2(\Gamma))$ is the right-regular representation of Γ, that is, $(\rho(g)f)_h = f_{hg}$.

A classical result of Alon and Roichman [2] states that if we choose $k \gg \log|\Gamma|$ generators, then the random Cayley graph is an expander, that is, $\lambda_2(X) = o(k)$ (see also [51] and the references therein for alternative proofs and extensions). The remarkable feature of this result is that it holds for *any* finite group Γ. Here we exhibit a new phenomenon: for many groups, choosing $k \gg (\log|\Gamma|)^4$ generators suffices to ensure that the random Cayley graph is nearly Ramanujan, that is, that $\lambda_2(X) \le (1 + o(1))2\sqrt{2k-1}$. This cannot happen for an arbitrary group, as the weaker estimates of [2] are essentially optimal for abelian groups. Rather, we show that this is the case for general nonabelian groups as soon as the dimensions of the nontrivial irreducible representations are not too small.

Theorem 3.6. *Denote by $d_0 \le d_1 \le \ldots \le d_m$ the dimensions of the isomorphism classes of irreducible representations of Γ, where $d_0 = 1$ corresponds to the trivial representation. Define the parameter*

$$\alpha(\Gamma) := \max_{1\le i\le m} \sqrt{\frac{\log(i+1)}{d_i}}.$$

Then for every $a > 0$, there is a constant $C > 0$ depending only on a so that

$$\|X^\perp\| \le \bigg(1 + C\alpha(\Gamma) + C\frac{(\log d_1)^{\frac{3}{4}}}{d_1^{\frac{1}{4}}} + C\frac{(\log|\Gamma|)^{\frac{2}{3}}}{k^{\frac{1}{6}}} + C\frac{\log|\Gamma|}{k^{\frac{1}{2}}}\bigg)2\sqrt{2k}$$

with probability at least $1 - d_1^{-a}$. In particular, $\lambda_2(X) \le (1 + o(1))2\sqrt{2k-1}$ with probability $1 - o(1)$ whenever $\alpha(\Gamma) = o(1)$ and $(\log|\Gamma|)^4 = o(k)$.

Proof. The Peter-Weyl theorem [44, Theorem 5.4.1] states that the right-regular representation decomposes as a direct sum $\rho = \bigoplus_{i=0}^m (\rho_i \otimes \mathbf{1}_{d_i})$ of non-isomorphic irreducible representations ρ_i of Γ, each of which appears with multiplicity d_i (here $\mathbf{1}_d$ is the identity matrix of dimension d). In other words, there is a choice of basis in which X is block-diagonal with blocks $X_i \otimes \mathbf{1}_{d_i}$, where each X_i is of the form (3.1) for a distinct irreducible representation of Γ. As the trivial representation accounts for the action on the eigenvector 1, we obtain $X^\perp = \bigoplus_{1\le i\le m} X_i \otimes \mathbf{1}_{d_i}$.

As in the proof of Theorem 3.4, we can compute using Lemma 3.1

$$\mathbf{E}[X_i] = 0, \quad \sigma(X_i)^2 \le 2k\bigg(1 + \frac{1}{d_i}\bigg), \quad \sigma_*(X_i)^2 \le v(X_i)^2 \le \frac{4k}{d_i}, \quad R(X_i) \le 2$$

for $1 \le i \le m$. Thus Corollary 2.15 yields

$$\mathbf{P}\bigg[\|X_i\| \ge \bigg(1 + C\frac{(\log d_i)^{\frac{3}{4}}}{d_i^{\frac{1}{4}}} + C\frac{t^{\frac{1}{2}}}{d_i^{\frac{1}{2}}} + C\frac{t^{\frac{2}{3}}}{k^{\frac{1}{6}}} + C\frac{t}{k^{\frac{1}{2}}}\bigg)2\sqrt{2k}\bigg] \le 4d_i e^{-t}$$

for all $1 \le i \le m$ and $t \ge 0$. Choosing $t = 2\log(i+1) + (a+3)\log d_i$ yields

$$\mathbf{P}\bigg[\|X_i\| \ge \bigg(1 + C\alpha(\Gamma) + C\frac{(\log d_1)^{\frac{3}{4}}}{d_1^{\frac{1}{4}}} + C\frac{(\log|\Gamma|)^{\frac{2}{3}}}{k^{\frac{1}{6}}} + C\frac{\log|\Gamma|}{k^{\frac{1}{2}}}\bigg)2\sqrt{2k}\bigg] \le \frac{d_1^{-a}}{(i+1)^2}$$

for all $1 \le i \le m$, where C depends only on a. (Here we used that $t \le (a+5)\log|\Gamma|$ as $d_i \le |\Gamma|$ and $m + 1 \le |\Gamma|$.) Using $\|X^\perp\| = \max_{1\le i\le m} \|X_i\|$, applying a union bound, and noting that $\sum_{i=1}^\infty (i+1)^{-2} < 1$ concludes the proof. □

The parameter $\alpha(\Gamma)$ in Theorem 3.6 controls the growth rate of the dimensions of the irreducible representations of Γ. The condition $\alpha(\Gamma) = o(1)$ holds as soon as the irreducible representations are sufficiently high-dimensional. This condition is satisfied in many examples. For example, the following was shown in [12].

Lemma 3.7. *For any sequence Γ_n of nonabelian finite simple groups such that $|\Gamma_n| \to \infty$, we have $\alpha(\Gamma_n) \to 0$.*

Proof. This can be verified directly by checking all the cases of the classicification of finite simple groups, as is done in the proof of [12, Lemma 9]. □

Theorem 3.6 therefore implies that for nonabelian finite simple groups, the Cayley graph defined by choosing $k \gg (\log|\Gamma|)^4$ random generators is nearly Ramanujan with high probability. When we are in addition in the domain of validity of the

Alon-Boppana theorem, that is, when $\log k \ll \log |\Gamma|$ (cf. Lemma 3.3), it follows that these random Cayley graphs are optimal expanders.

The formulation of Theorem 3.6 was inspired by [12], where a variant of the Gaussian model G associated to X was introduced on an ad-hoc basis to illustrate certain subtleties in the formulation of matrix concentration inequalities [9, §8.1]. The key point here, however, is that the universality principles of this paper make random matrices of this kind appear in a fundamental manner in the study of the expansion properties of random Cayley graphs.

Remark 3.8. To achieve a sharp bound, Theorem 3.6 requires at least that the smallest dimension d_1 of a nontrivial irreducible representation diverges; that is, Theorem 3.6 is concerned with quasirandom groups in the sense of [40]. It is clear that this condition is also necessary to obtain an optimal expander with high probability. For example, if $\Gamma = \mathrm{S}_d$ is the symmetric group, $d_1 = 1$ corresponds to the sign representation, and the 1×1 block X_1 in the proof of Theorem 3.6 equals twice the sum of k i.i.d. symmetric Bernoulli variables. Thus $\lambda_2(X) \geq \|X_1\|$ already exceeds, say, $3\sqrt{2k-1}$ with constant probability. A similar argument applies to any sequence of groups for which $d_1 \not\to \infty$. However, in this situation it is still possible to obtain $\lambda_2(X) \leq (1+o(1))2\sqrt{2k-1}$ with constant probability as long as the number of low-dimensional irreducible representations remains bounded. Conditions for this to hold follow along the same lines as in Theorem 3.6.

3.1.4. *Random lifts.* For an m-regular graph, the Alon-Boppana lower bound on the second eigenvalue arises from the fact that $2\sqrt{m-1}$ is the spectral radius of the infinite m-regular tree. This suggests that an analogue of Lemma 3.3 for a *non-regular* graph H should lower bound its second eigenvalue by the spectral radius $\varrho(\hat{H})$ of its universal covering tree $\hat{H}$. This is captured, at least qualitatively, by [42, Theorem 6.6]. A non-regular graph H may thus be viewed as an optimal expander if its second eigenvalue is bounded by $(1+o(1))\varrho(\hat{H})$ [42, §6].

Amit and Linial [3] and Friedman [37] proposed a model of random graphs that is designed to achieve such optimal expansion properties. Given any base graph $H = ([d], E_H)$ with d vertices, its *random n-lift* $H^{(n)} = ([d] \times [n], E_{H^{(n)}})$ is obtained by duplicating each vertex and edge of the base graph n times, and randomly scrambling the duplicate edges among the duplicate vertices. That is, for each $e \in E_H$ with $e = (i,j)$, $i \leq j$, we construct $e_k \in E_{H^{(n)}}$, $k = 1, \ldots, n$ with $e_k = ((i,k),(j,\sigma_e(k)))$, where σ_e is a random permutation that is chosen independently for each $e \in E_H$. The adjacency matrix $X^{(n)}$ of $H^{(n)}$ is

$$X^{(n)} = \sum_{e \in E_H} (A_e \otimes \Pi_e^{(n)} + A_e^* \otimes \Pi_e^{(n)*}),$$

where $\Pi_e^{(n)}$ are i.i.d. uniformly distributed $n \times n$ random permutation matrices and A_e are the $d \times d$ matrices $A_e = e_i e_j^*$ for $e = (i,j)$, $i \leq j$.

It is important to note that for every n, the restriction of $X^{(n)}$ to $\mathbb{C}^d \otimes \mathbb{C}1$ coincides with the adjacency matrix $X^{(1)}$ of H. Thus every eigenvalue of H is also an eigenvalue of $H^{(n)}$. The *new* eigenvalues that are introduced by the random lift are the eigenvalues of $X^{(n)\perp}$, the restriction of $X^{(n)}$ to $\mathbb{C}^d \otimes 1^\perp$. The long-standing conjecture that $\|X^{(n)\perp}\| \leq (1+o(1))\varrho(\hat{H})$ as $n \to \infty$ for fixed H was proved by Bordenave and Collins in [19]. This shows that random n-lifts are optimal expanders provided the base graph is an optimal expander.

As in the case of random regular graphs, however, it is far from clear whether this phenomenon persists if one considers random n-lifts of an unbounded sequence of base graphs H_n. The best bound to date [20, Theorem 1.4] is restricted to n-lifts of graphs H with $|E_H| \ll \frac{\log n}{(\log\log n)^2}$ edges. The following result addresses the complementary regime where the maximal degree of H grows at least polylogarithmically in the number of vertices nd of its n-lift $H^{(n)}$.

Theorem 3.9. *Let $H = ([d], E_H)$ be an (undirected, not necessarily simple) graph without self-loops. Denote by $\mathrm{D}(H)$ the maximal degree of a vertex of H and by $\mathrm{M}(H)$ the maximal multiplicity of an edge of H. Then for every $a > 0$, there is a constant $C > 0$ depending only on a so that the new eigenvalues of $H^{(n)}$ satisfy*

$$\|X^{(n)\perp}\| \le \left(1 + C\frac{\mathrm{M}(H)^{\frac{1}{4}}}{n^{\frac{1}{4}}}\frac{(\log nd)^{\frac{3}{4}}}{\mathrm{D}(H)^{\frac{1}{4}}} + C\frac{(\log nd)^{\frac{2}{3}}}{\mathrm{D}(H)^{\frac{1}{6}}} + C\frac{\log nd}{\mathrm{D}(H)^{\frac{1}{2}}}\right)\varrho(\hat{H})$$

with probability at least $1-(nd)^{-a}$. In particular, $\|X^{(n)\perp}\| \le (1+o(1))\varrho(\hat{H})$ with probability $1-o(1)$ whenever $(\log nd)^4 = o(\mathrm{D}(H))$ and $\mathrm{M}(H) = O(n)$.

A surprising aspect of Theorem 3.9 is that when H is a simple graph and $\mathrm{D}(H) \gg (\log d)^4$, the conclusion holds already for $n = 2$, that is, for random 2-lifts. This is stark contrast to the bounded degree case, where one must in general let $n \to \infty$ to achieve an $(1 + o(1))\varrho(\hat{H})$ upper bound. On the other hand, when $n \to \infty$, a quantitative Alon-Boppana type theorem of [20, Theorem 1.7] shows that there is a broad range of parameters where Theorem 3.9 yields the smallest possible new eigenvalues among all (not necessarily random) n-lifts.

Remark 3.10. The assumption that H has no self-loops was made for simplicity. The proof of Theorem 3.9 will allow for self-loops, but in this case we must let $n \to \infty$ to achieve a $(1 + o(1))\varrho(\hat{H})$ bound. We already discussed a special case of this setting: when H is the graph with 1 vertex and k self-loops, $H^{(n)}$ coincides with the permutation model of random regular graphs of section 3.1.2.

Remark 3.11. When H is a simple m-regular graph (for which $\varrho(\hat{H}) = 2\sqrt{m-1}$), a result along the lines of Theorem 3.9 can be obtained in a much simpler manner by comparing the norm of $X^\perp$ to that of a Wigner matrix [13, 11]. The primary interest of Theorem 3.9 is that it yields the correct upper bound for general H.

The proof of Theorem 3.9 is given in section 9.1. Let however briefly outline the argument. It is a basic fact (see, e.g., [19]) that $\varrho(\hat{H})$ may be computed as

$$\varrho(\hat{H}) = \left\| \sum_{e \in E_H} (A_e \otimes \lambda(g_e) + A_e^* \otimes \lambda(g_e)^*) \right\|, \tag{3.2}$$

where $(g_e)_{e\in E_H}$ are the free generators of the free group $\mathrm{F}_{|E_H|}$ and λ denotes the left-regular representation. On the other hand, Theorem 2.14 enables us to bound the norm of $X^{(n)\perp}$ by that of $X^{(n)\perp}_{\mathrm{free}}$. This almost yields the desired conclusion, except that in $X^{(n)\perp}_{\mathrm{free}}$ the free generators $\lambda(g_i)$ are replaced by certain deformed circular variables. While in general $\|X^{(n)\perp}_{\mathrm{free}}\| > \varrho(\hat{H})$, we will show these quantities coincide to leading order when $\mathrm{D}(H) \to \infty$, concluding the proof.

3.2. **Matrix concentration inequalities for smallest singular values.** The theory behind classical matrix concentration inequalities [67] is inherently limited to the extreme eigenvalues of random matrices. In contrast, our results control the entire spectrum. This makes it possible, for example, to obtain matrix concentration inequalities for the smallest singular value of non-self-adjoint random matrices. Such results are fundamentally outside the scope of classical matrix concentration inequalities for general models of the form (2.1) (unless one imposes special structure, see, e.g., [67, §5.2.1] for an example).

We presently state a general result of this kind. In the following, we define the smallest singular value of Y as $\mathrm{s}_{\min}(Y) := \inf \mathrm{sp}(|Y|)$.

Theorem 3.12. *Let $Y = Z_0 + \sum_{i=1}^n Z_i$, where Z_0 is a nonrandom $d \times m$ matrix and $Z_1, \ldots, Z_n$ are independent centered $d \times m$ random matrices, with $d \geq m$. Then*

$$\mathbf{P}\big[\mathrm{s}_{\min}(Y) \leq \mathrm{s}_{\min}(Y_{\mathrm{free}}) - C\big\{v(Y)^{\frac{1}{2}}\sigma(Y)^{\frac{1}{2}}(\log d)^{\frac{3}{4}} - \sigma_*(Y)t - R(Y)^{\frac{1}{3}}\sigma(Y)^{\frac{2}{3}}t^{\frac{2}{3}} - R(Y)t\big\}\big] \leq de^{-t}$$

for all $t \geq 0$, where C is universal constant. Here $\sigma(Y), \sigma_*(Y), v(Y), R(Y)$ are defined as in Corollary 2.15, and Y_{free} is the $d \times m$ matrix so that the real and imaginary parts of its entries is a semicircular family with the same mean and covariance as the real and imaginary parts of the entries of Y.

The proof of Theorem 3.12 will be given in section 9.2 below. A variant of Theorem 3.12 for unbounded random matrices can also be deduced along the same lines, by using Theorem 2.7 instead of Theorem 2.5 in the proof.

In the case that $\mathbf{E}[Y] = 0$, a simple "user-friendly" bound

$$\mathrm{s}_{\min}(Y_{\mathrm{free}}) \geq \mathrm{s}_{\min}(\mathbf{E}Y^*Y)^{\frac{1}{2}} - \|\mathbf{E}YY^*\|^{\frac{1}{2}} \tag{3.3}$$

was obtained in [9, Lemma 3.15]. This bound gives rise to very simple explicit estimates, but may be far from sharp for nonhomogeneous random matrices. In this case, the quantity $\mathrm{s}_{\min}(Y_{\mathrm{free}})$ can also be computed exactly using an explicit variational formula (in the spirit of [46]) that is obtained in [56].

Example 3.13 (Bipartite random graphs). Consider a bipartite random graph with vertex set $[d] \sqcup [m]$ $(d \geq m)$ in which each edge (i, j) with $i \in [d], j \in [m]$ is included independently with probability p_{ij}. This is a nonhomogeneous and bipartite analogue of the classical Erdős-Rényi model. The adjacency matrix A of this graph is the $d \times m$ matrix with independent entries $A_{ij} \sim \mathrm{Bern}(p_{ij})$.

A basic question of interest in this setting (cf. [35] and the references therein) is to bound the largest and smallest singular values of $A - \mathbf{E}A$.

Corollary 3.14. *Denote by $\rho := \min_j \sum_i p_{ij}(1 - p_{ij})$, $\gamma := \max_i \sum_j p_{ij}(1 - p_{ij})$, and $k := \max\big\{\max_i \sum_j p_{ij}(1 - p_{ij}), \max_j \sum_i p_{ij}(1 - p_{ij})\big\}$. Then for every $a > 0$, there is a constant $C > 0$ that depends only on a so that*

$$\|A - \mathbf{E}A\| \leq \sqrt{\rho} + \sqrt{\gamma} + Ck^{\frac{1}{3}}(\log d)^{\frac{2}{3}},$$
$$\mathrm{s}_{\min}(A - \mathbf{E}A) \geq \sqrt{\rho} - \sqrt{\gamma} - Ck^{\frac{1}{3}}(\log d)^{\frac{2}{3}}$$

with probability at least $1 - d^{-a}$, provided that $k \geq \log d$.

Proof. We can express $Y := A - \mathbf{E}A$ as $Y = \sum_{ij} Z_{ij}$ where $Z_{ij} = Y_{ij}e_ie_j^*$ are independent centered random matrices. Then we readily compute $\mathrm{s}_{\min}(\mathbf{E}Y^*Y) = \rho$,

$\|\mathbf{E}YY^*\| = \gamma$, $\sigma^2(X) \le k$, $\sigma_*(Y)^2 \le v(Y)^2 \le \max_{ij} p_{ij}(1-p_{ij})$, and $R(Y) \le 1$. The conclusion now follows directly from Corollary 2.15, Theorem 3.12, and (3.3). ☐

The simplest example of this result is the homogeneous case where $p_{ij} = p < 1$. In this case, the above bounds reduce to

$$1 - \sqrt{\frac{m}{d}} - \frac{C(\log d)^{\frac{2}{3}}}{(dp)^{\frac{1}{6}}} \le \frac{\mathrm{s}_{\min}(A - \mathbf{E}A)}{\sqrt{dp(1-p)}} \le \frac{\|A - \mathbf{E}A\|}{\sqrt{dp(1-p)}} \le 1 + \sqrt{\frac{m}{d}} + \frac{C(\log d)^{\frac{2}{3}}}{(dp)^{\frac{1}{6}}}.$$

This shows that the classical Bai-Yin law [7], which applies to dense graphs with constant $0 < p, \frac{m}{d} < 1$ as $d \to \infty$, remains valid for sparse graphs with average degree $dp \gg (\log d)^4$. In this homogeneous setting, the results of [35] establish the same conclusion in the slightly larger range $dp \gg \log d$. However, for nonhomogeneous graphs, the best known bounds due to [35] are already weaker than those of Corollary 3.14 to leading order, cf. [35, Remark 2.6].

On the other hand, we have formulated the simple bounds of Corollary 3.14 for sake of illustration only: the same proof yields bounds in which $\sqrt{\rho} + \sqrt{\gamma}$ and $\sqrt{\rho} - \sqrt{\gamma}$ are replaced by the optimal leading-order terms $\|Y_{\mathrm{free}}\|$ and $\mathrm{s}_{\min}(Y_{\mathrm{free}})$, respectively (where $Y = A - \mathbf{E}A$), which can be computed in terms of explicit variational principles [46, 56]. In other words, in contrast to previous results, we obtain sharp Bai-Yin laws for sparse nonohomogeneous random matrices.

Remark 3.15. More generally, Theorem 3.12 may be viewed as a nonasymptotic, nonhomogeneous Bai-Yin law that is sharp to leading order. It should be emphasized, however, that it can only locate the smallest singular value of Y near that of its noncommutative model Y_{free}. In particular, Theorem 3.12 sheds no light on the invertibility of Y when $\mathrm{s}_{\min}(Y_{\mathrm{free}}) = 0$, as is the case, e.g., for square matrices with i.i.d. entries. The latter question is of a fundamentally different nature, which is presently understood for nonhomogeneous models only under restrictive assumptions [62, 30] (see, however, [66] for significant recent progress in this direction).

3.3. **Sample covariance matrices.** Let $Y_1, \ldots, Y_n$ be independent, centered random vectors in $\mathbb{R}^d$. The $d \times d$ random matrix

$$S = \sum_{i=1}^{n} Y_i Y_i^* \tag{3.4}$$

is called the (nonhomogeneous) *sample covariance matrix* associated to the data $Y_1, \ldots, Y_n$. Equivalently, we may express $S = YY^*$, where

$$Y = \sum_{i=1}^{n} Y_i e_i^* \tag{3.5}$$

is the $d \times n$ random matrix with independent columns $Y_1, \ldots, Y_n$. In the classical setting where $Y_1, \ldots, Y_n$ are identically distributed, $\frac{1}{n}\mathbf{E}S$ is the covariance matrix of Y_i, and $\frac{1}{n}S$ may be viewed as a statistical estimator of this covariance matrix. A central problem is then to bound the deviation $\frac{1}{n}\|S - \mathbf{E}S\|$ of the estimated covariance matrix from the actual covariance matrix. Here we allow for a more general nonhomogeneous situation where the data $Y_1, \ldots, Y_n$ need not be identically distributed, which is of independent interest (see, e.g., [22]).

From the viewpoint of this paper, sample covariance matrices may approached in two different ways: we may either view S itself as a model of the form (2.1), or we may view Y as a model of the form (2.1). These two interpretations give rise to distinct universality principles. As we will see below, neither approach subsumes the other: they control the behavior of S in complementary regimes.

For simplicity, we focus in this section on "user-friendly" explicit bounds on the expected deviation $\mathbf{E}\|S-\mathbf{E}S\|$; sharp bounds in terms of $S_{\rm free}$ and $Y_{\rm free}$, as well as high-probability bounds, may be obtained analogously.

3.3.1. *Gaussian sample covariance matrices.* In this section we consider Gaussian sample covariance matrices, that is, (3.4) where $Y_1,\ldots,Y_n$ are independent Gaussian random vectors $Y_i \sim N(0,\Sigma_i)$. In this case, Y is a Gaussian random matrix, to which the Gaussian theory of [9] can be applied.

Theorem 3.16 (Gaussian bound). *Let $Y_i \sim N(0,\Sigma_i)$. Then we have*

$$\mathbf{E}\|S-\mathbf{E}S\| \le 2\Bigg\|\sum_{i=1}^n \mathrm{Tr}[\Sigma_i]\Sigma_i\Bigg\|^{\frac{1}{2}} + \max_{i\le n}\mathrm{Tr}\,\Sigma_i + C\Bigg(\Bigg\|\sum_{i=1}^n \Sigma_i\Bigg\| + \max_{i\le n}\mathrm{Tr}\,\Sigma_i\Bigg)^{\frac{3}{4}} \max_{i\le n}\|\Sigma_i\|^{\frac{1}{4}}\log^{\frac{3}{2}}(d+n).$$

Proof. By [9, Lemma 3.8], we have $\sigma(Y)^2 = \|\sum_i \Sigma_i\| \vee \max_i \mathrm{Tr}\,\Sigma_i$ and $v(Y)^2 = \max_i\|\Sigma_i\|$. As $v(Y)\le\sigma(Y)$, [9, Theorem 3.11 and Proposition 3.12] yield

$$\mathbf{E}\|S-\mathbf{E}S\| \le 2\|\mathbf{E}[Y\,\mathbf{E}[Y^*Y]\,Y^*\|^{\frac{1}{2}} + \|\mathbf{E}Y^*Y\| + C\sigma(Y)^{\frac{3}{2}}v(Y)^{\frac{1}{2}}\log^{\frac{3}{2}}(d+n)$$

for a universal constant C. The leading terms are readily computed. ☐

On the other hand, even when Y is Gaussian, we may view S as a non-Gaussian random matrix of the form (2.1) with $Z_i = Y_iY_i^*$, to which the universality principles of this paper may be applied. For example, applying Theorem 2.8 yields the following bound, whose proof is given in section 9.3.1.

Theorem 3.17 (S-universality bound). *Let $Y_i \sim N(0,\Sigma_i)$. Then for any $\varepsilon\in(0,1]$*

$$\mathbf{E}\|S-\mathbf{E}S\| \le (1+\varepsilon)\,2\Bigg\|\sum_{i=1}^n \mathrm{Tr}[\Sigma_i]\Sigma_i\Bigg\|^{\frac{1}{2}} + \frac{C}{\varepsilon^3}\Bigg(\Bigg\|\sum_{i=1}^n \Sigma_i^2\Bigg\|^{\frac{1}{2}} + \max_{i\le n}\mathrm{Tr}\,\Sigma_i\Bigg)\log^3(d+n).$$

The fundamental distinction between these bounds is that Theorem 3.17 models S by the noncommutative model $S_{\rm free}$, while Theorem 3.16 models $S=YY^*$ by the noncommutative model $Y_{\rm free}Y_{\rm free}^*$. Somewhat surprisingly, these distinct interpretations have complementary (partially overlapping) domains of validity, which is already illustrated by the simplest possible example.

Example 3.18. Suppose that $Y_1,\ldots,Y_n$ are i.i.d. standard Gaussian vectors in $\mathbb{R}^d$, that is, $\Sigma_i = \mathbf{1}$ for all i. In this setting, the classical Bai-Yin law [7] implies that $\mathbf{E}\|S-\mathbf{E}S\| = (1+o(1))(2\sqrt{nd}+d)$ when $n,d\to\infty$ with $\frac{n}{d}$ fixed.

Let us now verify what Theorems 3.16 and 3.17 yield for this model.

- First, note that the Gaussian bound of Theorem 3.16 yields

$$\mathbf{E}\|S-\mathbf{E}S\| \le 2\sqrt{nd} + d + C(n+d)^{\frac{3}{4}}\log^{\frac{3}{2}}(d+n).$$

  Here the leading terms agree with the Bai-Yin law, but the error term is of smaller order if and only if $n\to\infty$ and $d \gg n^{\frac{1}{2}}(\log n)^3$. This includes the $n\propto d$ setting of the classical Bai-Yin law, but excludes cases where n is much larger than d.

- On the other hand, the universality bound of Theorem 3.17 yields
$$\mathbf{E}\|S-\mathbf{E}S\| \le (1+\varepsilon)\,2\sqrt{nd} + C\varepsilon^{-3}(\sqrt{n}+d)\log^3(d+n).$$
In this bound, the leading term agrees with the Bai-Yin law only when $n \gg d$, and the error term is of smaller order if and only if $(\log n)^6 \ll d \ll \frac{n}{(\log n)^6}$. This regime excludes the setting of the classical Bai-Yin law, but covers precisely the situation that the Gaussian bound fails to capture.

Combining the above bounds yields $\mathbf{E}\|S-\mathbf{E}S\| \le (1+o(1))(2\sqrt{nd}+d)$ whenever $n\to\infty$ and $d \gg (\log n)^6$. This very general conclusion hides the fact that two complementary approaches were used to capture the large d and large n regimes.

The homogeneous setting of Example 3.18 is special in that $\Sigma_i = \mathbf{1}$ implies $\|S-\mathbf{E}S\| = \|YY^*-n\mathbf{1}\| = \max\{\|Y\|^2-n, n-\mathrm{s}_{\min}(Y^*)^2\}$, so that this case can also be approached using the methods of section 3.2. Such a reduction fails, however, for nonhomogeneous sample covariance matrices. In the general setting, Theorems 3.16 and 3.17 control the behavior of Gaussian sample covariance matrices in complementary regimes that together span a wide range of parameters.

Remark 3.19. Let us note for completeness that Gaussian sample covariance matrices always satisfy $\mathbf{E}\|S-\mathbf{E}S\| \gtrsim \|\sum_i \mathrm{Tr}[\Sigma_i]\Sigma_i\|^{\frac{1}{2}} + \max_i \mathrm{Tr}\,\Sigma_i$, cf. section 9.3.2. Thus the leading terms in Theorems 3.16 and 3.17 are also lower bounds on $\mathbf{E}\|S-\mathbf{E}S\|$ up to a universal constant. On the other hand, the proofs of these results can even capture the sharp leading term predicted by free probability.

3.3.2. *Non-Gaussian models.* We now consider more general models where the data $Y_1,\ldots,Y_n$ may be non-Gaussian. This makes little difference in the setting of Theorem 3.17: here we already interpreted S itself as a non-Gaussian matrix, and applied the universality principle to compare it with its Gaussian model (the assumption that Y is Gaussian was not used in a fundamental way in the proof).

On the other hand, in order to extend Theorem 3.16 to the non-Gaussian setting, we must compare $S = YY^*$ with HH^*, where H is the Gaussian model associated to Y. Such a comparison can be deduced from our universality principles by means of a linearization argument as in [9, §3.3]. Note that the setting of the following result is far more general than that of the model (3.4).

Theorem 3.20 (Y-universality)**.** *Let $Y = Z_0 + \sum_{i=1}^n Z_i$ be a $d\times m$ random matrix defined as in Theorem 3.12, and let H be the $d\times m$ random matrix so that the real and imaginary parts of its entries are jointly Gaussian with the same mean and covariance as the real and imaginary parts of the entries of Y. Then*
$$\big|\mathbf{E}\|YY^*-\mathbf{E}YY^*\| - \mathbf{E}\|HH^*-\mathbf{E}HH^*\|\big| \lesssim \delta\,\mathbf{E}\|H\| + \delta^2$$
with
$$\delta = \sigma_*(Y)\log^{\frac{1}{2}}(d+m) + R(Y)^{\frac{1}{3}}\sigma(Y)^{\frac{2}{3}}\log^{\frac{2}{3}}(d+m) + R(Y)\log(d+m).$$

The proof of Theorem 3.20 is in section 9.3.3. We state the result for bounded random matrices for simplicity; similar results for unbounded matrices are obtained by using Theorem 2.7 rather than Theorem 2.5 in the proof.

As Y is already expressed as a sum of independent random matrices in (3.5), it is tempting to attempt to apply Theorem 3.20 with $Z_i = Y_i e_i^*$. Unfortunately, as is illustrated in the following example, such a straightforward application of the universality principle fails to yield meaningful results. The reason is simple: (3.5) captures only the independence of the data $Y_1, \ldots, Y_n$, but independent data alone does not suffice to ensure universality of sample covariance matrices.

Example 3.21. Let us revisit the Bai-Yin setting of Example 3.18 in the non-Gaussian case, that is, we now assume only that $Y_1, \ldots, Y_n$ are i.i.d. centered random vectors with unit covariance matrix $\Sigma_i = \mathbf{1}$ for all i.

In this setting, the Gaussian model H associated to Y in Theorem 3.20 is precisely the $d \times n$ random matrix with i.i.d. standard Gaussian entries. In particular, the classical Bai-Yin law [7] implies that $\mathbf{E}\|HH^* - \mathbf{E}HH^*\| = (1+o(1))(2\sqrt{\gamma}+1)d$ when $n, d \to \infty$ with $\frac{n}{d} = \gamma$ fixed. On the other hand, if we write $Y = \sum_{i=1}^n Z_i$ with $Z_i = Y_i e_i^*$, we clearly have $R(Y)^2 \geq \max_i \mathbf{E}\|Z_i\|^2 = d$, so that

$$\delta\, \mathbf{E}\|H\| + \delta^2 \gtrsim \frac{(\log d)^2}{2\sqrt{\gamma}+1}\, \mathbf{E}\|HH^* - \mathbf{E}HH^*\|.$$

Thus Theorem 3.20 cannot yield universality of the Bai-Yin law in this manner, as its error term is always of larger order than the Gaussian quantity of interest.

The problem that arises here is not an inefficiency of our universality principles, however, but is a genuine phenomenon: at the present level of generality, universality of the Bai-Yin law is simply false. For example, let $Y_i = \sqrt{d}\,\varepsilon_i e_{I_i}$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random signs and $I_1, \ldots, I_n$ are i.i.d. uniformly distributed variables on $[d]$. Then $Y_1, \ldots, Y_n$ are i.i.d. with zero mean and unit covariance. However, as YY^* is diagonal with multinomially distributed diagonal entries, we have [60]

$$\mathbf{E}\|YY^* - \mathbf{E}YY^*\| = (1+o(1))\frac{d \log d}{\log(\gamma^{-1}\log d)} \gg \mathbf{E}\|HH^* - \mathbf{E}HH^*\|$$

when $n, d \to \infty$ with $\frac{n}{d} = \gamma$ fixed. Thus universality of the Bai-Yin law fails.

Example 3.21 illustrates that even in the classical setting of Bai-Yin law, some additional assumption on the distribution of the vectors Y_i is needed to achieve universality. The additional assumption that would enable us to apply Theorem 3.20 is that each Y_i is itself a sum of independent random vectors of sufficiently small norm. This situation arises naturally in random matrix theory: we presently provide one example of such a model, where the above results yield a considerable improvement on the best known nonasymptotic bounds.

Example 3.22 (Product of random and deterministic matrices). Let A be a $N \times n$ random matrix with independent real entries that have zero mean and unit variance, and let B be a $d \times N$ nonrandom matrix. We are interested in the sample covariance matrix $S = YY^*$ where $Y = BA$. The difficulty of analyzing such models is that even though A has independent entries, the matrix Y generally has highly dependent entries which renders many standard tools of nonasymptotic random matrix theory inapplicable. Here we obtain the following.

Theorem 3.23. *Let $S = YY^*$ with $Y = BA$, where A is a $N \times n$ random matrix with $\mathbf{E}[A_{ij}] = 0$, $\mathrm{Var}(A_{ij}) = 1$, and $\|A_{ij}\|_\infty \leq \alpha$, and let B be a $d \times N$ nonrandom matrix. Assume that $\alpha \leq \sqrt{n}$ and $\|B\|_{\mathrm{HS}} \geq \alpha\|B\|$. Then we have*

$$\mathbf{E}\|S - \mathbf{E}S\| \leq$$

$$\left(1 + C\left\{\left(\frac{\alpha}{\sqrt{n}}\right)^{\frac{1}{15}} + \left(\frac{\alpha\|B\|}{\|B\|_{\mathrm{HS}}}\right)^{\frac{1}{4}}\right\}\log^3(d+n)\right)\left(2\|B\|_{\mathrm{HS}}\|B\|\sqrt{n} + \|B\|_{\mathrm{HS}}^2\right),$$

where C is a universal constant, $\mathbf{E}S = nBB^*$, and $\|M\|_{\mathrm{HS}}^2 := \mathrm{Tr}\,|M|^2$.

The proof is given in section 9.3.4. While we have formulated a single bound, it should be emphasized that the proof is once again a combination of two distinct universality principles. (We have made no effort to optimize the exponents in the lower-order terms, which are not expected to be optimal.)

It is instructive to compare Theorem 3.23 with previous nonasymptotic results in this setting, which establish bounds analogous to Theorem 3.23 up to a multiplicative factor that depends on the moments of A_{ij}: see [75] and the references therein for subgaussian or subexponential entries, and [72] for a slightly weaker result for entries with bounded fourth moment. The advantage of Theorem 3.23 is twofold. First, Theorem 3.23 reproduces the correct leading-order behavior in the Bai-Yin law (i.e., the case $N = d, B = \mathbf{1}$), while previous results lose at least a multiplicative factor. Second, Theorem 3.23 is applicable to sparse random matrices, while previous bounds are fundamentally inefficient in the sparse setting.

To illustrate this point, suppose that A_{ij} are symmetric Bernoulli variables with $\mathbf{P}[A_{ij} = 0] = 1 - p$ and $\mathbf{P}[A_{ij} = p^{-\frac{1}{2}}] = \frac{p}{2}$. Then Theorem 3.23 yields

$$\mathbf{E}\|S - \mathbf{E}S\| \le (1 + o(1))\big(2\|B\|_{\mathrm{HS}}\|B\|\sqrt{n} + \|B\|_{\mathrm{HS}}^2\big) \quad \text{for} \quad p \gg \frac{\log^{\beta}(d+n)}{n \wedge r}$$

for a suitable β, where $r = \|B\|_{\mathrm{HS}}^2\|B\|^{-2}$ is the effective rank of B. On the other hand, as $\mathbf{E}|A_{ij}|^4 = \frac{1}{p}$ diverges as soon as $p \to 0$, the results of [72, 75] fail to achieve even the correct order of magnitude of the norm in the sparse setting.

Remark 3.24. We have formulated Theorem 3.23 for the case that the entries A_{ij} are uniformly bounded, while prior results [72, 75] also consider unbounded entries. However, our restriction to bounded entries was made for simplicity of exposition only, and is not a fundamental restriction of the proof of Theorem 3.23. A related inequality for the unbounded case may be found in Remark 9.16.

While Example 3.22 provides a natural model where each Y_i is itself a sum of independent random vectors, such an assumption can be restrictive for more general models of sample covariance matrices. On the other hand, in the special (homogeneous) setting of the Bai-Yin law, it was shown in [28] that a much weaker assumption suffices to achieve universal behavior: in this case one need only assume that each Y_i satsifies certain concentration of measure properties, which rules out the counterexample of Example 3.21. In forthcoming work [55], the universality principles of this paper are further refined to capture such concentration assumptions for general (nonhomogeneous) sample covariance matrices.

More generally, the above considerations highlight the broader question whether the universality principles of this paper extend to random matrices that admit more general dependence structures than can be captured by the model (2.1); such principles could enable the analysis of natural models that are outside the scope of this paper. Progress in this direction may be found in [55, 71].

3.4. **Strong asymptotic freeness.** The celebrated asymptotic freeness theorem of Voiculescu [73] states that if $X_1^N, \ldots, X_m^N$ are independent $N \times N$ Wigner matrices and $s_1, \ldots, s_m$ is a free semicircular family (i.e., a semicircular family as in Definition 2.2 with zero mean and unit covariance), then

$$\lim_{N\to\infty} \operatorname{tr} p(X_1^N, \ldots, X_m^N) = \tau(p(s_1, \ldots, s_m)) \quad \text{a.s.}$$

for every noncommutative polynomial p. This makes it possible to compute the limiting spectral distributions of polynomials Wigner matrices using tools of free probabilty; see, e.g., [5, Chapter 5]. That the convergence holds also in norm

$$\lim_{N\to\infty} \|p(X_1^N,\ldots,X_m^N)\| = \|p(s_1,\ldots,s_m)\| \quad \text{a.s.}$$

is a deep result of Haagerup and Thorbjørnsen [41], who proved it for GUE matrices. The latter strong asymptotic freeness property is of fundamental importance both to random matrices and in the theory of operator algebras.

The methods of [41] are rather delicate, and their extension even to random matrices with i.i.d. entries with bounded fourth moment requires considerable effort [4]. It was therefore long unclear whether the strong asymptotic freeness pheonomenon could be expected to hold in the absence of strong symmetry assumptions. That this is indeed the case is a notable application of the sharp matrix concentration theory of [9], which made it possible to establish strong asymptotic freeness of an extremely general class of Gaussian random matrix models.

Here we extend the latter results to an even more general family of non-Gaussian random matrices. To this end, we must show that our universality principles for random matrices of the form (2.1) imply universality for polynomials of such matrices. In the following (asymptotic) result, whose proof is given in section 9.4, this is accomplished by a direct application of known linearization arguments [41, 33] that we use as a black box. However, while we do not develop this direction systematically in this paper, our methods can also be used to obtain nonasymptotic bounds for polynomials of random matrices: for example, Theorem 3.20 may be viewed as a result of this kind for a certain quadratic polynomial.

Theorem 3.25 (Strong asymptotic freeness)**.** *Let $s_1,\ldots,s_m$ be a free semicircular family. For each $N\ge 1$, let Let $H_1^N,\ldots,H_m^N$ be independent self-adjoint random matrices of dimension $d_N\ge N$ defined by*

$$H_k^N = Z_{k0}^N + \sum_{i=1}^{M_N} Z_{ki}^N,$$

where Z_{k0}^N is a deterministic self-adjoint matrix and $Z_{k1}^N,\ldots,Z_{kM_N}^N$ are independent self-adjoint random matrices with zero mean. Suppose that

$$\lim_{N\to\infty}\|\mathbf{E}[H_k^N]\| = \lim_{N\to\infty}\|\mathbf{E}[(H_k^N)^2]-\mathbf{1}\| = \lim_{N\to\infty}\bar{R}(H_k^N) = 0$$

and that

$$\lim_{N\to\infty}(\log d_N)^{\frac{3}{2}} v(H_k^N) = 0, \qquad \lim_{N\to\infty}(\log d_N)^2 \max_{1\le i\le M_N}\|Z_{ki}^N\| = 0 \quad \text{a.s.}$$

for every $1\le k\le m$. Then

$$\lim_{N\to\infty}\operatorname{tr} p(H_1^N,\ldots,H_m^N) = \tau(p(s_1,\ldots,s_m)) \quad \text{a.s.},$$
$$\lim_{N\to\infty}\|p(H_1^N,\ldots,H_m^N)\| = \|p(s_1,\ldots,s_m)\| \quad \text{a.s.}$$

for every noncommutative polynomial p. Moreover, the analogous result holds if a.s. convergence is replaced by convergence in probability.

Theorem 3.25 applies to a large family of random matrices with non-Gaussian, nonhomogeneous, and dependent entries. In order to illustrate some characteristic features of this result, let us develop one example in more detail.

Example 3.26 (Sparse Wigner matrices). We consider matrices with a deterministic sparsity pattern, where all nonzero entries of the matrix are i.i.d. We emphasize that such random matrices may be highly nonhomogeneous.

Definition 3.27. Let $(\eta_{ij})_{1\le i\le j<\infty}$ be i.i.d. real-valued random variables with zero mean and unit variance, and let $\mathrm{G} = ([d], E)$ be a k-regular graph with d vertices. Then the (G, η)-*sparse Wigner matrix* is the $d \times d$ self-adjoint random matrix X with entries $X_{ij} = k^{-\frac{1}{2}}\eta_{ij}1_{\{i,j\}\in E}$ for $1 \le i \le j \le d$.

The proof of of the following result is given in section 9.4.3.

Corollary 3.28. *Let $(\eta_{rij})_{1\le r\le m, 1\le i\le j<\infty}$ be i.i.d. centered random variables with unit variance and $\mathbf{E}[|\eta_{kij}|^p] < \infty$ for some $p > 2$, and let G_N be a k_N-regular graph with $d_N \ge N$ vertices. Let H_r^N be the (G_N, η_r)-sparse Wigner matrix.*

a. If $k_N \gg d_N^{\frac{2}{p-2}}(\log d_N)^{\frac{4p}{p-2}}$, then

$$\lim_{N\to\infty} \mathrm{tr}\, p(H_1^N, \ldots, H_m^N) = \tau(p(s_1, \ldots, s_m)) \text{ in probability,}$$

$$\lim_{N\to\infty} \|p(H_1^N, \ldots, H_m^N)\| = \|p(s_1, \ldots, s_m)\| \text{ in probability}$$

for every noncommutative polynomial p. If the graphs $\mathrm{G}_N = ([d_N], E_N)$ are increasing (i.e., $E_N \subseteq E_{N+1}$ for all N), the convergence also holds a.s.

b. If $k_N \ll d_N^{\frac{2}{p-2}}(\log d_N)^{-\frac{2p}{p-2}}$, then the conclusion of part a. must fail for some entry distribution satisfying the assumptions.

A fundamental phenomenon that is captured by this result is that strong asymptotic freeness requires a tradeoff between sparsity and integrability of the entries. For dense Wigner matrices $k_N = d_N$, we obtain strong asymptotic freeness as soon as the entries have $4 + \varepsilon$ moments for some $\varepsilon > 0$, as was previously shown in [4]. On the other hand, as we bound more moments, increasingly sparse random matrices can still achieve strong asymptotic freeness. This tradeoff is captured nearly optimally by Corollary 3.28, up to logarithmic factors.

In the opposite extreme, when the entries η_{rij} are uniformly bounded, it follows directly from Theorem 3.25 that strong asymptotic freeness holds (in the a.s. sense) as soon as $k_N \gg (\log d_N)^4$. By taking η_{rij} to be symmetric Bernoulli variables, this shows that one can construct $d\times d$ random matrices that achieve strong asymptotic freeness using only $O(d \log^5 d)$ bits of randomness.

Remark 3.29. The sparse Wigner model of Example 3.26 is only one special case of the very general setting captured by Theorem 3.25, which also includes several of the examples that were discussed in the previous sections: e.g., random matrices defined by group representations as in section 3.1, or centered adjacency matrices of sparse Erdős-Rényi graphs. Such examples further extend the scope of the strong asymptotic freeness phenomenon beyond what was previously known.

3.5. **Phase transitions in spiked models.** A widely studied phenomenon in random matrix theory, which dates back to the work of Baik, Ben Arous and Péché [8], is that low-rank perturbations of random matrices (known as "spiked" models) give rise to phase transitions: small perturbations do not affect the limiting eigenvalue statistics, while large perturbations give rise to the appearance of outlier eigenvalues. Several closely related forms of this phenomenon have been investigated by many authors; see, e.g., the survey [24]. For sake of illustration, we focus here on the following prototypical phenomenon of this kind.[2]

Theorem 3.30 (BBP transition for spiked GOE [16]). *Let G_d be a $d\times d$ self-adjoint random matrix whose entries $(G_{dij})_{i\geq j}$ are independent real Gaussian variables with mean 0 and variance $\frac{1+1_{i=j}}{d}$. Let A_d be a nonrandom $d \times d$ positive semidefinite matrix of rank r whose eigenvalues $\theta_1 \geq \cdots \geq \theta_r > 0$ are independent of d.*

a. We have for $1 \leq i \leq r$

$$\lambda_i(A_d + G_d) \xrightarrow[a.s.]{d\to\infty} \begin{cases} \theta_i + \frac{1}{\theta_i} & \text{for } \theta_i > 1, \\ 2 & \text{for } \theta_i \leq 1, \end{cases} \qquad \lambda_{r+1}(A_d + G_d) \xrightarrow[a.s.]{d\to\infty} 2,$$

where $\lambda_1(M) \geq \cdots \geq \lambda_d(M)$ are the eigenvalues of M.

b. For every $1 \leq i, j \leq r$ such that $\theta_i > 1$ and $\theta_j \neq \theta_i$, we have

$$\|P_i(A_d)v_i(A_d + G_d)\|^2 \xrightarrow[a.s.]{d\to\infty} 1 - \frac{1}{\theta_i^2}, \qquad \|P_j(A_d)v_i(A_d + G_d)\|^2 \xrightarrow[a.s.]{d\to\infty} 0,$$

where $P_i(M)$ is the projection on the eigenspace of M associated to the eigenvalue $\lambda_i(M)$, and $v_i(M)$ is any unit norm eigenvector of M with eigenvalue $\lambda_i(M)$.

We aim to understand whether the phenomena described in Theorem 3.30 are universal: do the conclusions remain valid if the GOE matrix G_n is replaced by another random matrix H_d whose entries have the same mean and covariance? Previous results have extended Theorem 3.30 to the setting where H_d has i.i.d. entries above the diagonal under distributional assumptions that require at least some bounded moments of higher order, cf. [24] and the references therein. (Analogues of Theorem 3.30 are also known to hold for non-Gaussian homogeneous models where H_d is invariant under a symmetry group; such models are rather different in spirit from the kind of universality phenomena considered here.)

When applied to this setting, our universality principles can capture many new situations, including sparse and dependent models.

Theorem 3.31. *Let G_d, A_d be as in Theorem 3.30, and let H_d be any $d \times d$ self-adjoint real random matrix of the form (2.1) whose entries have the same mean and covariance as those of G_d. Suppose that $(\log d)^2 R(H_d) \to 0$ as $d \to \infty$. Then all the conclusions of Theorem 3.30 remain valid if G_d is replaced by H_d.*

The proof of this result is given in section 9.5. Let us however briefly outline the main ingredients of the proof. On the one hand, Theorem 2.5 shows that the eigenvalues of A_d+H_d concentrate at the locations predicted by Theorem 3.30. On the other hand, for $\theta_i > 1$, let φ_i be a mollification of the indicator function of a small interval around $\theta_i+\frac{1}{\theta_i}$. Then $\varphi_i(A_d+H_d)$ coincides with high probability with the projection onto the linear span of the eigenvectors of A_d+H_d whose eigenvalues concentrate at $\theta_i + \frac{1}{\theta_i}$. We can therefore apply the second part of Theorem 2.9 to establish universality of these eigenprojections.

To illustrate Theorem 3.31, we briefly discuss one simple example.

Example 3.32 (Planted clique in the permutation model). Let X_d be the adjacency matrix of a random $2k_d$-regular graph with d vertices in the permutation model

[2]The assumption that A_d is positive semidefinite is made here exclusively to simplify the notation; any negative eigenvalues of A_d exhibit a completely analogous transition at $\theta_i = -1$.

defined in section 3.1.2, where $(\log d)^4 \ll k_d \ll d^2$. Choose a subset $E_d \subset [d]$ of vertices so that $|E_d| = (1+o(1))\theta\sqrt{2k_d}$. Then $1_{E_d}1_{E_d}^* + X_d$ is the adjacency matrix of the random graph in which we planted a clique with vertices E_d.

By Lemma 3.1, the random matrix $(2k_d)^{-\frac{1}{2}}X_d^\perp$, where $X_d^\perp$ is the restriction of X_d to $1^\perp$, has the same mean and covariance as a GOE matrix of dimension $d-1$. Furthermore, the assumptions on k_d and E_d imply that there exist unit vectors $v_d \in 1^\perp$ so that $\|(2k_d)^{-\frac{1}{2}}1_{E_d}1_{E_d}^* - \theta v_d v_d^*\| \to 0$. Thus applying Theorem 3.31 with $H_{d-1} = (2k_d)^{-\frac{1}{2}}X_d^\perp$ and $A_{d-1} = \theta v_d v_d^*$ shows that the adjacency matrix of the planted model has an outlier eigenvalue (beside its Perron-Frobenius eigenvalue) if and only if $\theta > 1$. In other words, the detectability of a planted clique by an outlier in the spectrum exhibits a phase transition at $|E_d| = \sqrt{2k_d}$.

Let us emphasize that the random matrices that arise in this example are both dependent and may be highly sparse. (For the classical study of spectral detection of planted cliques in dense Erdős-Rényi graphs, see [1].)

Remark 3.33. Even if we consider H_d with i.i.d. entries, sparse matrices are not captured by previous extensions of Theorem 3.30 as their entries have unbounded moments of order $p > 2$ (cf. Example 3.22). Some results for sparse matrices were obtained very recently in [64], but rely on a special choice of A_d.

In this section we have used the classical Gaussian result of Theorem 3.30 as input for the universality theory of this paper. However, much more general results can be obtained in the Gaussian setting by applying the sharp matrix concentration theory of [9]. This approach has two key advantages: it is nonasymptotic, and it yields analogous phenomena in nonhomogeneous situations. The latter are of particular interest in many applications, but are much less well understood than the homogeneous setting [18, 45, 64, 6]. The development of the nonhomogeneous theory in the Gaussian case will be treated in the forthcoming work [10]; its extension to non-Gaussian situations proceeds precisely as we do here.

4. The cumulant method

The aim of this section is to introduce the basic device that we will use to prove universality throughout this paper. The general setting that will be considered in this section is the following. Let $Y_1, \ldots, Y_n$ be independent random vectors in $\mathbb{R}^N$, and let $U_1, \ldots, U_n$ be independent Gaussian random vectors such that Y_i and U_i have the same mean and covariance. Given a function $f : \mathbb{R}^{Nn} \to \mathbb{C}$, we aim to bound the deviation from the Gaussian model

$$\Delta := \mathbf{E}[f(Y_1, \ldots, Y_n)] - \mathbf{E}[f(U_1, \ldots, U_n)].$$

There are various classical approaches to such problems. For example, the Lindeberg method replaces Y_i by U_i one term at a time, and then uses Taylor expansion to third order to control the error of each term; similar bounds arise from Stein's method [29, §5]. Unfortunately, in the setting of this paper such methods appear to give rise to very poor bounds. For example, in the context of Theorem 2.8, classical methods yield bounds where the parameter $\sigma_q(X)^2 \le \sigma(X)^2 := \|\sum_{i=1}^n \mathbf{E}Z_i^2\|$ is replaced by at least $\sum_{i=1}^n \mathbf{E}\|Z_i\|^2$, which is typically much larger.

The reason for the inefficiency of classical approaches to universality is that they require the independent variables to be bounded term by term. In the present setting, bounding the contribution of each summand Z_i in (2.1) separately ignores the noncommutativity of the summands. To surmount this problem, we will work instead with an exact formula for the deviation Δ in terms of a series expansion in the cumulants of the underlying variables. For our purposes, the advantage of this exact formula is that it will enable us to keep the summands Z_i together, and estimate the resulting terms efficiently using trace inequalities without destroying their noncommutativity. The price we pay for this is that we must expand the deviation Δ to high order in order to obtain efficient estimates.

In the univariate case $N = 1$, the cumulant expansion dates back to the work of Barbour [14], and has been routinely applied to the study of random matrices with independent entries since the work of Lytova and Pastur [47]. In the remainder of this section, we recall the relevant arguments of [14, 47] and spell out their immediate extension to the multivariate case $N > 1$.

4.1. **Cumulants.** Let $W_1, \ldots, W_m$ be bounded real-valued random variables. Then their log-moment generating function is analytic with power series expansion

$$\log \mathbf{E}[e^{\sum_{i=1}^m t_i W_i}] = \sum_{k=1}^{\infty} \sum_{j_1,\ldots,j_k=1}^{m} \frac{1}{k!}\, \kappa(W_{j_1}, \ldots, W_{j_k})\, t_{j_1} \cdots t_{j_k}.$$

The coefficient $\kappa(W_1, \ldots, W_k)$ is called the *joint cumulant* of the random variables $W_1, \ldots, W_k$. Joint cumulants are multilinear in their arguments and invariant under permutation of their arguments. Moreover, for jointly Gaussian random variables, all joint cumulants of order $k \geq 3$ vanish.

For any subset $J \subseteq [m] := \{1, \ldots, m\}$, denote by $W_J := (W_j)_{j \in J}$ the associated subset of random variables. Moreover, denote by $\mathrm{P}([m])$ the collection of all partitions of $[m]$. The following fundamental result [57, Proposition 3.2.1] expresses the relation between joint cumulants and moments.

Lemma 4.1 (Leonov-Shiryaev)**.** *We can write*

$$\mathbf{E}[W_1 \cdots W_m] = \sum_{\pi \in \mathrm{P}([m])} \prod_{J \in \pi} \kappa(W_J).$$

Conversely, we have

$$\kappa(W_1, \ldots, W_m) = \sum_{\pi \in \mathrm{P}([m])} (-1)^{|\pi|-1} (|\pi| - 1)! \prod_{J \in \pi} \mathbf{E}\Bigg[\prod_{j \in J} W_j\Bigg].$$

The significance of cumulants for our purposes is the following identity. The univariate ($m = 1$) case was proved in [14, Lemma 1] and [47, Proposition 3.1]; the multivariate case follows precisely in the same manner.

Lemma 4.2. *For any polynomial $f : \mathbb{R}^m \to \mathbb{C}$ and $i \in [m]$, we have*

$$\mathbf{E}[W_i f(W_1, \ldots, W_m)] = \sum_{k=0}^{\infty} \sum_{j_1,\ldots,j_k=1}^{m} \frac{1}{k!}\, \kappa(W_i, W_{j_1}, \ldots, W_{j_k})\, \mathbf{E}\bigg[\frac{\partial^k f}{\partial x_{j_1} \cdots \partial x_{j_k}}(W_1, \ldots, W_k)\bigg].$$

Proof. Let $\varphi(x_1,\ldots,x_m) := e^{\sum_{j=1}^m t_j x_j}$. Then

$$\begin{aligned}\mathbf{E}[W_i\varphi(W_1,\ldots,W_m)] &= \mathbf{E}[\varphi(W_1,\ldots,W_m)]\frac{\partial}{\partial t_i}\log\mathbf{E}[e^{\sum_{j=1}^m t_j W_j}] \\ &= \sum_{k=0}^\infty \sum_{j_1,\ldots,j_k=1}^m \frac{1}{k!}\kappa(W_i,W_{j_1},\ldots,W_{j_k})\,t_{j_1}\cdots t_{j_k}\,\mathbf{E}[\varphi(W_1,\ldots,W_m)] \\ &= \sum_{k=0}^\infty \sum_{j_1,\ldots,j_k=1}^m \frac{1}{k!}\kappa(W_i,W_{j_1},\ldots,W_{j_k})\,\mathbf{E}\bigg[\frac{\partial^k\varphi}{\partial x_{j_1}\cdots\partial x_{j_k}}(W_1,\ldots,W_k)\bigg].\end{aligned}$$

As any monomial is given by $W_{i_1}\cdots W_{i_l} = \frac{\partial^l}{\partial t_{i_1}\cdots\partial t_{i_l}}\varphi(W_1,\ldots,W_m)\big|_{t_1,\ldots,t_m=0}$, the conclusion follows readily by differentiating the above identity. □

Note that the first two cumulants are given by $\kappa(W)=\mathbf{E}[W]$ and $\kappa(W_1,W_2)=\mathrm{Cov}(W_1,W_2)$. Thus if $W_1,\ldots,W_m$ are centered and jointly Gaussian (so that the cumulants of order $k\ge 3$ vanish), the identities of Lemmas 4.1 and 4.2 reduce to

$$\mathbf{E}[W_1\cdots W_m] = \sum_{\pi\in\mathrm{P}_2([m])}\prod_{\{i,j\}\in\pi}\mathrm{Cov}(W_i,W_j) \tag{4.1}$$

(where $\mathrm{P}_2([m])$ is the collection of pair partitions of $[m]$) and

$$\mathbf{E}[W_i f(W_1,\ldots,W_m)] = \sum_{j=1}^m \mathrm{Cov}(W_i,W_j)\,\mathbf{E}\bigg[\frac{\partial f}{\partial x_j}(W_1,\ldots,W_m)\bigg]. \tag{4.2}$$

These are none other than the well-known Wick formula and integration by parts formula for centered Gaussian measures.

4.2. **Cumulant expansion.** We can now express the basic principle that will be used to prove universality. This principle is a direct extension of the method of [14, 47] to the multivariate case; see, e.g., [47, Corollary 3.1].

Theorem 4.3. *Let $Y_1,\ldots,Y_n$ be independent centered and bounded random vectors in $\mathbb{R}^N$, and let $U_1,\ldots,U_n$ be independent centered Gaussian random vectors in $\mathbb{R}^N$ such that Y_i and U_i have the same covariance. Assume that $Y=(Y_1,\ldots,Y_n)$ and $U=(U_1,\ldots,U_n)$ are independent of each other, and define*

$$Y(t) := \sqrt{t}\,Y + \sqrt{1-t}\,U.$$

Then we have

$$\frac{d}{dt}\mathbf{E}[f(Y(t))] = \frac{1}{2}\sum_{k=3}^\infty\sum_{i=1}^n\sum_{j_1,\ldots,j_k=1}^N \frac{t^{\frac{k}{2}-1}}{(k-1)!}\kappa(Y_{ij_1},\ldots,Y_{ij_k})\,\mathbf{E}\bigg[\frac{\partial^k f}{\partial y_{ij_1}\cdots y_{ij_k}}(Y(t))\bigg]$$

for any polynomial $f:\mathbb{R}^{Nn}\to\mathbb{C}$ and $t\in[0,1]$.

Proof. We readily compute

$$\frac{d}{dt}\mathbf{E}[f(Y(t))] = \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^N\bigg\{\frac{1}{\sqrt{t}}\mathbf{E}\bigg[Y_{ij}\frac{\partial f}{\partial y_{ij}}(Y(t))\bigg] - \frac{1}{\sqrt{1-t}}\mathbf{E}\bigg[U_{ij}\frac{\partial f}{\partial y_{ij}}(Y(t))\bigg]\bigg\}.$$

The conclusion follows by applying Lemma 4.2 conditionally on $\{U,(Y_k)_{k\ne i}\}$ to compute the first term in the sum, and applying (4.2) conditionally on $\{Y,(U_k)_{k\ne i}\}$ to compute the second term in the sum. □

The model $Y(t)$ should be viewed as an interpolation between the original model Y and the associated Gaussian model U. In particular, Theorem 4.3 yields a bound on the Gaussian deviation by the fundamental theorem of calculus

$$\mathbf{E}[f(Y)] - \mathbf{E}[f(U)] = \int_0^1 \frac{d}{dt}\mathbf{E}[f(Y(t))]\,dt.$$

We will however often find it necessary to perform a change of variables before applying the fundamental theorem of calculus.

When the function f is not a polynomial, it must be approximated by a polynomial before we can apply Theorem 4.3. The following result is a straightforward combination of Theorem 4.3 with Taylor expansion to order $p-1$.

Corollary 4.4. *Let $Y_1,\ldots,Y_n$ be independent centered and bounded random vectors in $\mathbb{R}^N$, and let $U_1,\ldots,U_n$ be independent centered Gaussian random vectors in $\mathbb{R}^N$ such that Y_i and U_i have the same covariance. Assume that $Y=(Y_1,\ldots,Y_n)$ and $U=(U_1,\ldots,U_n)$ are independent of each other, and define*

$$Y(t) := \sqrt{t}\,Y + \sqrt{1-t}\,U.$$

Let $p\geq 3$ and $f:\mathbb{R}^{Nn}\to\mathbb{C}$ be a smooth function. Then we have

$$\frac{d}{dt}\mathbf{E}[f(Y(t))] = \frac{1}{2}\sum_{k=3}^{p-1}\sum_{i=1}^{n}\sum_{j_1,\ldots,j_k=1}^{N}\frac{t^{\frac{k}{2}-1}}{(k-1)!}\,\kappa(Y_{ij_1},\ldots,Y_{ij_k})\,\mathbf{E}\bigg[\frac{\partial^k f}{\partial y_{ij_1}\cdots y_{ij_k}}(Y(t))\bigg] + \mathcal{R}$$

for any $t\in[0,1]$, where the reminder term satisfies

$$|\mathcal{R}| \lesssim \sup_{s,t\in[0,1]}\Bigg\{\Bigg|\sum_{i=1}^{n}\sum_{j_1,\ldots,j_p=1}^{N}\mathbf{E}\bigg[Y_{ij_1}\cdots Y_{ij_p}\frac{\partial^p f}{\partial y_{ij_1}\cdots y_{ij_p}}(Y(t,i,s))\bigg]\Bigg| + \max_{2\leq k\leq p-1}\Bigg|\sum_{i=1}^{n}\sum_{j_1,\ldots,j_p=1}^{N}\frac{\kappa(Y_{ij_1},\ldots,Y_{ij_k})}{(k-1)!}\,\mathbf{E}\bigg[Y_{ij_{k+1}}\cdots Y_{ij_p}\frac{\partial^p f}{\partial y_{ij_1}\cdots y_{ij_p}}(Y(t,i,s))\bigg]\Bigg|\Bigg\}$$

with $Y_j(t,i,s) := s^{1_{i=j}}\sqrt{t}\,Y_j + \sqrt{1-t}\,U_j$.

Proof. Let $g:\mathbb{R}^{Nn}\to\mathbb{C}$ be a smooth function, and let g_i be the Taylor expansion of $t\mapsto g(y_1,\ldots,y_{i-1},ty_i,y_{i+1},\ldots,y_n)$ to order $p-1$ around 0 (evaluated at $t=1$):

$$g_i(y) := \sum_{l=0}^{p-1}\sum_{j_1,\ldots,j_l=1}^{N}\frac{1}{l!}\,y_{ij_1}\cdots y_{ij_l}\,\frac{\partial^l g}{\partial y_{ij_1}\cdots\partial y_{ij_l}}(y_1,\ldots,y_{i-1},0,y_{i+1},\ldots,y_n).$$

Then

$$\frac{\partial^k g}{\partial y_{ij_1}\cdots\partial y_{ij_k}}(y) = \frac{\partial^k g_i}{\partial y_{ij_1}\cdots\partial y_{ij_k}}(y) + \int_0^1\frac{(1-s)^{p-k-1}}{(p-k-1)!}\cdot \sum_{j_{k+1},\ldots,j_p=1}^{N} y_{ij_{k+1}}\cdots y_{ij_p}\,\frac{\partial^p g}{\partial y_{ij_1}\cdots\partial y_{ij_p}}(y_1,\ldots,y_{i-1},sy_i,y_{i+1},\ldots,y_n)\,ds$$

for all $0 \le k \le p-1$ and $j_1, \ldots, j_k$. Choosing $g(Y) := f(\sqrt{t}\,Y + \sqrt{1-t}\,U)$ yields

$$\begin{aligned}\frac{1}{2\sqrt{t}} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{E}\bigg[Y_{ij}\frac{\partial f}{\partial y_{ij}}(Y(t))\bigg] &= \frac{1}{2t}\sum_{i=1}^{n}\sum_{j=1}^{N}\mathbf{E}\bigg[Y_{ij}\frac{\partial g}{\partial y_{ij}}(Y)\bigg] \\ &= \frac{1}{2t}\sum_{i=1}^{n}\sum_{j=1}^{N}\mathbf{E}\bigg[Y_{ij}\frac{\partial g_i}{\partial y_{ij}}(Y)\bigg] + \mathcal{R}_1,\end{aligned}$$

where

$$\mathcal{R}_1 = \frac{t^{\frac{p}{2}-1}}{2}\int_0^1 \frac{(1-s)^{p-2}}{(p-2)!}\sum_{i=1}^{n}\sum_{j_1,\ldots,j_p=1}^{N}\mathbf{E}\bigg[Y_{ij_1}\cdots Y_{ij_p}\frac{\partial^p f}{\partial u_{ij_1}\cdots\partial y_{ij_p}}(Y(t,i,s))\bigg]\,ds.$$

As $y_i \mapsto g_i(y)$ is a polynomial of degree $p-1$ and $\kappa(Y_{ij}) = \mathbf{E}[Y_{ij}] = 0$, we can now apply Lemma 4.2 conditionally on $\{U, (Y_k)_{k\neq i}\}$ to compute

$$\begin{aligned}&\frac{1}{2t}\sum_{i=1}^{n}\sum_{j=1}^{N}\mathbf{E}\bigg[Y_{ij}\frac{\partial g_i}{\partial y_{ij}}(Y)\bigg] \\ &= \frac{1}{2t}\sum_{k=2}^{p-1}\sum_{i=1}^{n}\sum_{j_1,\ldots,j_k=1}^{N}\frac{1}{(k-1)!}\,\kappa(Y_{ij_1},\ldots,Y_{ij_k})\,\mathbf{E}\bigg[\frac{\partial g_i}{\partial y_{ij_1}\cdots\partial y_{ij_k}}(Y)\bigg] \\ &= \frac{1}{2}\sum_{k=2}^{p-1}\sum_{i=1}^{n}\sum_{j_1,\ldots,j_k=1}^{N}\frac{t^{\frac{k}{2}-1}}{(k-1)!}\,\kappa(Y_{ij_1},\ldots,Y_{ij_k})\,\mathbf{E}\bigg[\frac{\partial f}{\partial y_{ij_1}\cdots\partial y_{ij_k}}(Y(t))\bigg] - \mathcal{R}_2,\end{aligned}$$

where

$$\begin{aligned}\mathcal{R}_2 = \frac{t^{\frac{p}{2}-1}}{2}\sum_{k=2}^{p-1}\int_0^1\frac{(1-s)^{p-k-1}}{(p-k-1)!}\,\cdot \\ \sum_{i=1}^{n}\sum_{j_1,\ldots,j_p=1}^{N}\frac{\kappa(Y_{ij_1},\ldots,Y_{ij_k})}{(k-1)!}\,\mathbf{E}\bigg[Y_{ij_{k+1}}\cdots Y_{ij_p}\frac{\partial^p f}{\partial y_{ij_1}\cdots\partial y_{ij_p}}(Y(t,i,s))\bigg]\,ds.\end{aligned}$$

Thus the identity in the statement follows precisely as in the proof of Theorem 4.3 with $\mathcal{R} = \mathcal{R}_1 - \mathcal{R}_2$. The estimate on $|\mathcal{R}|$ now follows readily by noting that

$$\sum_{k=1}^{p-1}\int_0^1\frac{(1-s)^{p-k-1}}{(p-k-1)!}\,ds = \sum_{k=1}^{p-1}\frac{1}{(p-k)!} \le e-1,$$

concluding the proof. □

5. Basic tools

The aim of this section is to develop a two important tools that will be needed in the proofs of our main results. In section 5.1, we prove a trace inequality that will enable us to control the derivatives that arise in the cumulant expansion of various spectral statistics. In section 5.2, we develop concentration of measure inequalities for the resolvent and for more general spectral statistics.

5.1. **A trace inequality.** Let $L_p(S_p^d)$ be the Banach space of $d\times d$ random matrices M (that is, $\mathrm{M}_d(\mathbb{C})$-valued random variables on an underlying probability space that we consider fixed throughout the paper) such that $\|M\|_p<\infty$, where

$$\|M\|_p := \begin{cases} \mathbf{E}[\operatorname{tr}|M|^p]^{\frac{1}{p}} & \text{if } 1\le p<\infty,\\ \|\|\|M\|\|\|_\infty & \text{if } p=\infty.\end{cases}$$

In particular, we can write

$$\sigma_q(X) := \left\|\left(\sum_{i=1}^n \mathbf{E}Z_i^2\right)^{\frac{1}{2}}\right\|_q, \qquad R_q(X) := \left(\sum_{i=1}^n \|Z_i\|_q^q\right)^{\frac{1}{q}}$$

for $q<\infty$ (cf. section 2.1.4).

The following trace inequality will play a key role throughout this paper.

Proposition 5.1. *Fix $k\ge 2$. Let $(Z_{ij})_{i\in[n],j\in[k]}$ be a collection of (possibly dependent) $d\times d$ self-adjoint random matrices such that Z_{ij} has the same distribution as Z_i for each i,j. Let $1\le p_1,\ldots,p_k,q\le\infty$ satisfy $\sum_{j=1}^k \frac{1}{p_j}=1-\frac{k}{q}$. Then*

$$\left|\sum_{i=1}^n \mathbf{E}[\operatorname{tr} Z_{i1}Y_1Z_{i2}Y_2\cdots Z_{ik}Y_k]\right| \le R_q(X)^{\frac{(k-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-k)}{q-2}}\prod_{j=1}^k \|Y_j\|_{p_j}$$

for any (possibly dependent) $d\times d$ random matrices $Y_1,\ldots,Y_k$ that are independent of the random matrices $(Z_{ij})_{i\in[n],j\in[k]}$.

In preparation for the proof of this result, we recall some fundamental tools that will be needed below. We first state a variant of the Riesz-Thorin interpolation theorem for Schatten classes. (The application of complex interpolation in this context was inspired by [69], and was previously used in [9, Lemma 4.5].)

Lemma 5.2. *Let $F:(L_\infty(S_\infty^d))^k\to\mathbb{C}$ be a multilinear functional. Then the map*

$$\left(\frac{1}{p_1},\ldots,\frac{1}{p_k}\right)\mapsto \log\sup_{M_1,\ldots,M_k}\frac{|F(M_1,\ldots,M_k)|}{\|M_1\|_{p_1}\cdots\|M_k\|_{p_k}}$$

is convex on $[0,1]^k$.

Proof. This follows immediately from the classical complex interpolation theorem for multilinear maps [23, §10.1] and the fact that the spaces $L_p(S_p^d)$ form a complex interpolation scale $L_r(S_r^d)=(L_p(S_p^d),L_q(S_q^d))_\theta$ with $\frac{1}{r}=\frac{1-\theta}{p}+\frac{\theta}{q}$ [59, §2]. □

Next, we recall a Hölder inequality for Schatten classes. We include a proof in order to illustrate Lemma 5.2; the same method will be used again below.

Lemma 5.3. *Let $1\le p_1,\ldots,p_k\le\infty$ satisfy $\sum_{i=1}^k\frac{1}{p_i}=1$. Then*

$$|\mathbf{E}[\operatorname{tr} Y_1\cdots Y_k]|\le \|Y_1\|_{p_1}\cdots\|Y_k\|_{p_k}$$

for any $d\times d$ random matrices $Y_1,\ldots,Y_k$.

Proof. We must show that $F(Y_1,\ldots,Y_k):=\mathbf{E}[\operatorname{tr} Y_1\cdots Y_k]$ satisfies

$$\sup_{Y_1,\ldots,Y_k}\frac{|F(Y_1,\ldots,Y_k)|}{\|Y_1\|_{p_1}\cdots\|Y_k\|_{p_k}}\le 1 \qquad \text{for all } \left(\frac{1}{p_1},\ldots,\frac{1}{p_k}\right)\in\Delta,$$

where $\Delta:=\{x\in[0,1]:\sum_{i=1}^k x_i=1\}$. By Lemma 5.2, it suffices to prove the claim only for the extreme points of Δ, that is, when $p_i=1$ and $p_j=\infty$, $j\neq i$ for some i. But the latter case is elementary, as $|XY|^2 = Y^*X^*XY \le \|X\|^2|Y|^2$ and thus $|F(Y_1,\ldots,Y_k)| \le \|Y_{i+1}\cdots Y_kY_1\cdots Y_{i-1}\|_\infty\|Y_i\|_1 \le \|Y_i\|_1\prod_{j\ne i}\|Y_j\|_\infty$. □

Finally, we recall without proof the Lieb-Thirring inequality [26, Theorem 7.4].

Lemma 5.4. *Let Y, Z be $d\times d$ positive semidefinite random matrices. Then*

$$\mathbf{E}[\mathrm{tr}\,(ZYZ)^r] \le \mathbf{E}[\mathrm{tr}\,Z^rY^rZ^r]$$

for every $1 \le r < \infty$.

We can now proceed to the proof of Proposition 5.1.

Proof of Proposition 5.1. Throughout the proof we will assume without loss of generality that $R_q(X) < \infty$, as the conclusion is trivial otherwise.

Step 1. The assumption $R_q(X) < \infty$ implies that

$$F(Y_1,\ldots,Y_k) := \sum_{i=1}^n \mathbf{E}[\mathrm{tr}\,Z_{i1}Y_1Z_{i2}Y_2\cdots Z_{ik}Y_k]$$

defines a multilinear functional on $L_\infty(S_\infty^d)$, where it is implicit in the notation that $(Y_1,\ldots,Y_k)$ are taken to be independent of (Z_{ij}). Our aim is to show that

$$\sup_{Y_1,\ldots,Y_k}\frac{|F(Y_1,\ldots,Y_k)|}{\|Y_1\|_{p_1}\cdots\|Y_k\|_{p_k}} \le R_q(X)^{\frac{(k-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-k)}{q-2}}$$

for all $(\frac{1}{p_1},\ldots,\frac{1}{p_k}) \in \Delta := \{x\in[0,1]^k : \sum_{i=1}^k x_i = 1-\frac{k}{q}\}$. By Lemma 5.2, it suffices to prove the claim only for $(\frac{1}{p_1},\ldots,\frac{1}{p_k})$ that are extreme points of the simplex Δ, that is, when $p_i = \frac{q}{q-k}$ and $p_j=\infty$, $j\ne i$ holds for some i.

By cyclic permutation of the trace, it suffices to consider the case $p_1,\ldots,p_{k-1} = \infty$ and $p_k = \frac{q}{q-k}$. To further simplify the statement to be proved, let I be a random variable that is uniformly distributed on $[n]$ and is independent of (Y_j, Z_{ij}), and define the random matrices $\mathbf{Z}_j := Z_{Ij}$. Then it suffices to show that

$$n|\mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kY_k]| \le R_q(X)^{\frac{(k-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-k)}{q-2}}$$

whenever $\|Y_1\|_\infty = \cdots = \|Y_{k-1}\|_\infty = 1$ and $\|Y_k\|_{\frac{q}{q-k}} = 1$. In the remainder of the proof, we fix $Y_1,\ldots,Y_k$ satisfying the latter assumptions.

Step 2. The assumptions on $k, p_1,\ldots,p_k, q$ imply that $q\ge k\ge 2$. In the case that $q=k$, we can estimate using Lemma 5.3

$$n|\mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kY_k]| \le n\|\mathbf{Z}_1\|_k\|Y_1\|_\infty\cdots\|\mathbf{Z}_k\|_k\|Y_k\|_\infty = R_k(X)^k,$$

completing the proof. We therefore assume in the rest of the proof that $q>k$.

Step 3. Suppose k is even. Denote by $\mathbf{Z}_j = \mathbf{U}_j|\mathbf{Z}_j|$ and $Y_k = V_k|Y_k|$ the polar decompositions of $\mathbf{Z}_j$ and Y_k, respectively. Then we can estimate for $r \geq 1$

$$
\begin{aligned}
&|\mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kY_k]| = |\mathbf{E}[\mathrm{tr}\,|Y_k|^{\frac{1}{2}}\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kV_k|Y_k|^{\frac{1}{2}}]| \\
&\leq \mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\cdots\mathbf{Z}_{\frac{k}{2}}Y_{\frac{k}{2}}Y_{\frac{k}{2}}^*\mathbf{Z}_{\frac{k}{2}}\cdots Y_1^*\mathbf{Z}_1|Y_k|]^{\frac{1}{2}}\cdot \\
&\qquad \mathbf{E}[\mathrm{tr}\,\mathbf{Z}_kY_{k-1}^*\mathbf{Z}_{k-1}\cdots Y_{\frac{k}{2}+1}^*\mathbf{Z}_{\frac{k}{2}+1}\mathbf{Z}_{\frac{k}{2}+1}Y_{\frac{k}{2}+1}\cdots\mathbf{Z}_{k-1}Y_{k-1}\mathbf{Z}_kV_k|Y_k|V_k^*]^{\frac{1}{2}} \\
&= \mathbf{E}[\mathrm{tr}\,|\mathbf{Z}_1|^{1-\frac{1}{r}}\mathbf{U}_1^*Y_1\mathbf{Z}_2\cdots Y_{\frac{k}{2}-1}\mathbf{Z}_{\frac{k}{2}}Y_{\frac{k}{2}}\cdot \\
&\qquad\qquad Y_{\frac{k}{2}}^*\mathbf{Z}_{\frac{k}{2}}Y_{\frac{k}{2}-1}^*\cdots\mathbf{Z}_2Y_1^*\mathbf{U}_1|\mathbf{Z}_1|^{1-\frac{1}{r}}|\mathbf{Z}_1|^{\frac{1}{r}}|Y_k||\mathbf{Z}_1|^{\frac{1}{r}}]^{\frac{1}{2}}\cdot \\
&\quad \mathbf{E}[\mathrm{tr}\,|\mathbf{Z}_k|^{1-\frac{1}{r}}\mathbf{U}_k^*Y_{k-1}^*\mathbf{Z}_{k-1}\cdots Y_{\frac{k}{2}+1}^*\mathbf{Z}_{\frac{k}{2}+1}\cdot \\
&\qquad\qquad \mathbf{Z}_{\frac{k}{2}+1}Y_{\frac{k}{2}+1}\cdots\mathbf{Z}_{k-1}Y_{k-1}\mathbf{U}_k|\mathbf{Z}_k|^{1-\frac{1}{r}}|\mathbf{Z}_k|^{\frac{1}{r}}V_k|Y_k|V_k^*|\mathbf{Z}_k|^{\frac{1}{r}}]^{\frac{1}{2}}
\end{aligned}
$$

by Cauchy-Schwarz. Now let

$$r = \frac{q-2}{q-k} \in [1,\infty).$$

Then we have $2\frac{1-\frac{1}{r}}{q} + (k-2)\frac{1}{q} + \frac{1}{r} = 1$. We can therefore estimate

$$
\begin{aligned}
&|\mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kY_k]| \\
&\leq \|\mathbf{Z}_1\|_q^{1-\frac{1}{r}}\|\mathbf{Z}_2\|_q\cdots\|\mathbf{Z}_{\frac{k}{2}}\|_q\,\||\mathbf{Z}_1|^{\frac{1}{r}}|Y_k||\mathbf{Z}_1|^{\frac{1}{r}}\|_r^{\frac{1}{2}}\cdot \\
&\qquad \|\mathbf{Z}_k\|_q^{1-\frac{1}{r}}\|\mathbf{Z}_{\frac{k}{2}+1}\|_q\cdots\|\mathbf{Z}_{k-1}\|_q\,\||\mathbf{Z}_k|^{\frac{1}{r}}V_k|Y_k|V_k^*|\mathbf{Z}_k|^{\frac{1}{r}}\|_r^{\frac{1}{2}} \\
&= n^{-\frac{k-2}{q-2}}R_q(X)^{\frac{(k-2)q}{q-2}}\,\||\mathbf{Z}_1|^{\frac{1}{r}}|Y_k||\mathbf{Z}_1|^{\frac{1}{r}}\|_r^{\frac{1}{2}}\,\||\mathbf{Z}_k|^{\frac{1}{r}}V_k|Y_k|V_k^*|\mathbf{Z}_k|^{\frac{1}{r}}\|_r^{\frac{1}{2}}
\end{aligned}
$$

by Lemma 5.3, where we used that $\|\mathbf{Z}_j\|_q = n^{-\frac{1}{q}}R_q(X)$ and that $k - \frac{2}{r} = \frac{(k-2)q}{q-2}$. On the other hand, using Lemma 5.4 we obtain

$$\||\mathbf{Z}_1|^{\frac{1}{r}}|Y_k||\mathbf{Z}_1|^{\frac{1}{r}}\|_r^r \leq \mathbf{E}[\mathrm{tr}\,|Y_k|^r\mathbf{Z}_1^2] = \mathbf{E}[\mathrm{tr}\,|Y_k|^r\mathbf{E}[\mathbf{Z}_1^2]] \leq \|\mathbf{E}[\mathbf{Z}_1^2]\|_{\frac{q}{2}},$$

where we used that Y_k and $\mathbf{Z}_1$ are independent and $\||Y_k|^r\|_{\frac{q}{q-2}} = \|Y_k\|_{\frac{q}{q-k}}^r = 1$. The analogous term involving $\mathbf{Z}_k$ is estimated identically. We therefore obtain

$$|\mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kY_k]| \leq n^{-1}R_q(X)^{\frac{(k-2)q}{q-2}}\,\sigma_q(X)^{\frac{2(q-k)}{q-2}},$$

where we used that $\|\mathbf{E}[\mathbf{Z}_j^2]\|_{\frac{q}{2}} = n^{-1}\sigma_q(X)^2$ for all j. This concludes the proof of the inequality for the case that k is even.

Step 4. Finally, suppose k is odd. Then we apply Cauchy-Schwarz as follows:

$$
\begin{aligned}
&|\mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kY_k]| \\
&= |\mathbf{E}[\mathrm{tr}\,|Y_k|^{\frac{1}{2}}\mathbf{Z}_1Y_1\cdots\mathbf{Z}_{\frac{k-1}{2}}Y_{\frac{k-1}{2}}\mathbf{U}_{\frac{k+1}{2}}|\mathbf{Z}_{\frac{k+1}{2}}|^{\frac{1}{2}}\cdot \\
&\qquad\qquad |\mathbf{Z}_{\frac{k+1}{2}}|^{\frac{1}{2}}Y_{\frac{k+1}{2}}\mathbf{Z}_{\frac{k+3}{2}}\cdots Y_{k-1}\mathbf{Z}_kV_k|Y_k|^{\frac{1}{2}}]| \\
&\leq \mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\cdots\mathbf{Z}_{\frac{k-1}{2}}Y_{\frac{k-1}{2}}\mathbf{U}_{\frac{k+1}{2}}|\mathbf{Z}_{\frac{k+1}{2}}|\mathbf{U}_{\frac{k+1}{2}}^*Y_{\frac{k-1}{2}}^*\mathbf{Z}_{\frac{k-1}{2}}\cdots Y_1^*\mathbf{Z}_1|Y_k|]^{\frac{1}{2}}\cdot \\
&\qquad \mathbf{E}[\mathrm{tr}\,\mathbf{Z}_kY_{k-1}^*\cdots\mathbf{Z}_{\frac{k+3}{2}}Y_{\frac{k+1}{2}}^*|\mathbf{Z}_{\frac{k+1}{2}}|Y_{\frac{k+1}{2}}\mathbf{Z}_{\frac{k+3}{2}}\cdots Y_{k-1}\mathbf{Z}_kV_k|Y_k|V_k^*]^{\frac{1}{2}}.
\end{aligned}
$$

The rest of the proof proceeds exactly as in the case that k is even. □

5.2. **Concentration of measure.** In the proof of our main results, it will be necessary to control the norms of the resolvents $\|(z\mathbf{1}-X)^{-1}\|$ and $\|(z\mathbf{1}-G)^{-1}\|$ simultaneously over many points $z \in \mathbb{C}$. To this end, we will exploit the fact that these quantities are strongly concentrated around their means.

For the Gaussian model G, such concentration inequalities follow from a routine application of Gaussian concentration, as we recall in section 5.2.1. However, the non-Gaussian model X does not appear to be amenable to off-the-shelf concentration inequalities: while convex Lipschitz functions of sums of independent random matrices (such as the norm $\|X\|$) can be treated using concentration inequalities due to Talagrand, such methods do not apply to the non-convex function $(Z_1, \ldots, Z_n) \mapsto \|(z\mathbf{1}-X)^{-1}\|$. In section 5.2.2, we develop a specialized concentration inequality that will play a key role in the proofs of our main results.

Finally, in section 5.2.3, we obtain concentration inequalities for the spectral statistics $\langle v, \varphi(X)w\rangle$ both in the Gaussian and non-Gaussian situations, which may be used in conjunction with Theorem 2.9 to obtain high probability universality bounds for spectral statistics. The proofs of these concentration inequalities rely on concentration of the resolvent as derived in the previous sections. An analogous concentration inequality for moments, which may be used in conjunction with Theorem 2.8, is much simpler and follows from a routine application of Talagrand's concentration inequality; such an inequality is given in Lemma 9.20.

5.2.1. *Resolvent norm: the Gaussian case.* The Gaussian random matrix G is amenable to a routine application of Gaussian concentration [21, Theorem 5.6] as in [9, Lemma 6.5]. For completeness, we spell out the argument.

Lemma 5.5. *Fix $z \in \mathbb{C}$ with $\operatorname{Im} z > 0$. Then we have for any $x \geq 0$*

$$\mathbf{P}\left[\big|\|(z\mathbf{1}-G)^{-1}\| - \mathbf{E}\|(z\mathbf{1}-G)^{-1}\|\big| \geq \frac{\sigma_*(X)}{(\operatorname{Im} z)^2}\, x\right] \leq 2e^{-x^2/2}.$$

Proof. Without loss of generality, we may express

$$G = A_0 + \sum_{i=1}^N g_i A_i$$

for some deterministic $A_0, \ldots, A_N \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ and i.i.d. standard Gaussian variables $g_1, \ldots, g_N$ (cf. Remark 2.3). Now consider the function $f : \mathbb{R}^N \to \mathbb{R}$ defined by

$$f(x) := \left\| \left(z\mathbf{1} - A_0 - \sum_{i=1}^N x_i A_i \right)^{-1} \right\|.$$

As $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$ for invertible matrices A, B, and as $\|(z\mathbf{1}-Y)^{-1}\| \leq (\operatorname{Im} z)^{-1}$ for any self-adjoint matrix Y, we obtain

$$\begin{aligned} |f(x) - f(y)| &\leq \left\| \left(z\mathbf{1} - A_0 - \sum_{i=1}^N x_i A_i \right)^{-1} - \left(z\mathbf{1} - A_0 - \sum_{i=1}^N y_i A_i \right)^{-1} \right\| \\ &\leq \frac{1}{(\operatorname{Im} z)^2} \left\| \sum_{i=1}^N (x_i - y_i) A_i \right\| \leq \frac{\sigma_*(X)}{(\operatorname{Im} z)^2}\, \|x - y\|, \end{aligned}$$

where we used that

$$\begin{aligned}\left\|\sum_{i=1}^N (x_i-y_i)A_i\right\| &= \sup_{\|v\|=\|w\|=1}\left|\sum_{i=1}^N (x_i-y_i)\langle v, A_i w\rangle\right| \\ &\le \sup_{\|v\|=\|w\|=1}\left(\sum_{i=1}^N |\langle v, A_i w\rangle|^2\right)^{\frac12}\|x-y\| \\ &= \sup_{\|v\|=\|w\|=1} \mathbf{E}\big[|\langle v,(G-\mathbf{E}G)w\rangle|^2\big]^{\frac12}\|x-y\| \\ &= \sigma_*(X)\,\|x-y\|.\end{aligned}$$

Thus $\|(z\mathbf{1}-G)^{-1}\| = f(g_1,\ldots,g_N)$ is a $\frac{\sigma_*(X)}{(\mathrm{Im}\, z)^2}$-Lipschitz function of a standard Gaussian vector. The conclusion is therefore immediate from the Gaussian concentration inequality [21, Theorem 5.6], which states than an L-Lipschitz function of a standard Gaussian vector is L^2-subgaussian. □

5.2.2. *Resolvent norm: the non-Gaussian case.* We now aim to prove an analogue of Lemma 5.5 for the non-Gaussian model X. To this end, we exploit the resolvent identity to prove a specialized concentration inequality using the entropy method [21, Chapter 6]. The result takes a more complicated form than Lemma 5.5, but will nonetheless suffice for the purposes of this paper.

Proposition 5.6. *Fix $z\in\mathbb{C}$ with $\mathrm{Im}\, z>0$. Then we have*

$$\mathbf{P}\Bigg[\big|\|(z\mathbf{1}-X)^{-1}\| - \mathbf{E}\|(z\mathbf{1}-X)^{-1}\|\big| \ge \frac{\sigma_*(X)}{(\mathrm{Im}\, z)^2}\sqrt{x} + \left\{\frac{R(X)}{(\mathrm{Im}\, z)^2}+\frac{R(X)^2}{(\mathrm{Im}\, z)^3}\right\}x + \left\{\frac{R(X)^{\frac12}(\mathbf{E}\|X-\mathbf{E}X\|)^{\frac12}}{(\mathrm{Im}\, z)^2} + \frac{R(X)(\mathbf{E}\|X-\mathbf{E}X\|^2)^{\frac12}}{(\mathrm{Im}\, z)^3}\right\}\sqrt{x}\Bigg] \le 2e^{-Cx}$$

for any $x\ge 0$, where C is a universal constant.

In preparation for the proof, we begin by estimating a type of discrete gradient of the function $(Z_1,\ldots,Z_n)\mapsto \|(z\mathbf{1}-X)^{-1}\|$.

Lemma 5.7. *Let $(Z_1',\ldots,Z_n')$ be an independent copy of $(Z_1,\ldots,Z_n)$. Let X be as in (2.1) and let $X^{\sim i} := Z_0 + \sum_{j\ne i} Z_j + Z_i'$. Then*

$$\|(z\mathbf{1}-X)^{-1}\| - \|(z\mathbf{1}-X^{\sim i})^{-1}\| \le \frac{2R(X)}{(\mathrm{Im}\, z)^2}$$

for all i, and

$$\sum_{i=1}^n (\|(z\mathbf{1}-X)^{-1}\| - \|(z\mathbf{1}-X^{\sim i})^{-1}\|)_+^2 \le W$$

with

$$W := \frac{2}{(\mathrm{Im}\, z)^4}\sup_{\|v\|=\|w\|=1}\sum_{i=1}^n |\langle v,(Z_i-Z_i')w\rangle|^2 + \frac{8}{(\mathrm{Im}\, z)^6}R(X)^2\left\|\sum_{i=1}^n (Z_i-Z_i')^2\right\|.$$

Proof. The first part of the statement follows as

$$\|(z\mathbf{1}-X)^{-1}\| - \|(z\mathbf{1}-X^{\sim i})^{-1}\| \le \|(z\mathbf{1}-X)^{-1}(Z_i-Z_i')(z\mathbf{1}-X^{\sim i})^{-1}\| \le \frac{2R(X)}{(\mathrm{Im}\, z)^2}$$

using the reverse triangle inequality and $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ in the first inequality, and $\|(z\mathbf{1} - A)^{-1}\| \le (\operatorname{Im} z)^{-1}$ in the second inequality.

To prove the second part of the statement, we must estimate more carefully. Let v_*, w_* be (random) vectors in the unit sphere such that

$$\|(z\mathbf{1} - X)^{-1}\| = \sup_{\|v\|=\|w\|=1} |\langle v, (z\mathbf{1} - X)^{-1} w\rangle| = |\langle v_*, (z\mathbf{1} - X)^{-1} w_*\rangle|.$$

Then

$$\begin{aligned}
&\|(z\mathbf{1} - X)^{-1}\| - \|(z\mathbf{1} - X^{\sim i})^{-1}\| \\
&\le |\langle v_*, (z\mathbf{1} - X)^{-1} w_*\rangle| - |\langle v_*, (z\mathbf{1} - X^{\sim i})^{-1} w_*\rangle| \\
&\le |\langle v_*, (z\mathbf{1} - X)^{-1}(Z_i - Z_i')(z\mathbf{1} - X^{\sim i})^{-1} w_*\rangle| \\
&\le |\langle v_*, (z\mathbf{1} - X)^{-1}(Z_i - Z_i')(z\mathbf{1} - X)^{-1} w_*\rangle| \\
&\qquad + |\langle v_*, (z\mathbf{1} - X)^{-1}(Z_i - Z_i')(z\mathbf{1} - X^{\sim i})^{-1}(Z_i - Z_i')(z\mathbf{1} - X)^{-1} w_*\rangle|,
\end{aligned}$$

where we used twice the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$. But as we have $\|(z\mathbf{1} - X)^{-1} w_*\| \le (\operatorname{Im} z)^{-1}$ and $\|(\bar{z}\mathbf{1} - X)^{-1} v_*\| \le (\operatorname{Im} z)^{-1}$, we can estimate

$$\begin{aligned}
&\sum_{i=1}^n |\langle v_*, (z\mathbf{1} - X)^{-1}(Z_i - Z_i')(z\mathbf{1} - X)^{-1} w_*\rangle|^2 \\
&\qquad\qquad \le \frac{1}{(\operatorname{Im} z)^4} \sup_{\|v\|=\|w\|=1} \sum_{i=1}^n |\langle v, (Z_i - Z_i')w\rangle|^2.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&\sum_{i=1}^n |\langle v_*, (z\mathbf{1} - X)^{-1}(Z_i - Z_i')(z\mathbf{1} - X^{\sim i})^{-1}(Z_i - Z_i')(z\mathbf{1} - X)^{-1} w_*\rangle|^2 \\
&\le \sum_{i=1}^n \|(Z_i - Z_i')(\bar{z}\mathbf{1} - X)^{-1} v_*\|^2 \|(z\mathbf{1} - X^{\sim i})^{-1}(Z_i - Z_i')(z\mathbf{1} - X)^{-1} w_*\|^2 \\
&\le \frac{4R(X)^2}{(\operatorname{Im} z)^4} \sum_{i=1}^n \langle (z\mathbf{1} - X)^{-1} w_*, (Z_i - Z_i')^2 (z\mathbf{1} - X)^{-1} w_*\rangle \\
&\le \frac{4R(X)^2}{(\operatorname{Im} z)^6} \left\| \sum_{i=1}^n (Z_i - Z_i')^2 \right\|.
\end{aligned}$$

The conclusion follows readily using $(a + b)^2 \le 2a^2 + 2b^2$. □

Next, we bound the expectation of the random variable W.

Lemma 5.8. *Let W be defined as in Lemma 5.7. Then*

$$\mathbf{E}[W] \lesssim \frac{\sigma_*(X)^2}{(\operatorname{Im} z)^4} + \frac{R(X)\mathbf{E}\|X - \mathbf{E}X\|}{(\operatorname{Im} z)^4} + \frac{R(X)^2 \mathbf{E}\|X - \mathbf{E}X\|^2}{(\operatorname{Im} z)^6}.$$

Proof. First note that as $(a - b)^2 \le 2a^2 + 2b^2$ and as $(Z_1, \ldots, Z_n)$ and $(Z_1', \ldots, Z_n')$ are equidistributed, we can estimate

$$\mathbf{E}[W] \le \frac{4}{(\operatorname{Im} z)^4} \mathbf{E}\left[\sup_{\|v\|=\|w\|=1} \sum_{i=1}^n |\langle v, Z_i w\rangle|^2 \right] + \frac{8}{(\operatorname{Im} z)^6} R(X)^2 \mathbf{E}\left\| \sum_{i=1}^n (Z_i - Z_i')^2 \right\|.$$

To estimate the first term, we apply [21, Theorem 11.8] to obtain

$$\begin{aligned}
&\mathbf{E}\Bigg[\sup_{\|v\|=\|w\|=1}\sum_{i=1}^n (\mathrm{Re}\,\langle v, Z_i w\rangle)^2\Bigg] \\
&\le 8R(X)\,\mathbf{E}\Bigg[\sup_{\|v\|=\|w\|=1}\sum_{i=1}^n \mathrm{Re}\,\langle v, Z_i w\rangle\Bigg] + \sup_{\|v\|=\|w\|=1}\sum_{i=1}^n \mathbf{E}[(\mathrm{Re}\,\langle v, Z_i w\rangle)^2], \\
&\le 8R(X)\,\mathbf{E}\|X-\mathbf{E}X\| + \sigma_*(X)^2,
\end{aligned}$$

and analogously when the real part is replaced by the imaginary part. Thus

$$\mathbf{E}\Bigg[\sup_{\|v\|=\|w\|=1}\sum_{i=1}^n |\langle v, Z_i w\rangle|^2\Bigg] \le 2\sigma_*(X)^2 + 16R(X)\mathbf{E}\|X-\mathbf{E}X\|.$$

To estimate the second term, note that

$$\begin{aligned}
\mathbf{E}\Bigg\|\sum_{i=1}^n (Z_i - Z_i')^2\Bigg\| &= \mathbf{E}\Bigg\|\mathbf{E}_\varepsilon\Bigg[\Bigg(\sum_{i=1}^n \varepsilon_i (Z_i - Z_i')\Bigg)^2\Bigg]\Bigg\| \le \mathbf{E}\Bigg\|\sum_{i=1}^n \varepsilon_i(Z_i - Z_i')\Bigg\|^2 \\
&= \mathbf{E}\Bigg\|\sum_{i=1}^n (Z_i - Z_i')\Bigg\|^2 \le 4\mathbf{E}\|X-\mathbf{E}X\|^2,
\end{aligned}$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. random signs independent of Z, Z' and $\mathbf{E}_\varepsilon$ denotes the expectation with respect to the variables ε_i only. The first equality is trivial, the first inequality is by Jensen, the second equality holds by the exchangeability of (Z_i, Z_i'), and the second inequality follows by the triangle inequality and $(a+b)^2 \le 2a^2+2b^2$. Combining the above estimates completes the proof. □

Finally, we show that W has a self-bounding property.

Lemma 5.9. *Let W be defined as in Lemma 5.7, and define*

$$W^{\sim i} := \frac{2}{(\mathrm{Im}\,z)^4}\sup_{\|v\|=\|w\|=1}\sum_{j\ne i}|\langle v, (Z_j - Z_j')w\rangle|^2 + \frac{8}{(\mathrm{Im}\,z)^6}R(X)^2\Bigg\|\sum_{j\ne i}(Z_j - Z_j')^2\Bigg\|.$$

Then $W^{\sim i} \le W$ and

$$\sum_{i=1}^n (W - W^{\sim i})^2 \le \Bigg\{\frac{16R(X)^2}{(\mathrm{Im}\,z)^4} + \frac{64R(X)^4}{(\mathrm{Im}\,z)^6}\Bigg\}W.$$

Proof. That $W^{\sim i} \le W$ is obvious. To prove the self-bounding inequality, let u_*, v_*, w_* be (random) vectors in the unit sphere such that

$$\sup_{\|v\|=\|w\|=1}\sum_{i=1}^n |\langle v, (Z_i - Z_i')w\rangle|^2 = \sum_{i=1}^n |\langle v_*, (Z_i - Z_i')w_*\rangle|^2$$

and

$$\Bigg\|\sum_{i=1}^n (Z_i - Z_i')^2\Bigg\| = \sup_{\|u\|=1}\sum_{i=1}^n \|(Z_i - Z_i')u\|^2 = \sum_{i=1}^n \|(Z_i - Z_i')u_*\|^2.$$

Then

$$W - W^{\sim i} \le \frac{2}{(\mathrm{Im}\,z)^4}|\langle v_*, (Z_i - Z_i')w_*\rangle|^2 + \frac{8}{(\mathrm{Im}\,z)^6}R(X)^2\|(Z_i - Z_i')u_*\|^2.$$

Therefore,

$$\begin{aligned}
&\sum_{i=1}^n (W-W^{\sim i})^2 \\
&\le \frac{8}{(\mathrm{Im}\,z)^8}\sum_{i=1}^n |\langle v_*,(Z_i-Z_i')w_*\rangle|^4 + \frac{128}{(\mathrm{Im}\,z)^{12}}R(X)^4\sum_{i=1}^n \|(Z_i-Z_i')u_*\|^4 \\
&\le \frac{32}{(\mathrm{Im}\,z)^8}R(X)^2\sum_{i=1}^n |\langle v_*,(Z_i-Z_i')w_*\rangle|^2 + \frac{512}{(\mathrm{Im}\,z)^{12}}R(X)^6\sum_{i=1}^n \|(Z_i-Z_i')u_*\|^2 \\
&\le \frac{16R(X)^2}{(\mathrm{Im}\,z)^4}\cdot\frac{2}{(\mathrm{Im}\,z)^4}\sup_{\|v\|=\|w\|=1}\sum_{i=1}^n |\langle v,(Z_i-Z_i')w\rangle|^2 \\
&\qquad + \frac{64R(X)^4}{(\mathrm{Im}\,z)^6}\cdot\frac{8}{(\mathrm{Im}\,z)^6}R(X)^2\left\|\sum_{i=1}^n (Z_i-Z_i')^2\right\|.
\end{aligned}$$

The conclusion follows from the definition of W. □

We can now complete the proof of Proposition 5.6.

Proof of Proposition 5.6. We begin by noting that the self-bounding property established in Lemma 5.9 implies, by [21, Theorem 6.19], that

$$\log \mathbf{E}[e^{W/a}] \le \frac{2}{a}\,\mathbf{E}[W], \qquad a = \frac{16R(X)^2}{(\mathrm{Im}\,z)^4} + \frac{64R(X)^4}{(\mathrm{Im}\,z)^6}. \tag{5.1}$$

On the other hand, the estimate of Lemma 5.7 implies, by the exponential Poincaré inequality [21, Theorem 6.16], that for $0 \le \lambda < a^{-\frac12}$

$$\log \mathbf{E}[e^{\lambda\{\|(z\mathbf{1}-X)^{-1}\|-\mathbf{E}\|(z\mathbf{1}-X)^{-1}\|\}}] \le \frac{\lambda^2 a}{1-\lambda^2 a}\log \mathbf{E}[e^{W/a}].$$

Combining these estimates with a Chernoff bound [21, p. 29] yields

$$\mathbf{P}\big[\|(z\mathbf{1}-X)^{-1}\| \ge \mathbf{E}\|(z\mathbf{1}-X)^{-1}\| + \sqrt{8\mathbf{E}[W]x} + \sqrt{a}\,x\big] \le e^{-x}$$

for all $x \ge 0$. This yields a tail bound for deviation above the mean.

We must now prove a tail bound for deviation below the mean. This requires a variant of the second inequality of [21, Theorem 6.16], whose proof we spell out for completeness. The last inequality of [21, Theorem 6.15] and Lemma 5.7 imply

$$\mathrm{Ent}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big] \le \lambda^2\vartheta(\lambda b)\,\mathbf{E}\big[We^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big], \qquad b = \frac{2R(X)}{(\mathrm{Im}\,z)^2}$$

for $\lambda \ge 0$, where $\mathrm{Ent}(Z) := \mathbf{E}[Z\log Z] - \mathbf{E}[Z]\log\mathbf{E}[Z]$ and we used that $\vartheta(x) := \frac{e^x-1}{x}$ is a positive increasing function. In particular, as $b^2 \le a$ and $\vartheta(1) \le 2$,

$$\mathrm{Ent}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big] \le 2\lambda^2\,\mathbf{E}\big[We^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big]$$

for $0 \le \lambda \le a^{-\frac12}$. Applying the duality formula of entropy as in [21, p. 187] yields

$$\mathrm{Ent}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big] \le \frac{2\lambda^2 a}{1-2\lambda^2 a}\log\mathbf{E}[e^{W/a}]\,\mathbf{E}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big]$$

for $0 \le \lambda \le (2a)^{-\frac12}$. Therefore

$$\frac{d}{d\lambda}\left(\frac{1}{\lambda}\log\mathbf{E}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big]\right) = \frac{\mathrm{Ent}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big]}{\lambda^2\,\mathbf{E}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big]} \le \frac{2a}{1-2\lambda^2 a}\log\mathbf{E}[e^{W/a}]$$

for $0 \le \lambda \le (2a)^{-\frac{1}{2}}$. Integrating both sides yields

$$\log \mathbf{E}\big[e^{-\lambda\{\|(z\mathbf{1}-X)^{-1}\|-\mathbf{E}\|(z\mathbf{1}-X)^{-1}\|\}}\big] \le \lambda\sqrt{2a}\,\mathrm{arctanh}(\lambda\sqrt{2a})\log \mathbf{E}[e^{W/a}] \le \frac{2\lambda^2 a}{1-\lambda\sqrt{2a}}\log \mathbf{E}[e^{W/a}] \le \frac{4\mathbf{E}[W]\lambda^2}{1-\lambda\sqrt{2a}},$$

where we used $\mathrm{arctanh}(x) \le \frac{x}{1-x}$ in the second inequality and (5.1) in the last inequality. We can now apply a Chernoff bound [21, p. 29] to obtain

$$\mathbf{P}\big[\|(z\mathbf{1}-X)^{-1}\| \le \mathbf{E}\|(z\mathbf{1}-X)^{-1}\| - 4\sqrt{\mathbf{E}[W]x} - \sqrt{2a}\,x\big] \le e^{-x}$$

for all $x \ge 0$. This yields a tail bound for deviation below the mean.

To conclude the proof, it remains to combine the upper and lower tail bounds by the union bound, and to use Lemma 5.8 to estimate $\mathbf{E}[W]$. ☐

5.2.3. *Spectral statistics.* Theorem 2.9 establishes universality of the expectations of spectral statistics of the form $\langle v, \varphi(X)w\rangle$. A corresponding tail bound would follow if we can prove a concentration inequality for such spectral statistics. This problem turns out to be subtle even in the Gaussian case: even when φ is Lipschitz, the Lipschitz property of $(g_1,\ldots,g_N) \mapsto \langle v, \varphi(G)w\rangle$ is not obvious. Deep results on the latter problem [27] could be applied in the Gaussian setting, but do not appear to be sufficiently powerful to handle the non-Gaussian case.

Here we take a different approach. Using functional calculus [32, §2.2], $\varphi(X)$ can be expressed as an integral of the resolvent of X. (We will use an essentially equivalent formulation that appears in the proof of [41, Theorem 6.2].) With this representation in hand, we can readily repeat the proof of Proposition 5.6 to obtain a concentration inequality. The main result of this section is the following.

Proposition 5.10. *For $\varphi \in W^{4,1}(\mathbb{R})$ and $v, w \in \mathbb{R}^d$ with $\|v\| = \|w\| = 1$, we have*

$$\mathbf{P}\big[|\langle v, \varphi(G)w\rangle - \mathbf{E}[\langle v, \varphi(G)w\rangle]| \ge \|\varphi\|_{W^{3,1}}\sigma_*(X)\sqrt{x}\big] \le 4e^{-Cx}$$

and

$$\mathbf{P}\big[|\langle v, \varphi(X)w\rangle - \mathbf{E}[\langle v, \varphi(X)w\rangle]| \ge \|\varphi\|_{W^{4,1}}\big\{(R(X)+R(X)^2)x + (\sigma_*(X) + R(X)^{\frac{1}{2}}(\mathbf{E}\|X-\mathbf{E}X\|)^{\frac{1}{2}} + R(X)(\mathbf{E}\|X-\mathbf{E}X\|^2)^{\frac{1}{2}})\sqrt{x}\big\}\big] \le 4e^{-Cx}$$

for all $x \ge 0$, where C is a universal constant.

The basis for the proof is the following identity.

Lemma 5.11. *For any $M \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$, $p \in \mathbb{N}$, and $\varphi \in C_c^\infty(\mathbb{R})$, we have*

$$\varphi(M) = -\frac{1}{\pi}\lim_{\varepsilon\downarrow 0}\mathrm{Im}\int_{-\infty}^{\infty}\left(1+\frac{d}{dx}\right)^p \varphi(x)\times \frac{(1+i)^p}{(p-1)!}\int_0^\infty ((x+t+i(\varepsilon+t))\mathbf{1}-M)^{-1}\,t^{p-1}e^{-(1+i)t}\,dt\,dx.$$

Proof. The identity follows by following verbatim the proof of [41, Theorem 6.2], noting that we can express $\langle v, \varphi(M)v\rangle = \int \varphi(x)\,\mu_v(dx)$ for some measure μ_v (the spectral distribution of M with respect to the state $\tau_v(M) := \langle v, Mv\rangle$). ☐

We can now use the above representation to establish discrete (and continuous) gradient bounds of $\langle v, \varphi(X)w\rangle$ along the lines of Lemma 5.7.

Lemma 5.12. *Let $X = Z_0 + \sum_{j=1}^n Z_j$ as in (2.1), define $X^{\sim i}$ as in Lemma 5.7, let $G = A_0 + \sum_{i=1}^N g_i A_i$ as in the proof of Lemma 5.5, and let $\|v\| = \|w\| = 1$.*

a. The function $(g_1, \ldots, g_N) \mapsto \langle v, \varphi(G)w\rangle$ is $C\|\varphi\|_{W^{3,1}}\sigma_*(X)$-Lipschitz.

b. $|\langle v, \varphi(X)w\rangle - \langle v, \varphi(X^{\sim i})w\rangle| \lesssim \|\varphi\|_{W^{3,1}} R(X)$.

c. $\sum_{i=1}^n |\langle v, \varphi(X)w\rangle - \langle v, \varphi(X^{\sim i})w\rangle|^2 \lesssim \|\varphi\|^2_{W^{4,1}} V$.

Here C is a universal constant, and V is defined as W in Lemma 5.7 with $\mathrm{Im}\, z \leftarrow 1$.

Proof. By a routine approximation argument, we may assume that $\varphi \in C_c^\infty(\mathbb{R})$. Now note that $\langle v, (z\mathbf{1} - G)^{-1}w\rangle$ is $\frac{\sigma_*(X)}{(\mathrm{Im}\, z)^2}$-Lipschitz as in the proof of Lemma 5.5. Thus Lemma 5.11 with $p = 3$ shows that $\langle v, \varphi(G)w\rangle$ is Lipschitz with constant

$$\frac{1}{\pi} \cdot 3\|\varphi\|_{W^{3,1}} \cdot \lim_{\varepsilon \downarrow 0} \frac{(\sqrt{2})^3}{2!} \int_0^\infty \frac{\sigma_*(X)}{(\varepsilon + t)^2}\, t^2 e^{-t}\, dt \lesssim \|\varphi\|_{W^{3,1}}\sigma_*(X),$$

which establishes part a. Part b. follows in precisely the same manner using that $|\langle v, (z\mathbf{1} - X)^{-1}w\rangle - \langle v, (z\mathbf{1} - X^{\sim i})^{-1}w\rangle| \le \frac{2R(X)}{(\mathrm{Im}\, z)^2}$ as in the proof of Lemma 5.7.

For part c., we begin by applying Lemma 5.11 with $p = 4$ to estimate

$$\left(\sum_{i=1}^n |\langle v, \varphi(X)w\rangle - \langle v, \varphi(X^{\sim i})w\rangle|^2\right)^{\frac{1}{2}} \le \frac{4}{3!\pi} \int_{-\infty}^\infty \left|\left(1 + \frac{d}{dx}\right)^4 \varphi(x)\right| \times$$
$$\int_0^\infty \left(\sum_{i=1}^n |\langle v, ((x+t+it)\mathbf{1} - X)^{-1}w\rangle - \langle v, ((x+t+it)\mathbf{1} - X^{\sim i})^{-1}w\rangle|^2\right)^{\frac{1}{2}} t^3 e^{-t}\, dt\, dx,$$

where we used the triangle inequality to bring the Euclidean norm with respect to the index i inside the integral. But it follows as in the proof of Lemma 5.7 that the quantity inside the brackets on the second line is bounded by the random variable W of Lemma 5.7 with $\mathrm{Im}\, z \leftarrow t$. The conclusion follows readily. □

We can now complete the proof of Proposition 5.10.

Proof of Proposition 5.10. By rescaling φ, we may assume without loss of generality that $\|\varphi\|_{W^{3,1}} = 1$ (Gaussian case) or $\|\varphi\|_{W^{4,1}} = 1$ (non-Gaussian case). By a union bound, it suffices to consider separately the real and imaginary parts of $\langle v, \varphi(G)w\rangle$ and $\langle v, \varphi(X)w\rangle$, respectively. The conclusion now follows using Lemma 5.12 by repeating the proofs of Lemma 5.5 and Proposition 5.6 verbatim with $\mathrm{Im}\, z \leftarrow 1$. □

6. Universality of spectral statistics

The aim of this section is to prove our main universality principles for spectral statistics. The basic idea behind the proofs is that we will interpolate between the non-Gaussian and Gaussian models, and estimate the rate of change along the interpolation by means of the cumulant expansion and trace inequalities. This program will be implemented for the moments, resolvent moments, and resolvent in sections 6.2, 6.3, and 6.4, respectively. Before we do so, however, we first introduce some basic constructions that are common to all the proofs.

6.1. **Preliminaries.** We always fix a random matrix X as in (2.1), and let G be its Gaussian model. Throughout this section, we will further assume that X and G are independent of each other; this will entail no loss of generality, as the universality results that are proved in this section—Theorems 2.8, 2.9, and 6.8—are independent of the joint distribution of X and G. Define

$$X(t) := \mathbf{E}X + \sqrt{t}\,(X - \mathbf{E}X) + \sqrt{1-t}\,(G - \mathbf{E}G), \qquad t \in [0,1].$$

The random matrix $X(t)$ interpolates between $X(1) = X$ and $X(0) = G$, where the interpolation is chosen so that $\mathbf{E}X(t)$ and $\mathrm{Cov}(X(t))$ are independent of t. The basic principle behind all the proofs of this section is that we aim to compute $\frac{d}{dt}\mathbf{E}[f(X(t))]$ for the relevant spectral statistic $f : \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}} \to \mathbb{C}$ using the cumulant expansion. To this end, we will choose the random vector Y_i in the statements of Theorem 4.3 and Corollary 4.4 to be the $2d^2$-dimensional vector of the real and imaginary parts of the entries of the random matrix Z_i.

For our purposes, it will be convenient to reformulate the resulting expansions by combining them with the cumulant formula of Lemma 4.1. Before we can do so, we must introduce a simple construction that will facilitate working with the second identity of Lemma 4.1. For any $k \in \mathbb{N}$ and partition $\pi \in \mathrm{P}([k])$, define random matrices $Z_{i1|\pi}, \ldots, Z_{ik|\pi}$ ($i \in [n]$) with the following properties:

1. $(Z_{ij|\pi})_{i\in[n]}$ has the same distribution as $(Z_i)_{i\in[n]}$.
2. $(Z_{ij|\pi})_{i\in[n]} = (Z_{ik|\pi})_{i\in[n]}$ for indices j,k that belong to the same element of π.
3. $(Z_{ij|\pi})_{i\in[n]}$ are independent for indices j that belong to distinct elements of π.
4. $(Z_{ij|\pi})_{i\in[n],j\in[k]}$ is independent of X and G.

This construction will be fixed in the sequel. (We do not specify the joint distribution of these matrices for different k,π as these will not arise in the analysis.) We can now state a version of Theorem 4.3 in the present setting.

Corollary 6.1. *For any polynomial $f : \mathrm{M}_d(\mathbb{C}) \to \mathbb{C}$, we have*

$$\frac{d}{dt}\mathbf{E}[f(X(t))] = \frac{1}{2}\sum_{k=3}^{\infty}\frac{t^{\frac{k}{2}-1}}{(k-1)!}\sum_{\pi\in\mathrm{P}([k])}(-1)^{|\pi|-1}(|\pi|-1)!\,\mathbf{E}\Bigg[\sum_{i=1}^{n}\partial_{Z_{i1|\pi}}\cdots\partial_{Z_{ik|\pi}}f(X(t))\Bigg],$$

where $\partial_B f$ denotes the directional derivative of f in the direction $B \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$.

Proof. Let $\iota : \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}} \to \mathbb{R}^{2d^2}$ be defined by $\iota(M) := (\mathrm{Re}\,M_{uv}, \mathrm{Im}\,M_{uv})_{u,v\in[d]}$. Let $Y_i = \iota(Z_i)$, and define $Y(t)$ as in Theorem 4.3. Then we may write

$$\iota(X(t)) = \iota(\mathbf{E}X) + \sum_{i=1}^{n} Y_i(t).$$

In particular, we can equivalently view $f(X(t)) = f\big(\mathbf{E}X + \sum_{i=1}^{n}\iota^{-1}(Y_i(t))\big)$ as a function of $Y(t)$. Applying Theorem 4.3 to the latter yields

$$\frac{d}{dt}\mathbf{E}[f(X(t))] = \frac{1}{2}\sum_{k=3}^{\infty}\frac{t^{\frac{k}{2}-1}}{(k-1)!}\times \sum_{i=1}^{n}\sum_{\substack{(u_j,v_j,\alpha_j)\in\mathcal{I}\\ j=1,\ldots,k}}\kappa((Z_i)^{\alpha_1}_{u_1v_1},\ldots,(Z_i)^{\alpha_k}_{u_kv_k})\,\mathbf{E}\Bigg[\frac{\partial^k f}{\partial M^{\alpha_1}_{u_1v_1}\cdots\partial M^{\alpha_k}_{u_kv_k}}(X(t))\Bigg],$$

where $\mathcal{I} := [d]\times[d]\times\{\mathrm{R},\mathrm{I}\}$ and we denote $M^{\mathrm{R}}_{uv} := \mathrm{Re}\, M_{uv}$ and $M^{\mathrm{I}}_{uv} := \mathrm{Im}\, M_{uv}$. The conclusion follows by applying the second identity of Lemma 4.1 to the cumulant, and using the independence structure of $Z_{ij|\pi}$ to merge the product of expectations in the resulting identity into a single expectation. □

The following is the analogous version of Corollary 4.4.

Corollary 6.2. *For any $p \ge 3$ and smooth function $f : \mathrm{M}_d(\mathbb{C}) \to \mathbb{C}$ we have*

$$\frac{d}{dt}\mathbf{E}[f(X(t))] = \frac{1}{2}\sum_{k=3}^{p-1} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi\in\mathrm{P}([k])} (-1)^{|\pi|-1}(|\pi|-1)!\,\mathbf{E}\left[\sum_{i=1}^n \partial_{Z_{i1|\pi}}\cdots\partial_{Z_{ik|\pi}} f(X(t))\right] + \mathcal{R},$$

where the remainder term satisfies

$$|\mathcal{R}| \lesssim \sup_{s,t\in[0,1]}\left\{\left|\sum_{i=1}^n \mathbf{E}[\partial^p_{Z_i} f(X(t,i,s))]\right| + \max_{2\le k\le p-1}\left|\sum_{\pi\in\mathrm{P}([k])} \frac{(-1)^{|\pi|-1}(|\pi|-1)!}{(k-1)!}\sum_{i=1}^n \mathbf{E}[\partial^{p-k}_{Z_i}\partial_{Z_{i1|\pi}}\cdots\partial_{Z_{ik|\pi}} f(X(t,i,s))]\right|\right\}$$

with $X(t,i,s) := X(t) - (1-s)\sqrt{t}Z_i$.

Proof. The conclusion follows from Corollary 4.4 in exactly the same manner as we derived Corollary 6.1 from Theorem 4.3. □

6.2. **Moments.** The aim of this section is to show that the moments $\mathbf{E}[\mathrm{tr}\, X^{2p}]$ are close to their Gaussian analogues $\mathbf{E}[\mathrm{tr}\, G^{2p}]$. To this end, we will first compute $\frac{d}{dt}\mathbf{E}[\mathrm{tr}\, X(t)^{2p}]$ by means of the cumulant expansion, and then estimate the individual terms to obtain a differential inequality.

We begin by computing the derivatives of the moment function $M \mapsto \mathrm{tr}[M^{2p}]$.

Lemma 6.3. *Let $p \in \mathbb{N}$ and $B_1,\ldots,B_k \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$. Then*

$$\partial_{B_1}\cdots\partial_{B_k}\,\mathrm{tr}[M^{2p}] = \sum_{\sigma\in\mathrm{Sym}(k)} \sum_{\substack{r_1,\ldots,r_{k+1}\ge 0\\ r_1+\cdots+r_{k+1}=2p-k}} \mathrm{tr}[M^{r_1}B_{\sigma(1)}M^{r_2}B_{\sigma(2)}\cdots M^{r_k}B_{\sigma(k)}M^{r_{k+1}}].$$

Proof. This follows by applying the product rule k times. □

We will also need the following estimate.

Lemma 6.4. *For any $k \in \mathbb{N}$, we have*

$$\sum_{\pi\in\mathrm{P}([k])} (|\pi|-1)! \le 2^k(k-1)!.$$

Proof. We first crudely estimate

$$\sum_{\pi\in\mathrm{P}([k])} (|\pi|-1)! \le \sum_{\pi\in\mathrm{P}([k])} (|\pi|-1)! \prod_{J\in\pi} |J|!.$$

Now note that any partition of $[k]$ into m parts can be generated by first choosing $r_1,\ldots,r_m \ge 1$ such that $r_1+\cdots+r_m = k$, and then choosing disjoint sets $J_1,\ldots,J_m$

with $|J_i| = r_i$. Moreover, each distinct partition is generated precisely $m!$ times in this manner, as relabeling the sets J_i does not change the partition. Therefore

$$\sum_{\pi\in\mathrm{P}([k])}(|\pi|-1)!\prod_{J\in\pi}|J|! = \sum_{m=1}^{k}(m-1)!\,\frac{1}{m!}\sum_{\substack{r_1,\ldots,r_m\geq 1\\ r_1+\cdots+r_m=k}}\binom{k}{r_1,\ldots,r_m}\prod_{j=1}^{m}r_j!$$
$$= (k-1)!\sum_{m=1}^{k}\binom{k}{m} = (2^k-1)(k-1)!,$$

where the second equality follows as the number of m-tuples of positive integers that sum to k is $\binom{k-1}{m-1}$, and the last equality holds by the binomial theorem. □

We are now ready to apply the cumulant expansion.

Proposition 6.5. *For any $p \in \mathbb{N}$ with $p \geq 2$, $2p \leq q \leq \infty$, and $t \in [0, 1]$, we have*

$$\left|\frac{d}{dt}\mathbf{E}[\operatorname{tr} X(t)^{2p}]\right| \leq 64p^3 \max\{R_q(X)\sigma_q(X)^2, R_q(X)^3\}\times$$
$$\max\left\{\mathbf{E}[\operatorname{tr} X(t)^{2p}]^{1-\frac{3}{2p}}, (8pR_q(X))^{2p-3}\right\}.$$

Proof. Combining Corollary 6.1 and Lemma 6.3 yields

$$\frac{d}{dt}\mathbf{E}[\operatorname{tr} X(t)^{2p}] = \frac{1}{2}\sum_{k=3}^{2p}kt^{\frac{k}{2}-1}\sum_{\pi\in\mathrm{P}([k])}(-1)^{|\pi|-1}(|\pi|-1)!\times$$
$$\sum_{\substack{r_1,\ldots,r_{k+1}\geq 0\\ r_1+\cdots+r_{k+1}=2p-k}}\sum_{i=1}^{n}\mathbf{E}[\operatorname{tr} X(t)^{r_1}Z_{i1|\pi}X(t)^{r_2}Z_{i2|\pi}\cdots X(t)^{r_k}Z_{ik|\pi}X(t)^{r_{k+1}}].$$

Here we used that as $(Z_{i\sigma(j)|\pi})_{i\in[n],j\in[k]}$ and $(Z_{ij|\sigma^{-1}(\pi)})_{i\in[n],j\in[k]}$ are equidistributed for any permutation σ, we can eliminate the sum over σ in Lemma 6.3 by symmetry.

Now let $r = \frac{(2p-k)q}{q-k}$, so that $2p-k \leq r \leq 2p$. Let $p_j = \frac{r}{r_{j+1}}$ for $j < k$ and $p_k = \frac{r}{r_{k+1}+r_1}$. Then we can apply Proposition 5.1 to estimate

$$\left|\sum_{i=1}^{n}\mathbf{E}[\operatorname{tr} X(t)^{r_1}Z_{i1|\pi}X(t)^{r_2}Z_{i2|\pi}\cdots X(t)^{r_k}Z_{ik|\pi}X(t)^{r_{k+1}}]\right|$$
$$\leq R_q(X)^{\frac{(k-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-k)}{q-2}}\mathbf{E}[\operatorname{tr} X(t)^r]^{\frac{2p-k}{r}}$$
$$\leq R_q(X)^{\frac{(k-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-k)}{q-2}}\mathbf{E}[\operatorname{tr} X(t)^{2p}]^{1-\frac{k}{2p}}$$

for any $r_1, \ldots, r_{k+1} \geq 0$ with $r_1 + \cdots + r_{k+1} = 2p - k$. It follows that

$$\left|\frac{d}{dt}\mathbf{E}[\operatorname{tr} X(t)^{2p}]\right| \leq \frac{1}{2}\sum_{k=3}^{2p}(4p)^k R_q(X)^{\frac{(k-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-k)}{q-2}}\mathbf{E}[\operatorname{tr} X(t)^{2p}]^{1-\frac{k}{2p}}$$

using Lemma 6.4, $t \le 1$, and that the number of $(k+1)$-tuples of nonnegative integers that sum to $2p-k$ is $\binom{2p}{k} \le \frac{(2p)^k}{k!}$. To simplify the expression, we estimate

$$\begin{aligned}\left|\frac{d}{dt}\mathbf{E}[\mathrm{tr}\,X(t)^{2p}]\right| &\le \frac{1}{2}\sum_{k=3}^{2p} 2^{-k}(8p)^k R_q(X)^{\frac{(k-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-k)}{q-2}}\mathbf{E}[\mathrm{tr}\,X(t)^{2p}]^{1-\frac{k}{2p}} \\ &\le \frac{1}{8}\max_{3\le k\le 2p}(8p)^k R_q(X)^{\frac{(k-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-k)}{q-2}}\mathbf{E}[\mathrm{tr}\,X(t)^{2p}]^{1-\frac{k}{2p}} \\ &\le 64p^3\max\big\{R_q(X)^{\frac{q}{q-2}}\sigma_q(X)^{\frac{2(q-3)}{q-2}}\mathbf{E}[\mathrm{tr}\,X(t)^{2p}]^{1-\frac{3}{2p}}, \\ &\qquad\qquad (8p)^{2p-3}R_q(X)^{\frac{(2p-2)q}{q-2}}\sigma_q(X)^{\frac{2(q-2p)}{q-2}}\big\}.\end{aligned}$$

Here we used that the term inside the maximum on the second line is convex as a function of k, so that the maximum is attained at one of the endpoints $k \in \{3, 2p\}$. The proof is readily concluded using that $R^{\frac{q}{q-2}-1}\sigma^{\frac{2(q-3)}{q-2}} \le (\max\{R,\sigma\})^2$ and $R^{\frac{(2p-2)q}{q-2}-(2p-3)-1}\sigma^{\frac{2(q-2p)}{q-2}} \le (\max\{R,\sigma\})^2$ (as $q \ge 2p \ge 4$ implies that in both cases the exponents on the left-hand side are positive and sum to 2). ☐

It remains to solve the differential inequality in the statement of Proposition 6.5. To this end we will use the following simple lemma.

Lemma 6.6. *Let $f : [0,1] \to \mathbb{R}_+$, $C, K \ge 0$, and $\alpha \in [0,1]$. Suppose that*

$$\left|\frac{d}{dt}f(t)\right| \le C\max\{f(t)^{1-\alpha}, K^{1-\alpha}\}$$

for all $t \in [0,1]$. Then

$$|f(1)^\alpha - f(0)^\alpha| \le C\alpha + K^\alpha.$$

Proof. It follows readily by the chain rule that

$$\left|\frac{d}{dt}(f(t)+K)^\alpha\right| = \alpha(f(t)+K)^{\alpha-1}\left|\frac{d}{dt}f(t)\right| \le C\alpha,$$

so that

$$|(f(1)+K)^\alpha - (f(0)+K)^\alpha| = \left|\int_0^1 \frac{d}{dt}(f(t)+K)^\alpha\,dt\right| \le C\alpha.$$

The conclusion follows as $x^\alpha - y^\alpha \le (x+K)^\alpha - (y+K)^\alpha + K^\alpha$ for any $x, y \ge 0$. ☐

We can now conclude the proof of Theorem 2.8.

Proof of Theorem 2.8: first inequality. If $p = 1$, then $\mathbf{E}[\mathrm{tr}\,X(t)^{2p}] = \mathbf{E}[\mathrm{tr}\,X^2]$ is independent of t by construction, and the conclusion is trivial. If $p \ge 2$, we can apply Lemma 6.6 with $\alpha = \frac{3}{2p} \in [0,1]$ and Proposition 6.5 to obtain

$$|\mathbf{E}[\mathrm{tr}\,X^{2p}]^{\frac{3}{2p}} - \mathbf{E}[\mathrm{tr}\,G^{2p}]^{\frac{3}{2p}}| \le 96p^2\max\{R_q(X)\sigma_q(X)^2, R_q(X)^3\} + (8pR_q(X))^3.$$

The conclusion follows as $|x^{\frac{1}{3}} - y^{\frac{1}{3}}| \le |x-y|^{\frac{1}{3}}$ for $x, y \ge 0$. ☐

The second inequality of Theorem 2.8 follows by a slight variation of the proof.

Proof of Theorem 2.8: second inequality. Note that $\sigma_{2p}(X) = \sigma_{2p}(X(t))$ for all t, as the definition of $\sigma_{2p}(X)$ depends only on $\mathrm{Cov}(X)$. We can therefore estimate

$$\sigma_{2p}(X)^{2p} = \mathrm{tr}\,(\mathbf{E}[X(t)^2] - \mathbf{E}[X(t)]^2)^p \le \mathrm{tr}\,\mathbf{E}[X(t)^2]^p \le \mathbf{E}[\mathrm{tr}\,X(t)^{2p}]$$

for every $t \in [0,1]$. Here we used that $\mathrm{tr}\, A^p \le \mathrm{tr}\, B^p$ for $B \ge A \ge 0$, and that $A \mapsto \mathrm{tr}\, A^p$ is convex for $A \ge 0$ [26, Theorem 2.10]. Furthermore, note that

$$R_{2p}(X) \le R_2(X)^{\frac{q-2p}{p(q-2)}} R_q(X)^{\frac{(p-1)q}{p(q-2)}} \le \sigma_{2p}(X)^{\frac{q-2p}{p(q-2)}} R_q(X)^{\frac{(p-1)q}{p(q-2)}}$$

by the Riesz convexity theorem and as $R_2(X) = \sigma_2(X) \le \sigma_{2p}(X)$. Consequently, we can bound the differential inequality in the proof of Proposition 6.5 as

$$\begin{aligned}
\left|\frac{d}{dt}\mathbf{E}[\mathrm{tr}\, X(t)^{2p}]\right| &\le \frac{1}{2}\sum_{k=3}^{2p} (4p)^k R_{2p}(X)^{\frac{(k-2)p}{p-1}} \sigma_{2p}(X)^{\frac{2p-k}{p-1}} \mathbf{E}[\mathrm{tr}\, X(t)^{2p}]^{1-\frac{k}{2p}} \\
&\le \frac{1}{2}\sum_{k=3}^{2p} (4p)^k R_q(X)^{\frac{(k-2)q}{q-2}} \mathbf{E}[\mathrm{tr}\, X(t)^{2p}]^{1-\frac{k}{2p}+\frac{q-k}{p(q-2)}} \\
&\le \frac{1}{8}\max_{3\le k\le 2p} (8p)^k R_q(X)^{\frac{(k-2)q}{q-2}} \mathbf{E}[\mathrm{tr}\, X(t)^{2p}]^{1-\frac{k}{2p}+\frac{q-k}{p(q-2)}}.
\end{aligned}$$

By convexity, we may bound all terms in the maximum by their value at either $k = 2 + \frac{q-2}{q} \le 3$ or $k = 2 + 2p\frac{q-2}{q} \ge 2p$, so that

$$\left|\frac{d}{dt}\mathbf{E}[\mathrm{tr}\, X(t)^{2p}]\right| \le \frac{(8p)^{2+\frac{q-2}{q}} R_q(X)}{8} \max\left\{\mathbf{E}[\mathrm{tr}\, X(t)^{2p}]^{1-\frac{1}{2p}}, \left((8p)^{\frac{q-2}{q}} R_q(X)\right)^{2p-1}\right\}.$$

The conclusion follows from Lemma 6.6. □

Remark 6.7. That there is considerable room in the proof of Theorem 2.8 is evident from the crude inequality in the first equation display of the proof of Lemma 6.4. This additional room can be used to capture models whose summands Z_i are not uniformly bounded, but have subexponential tails. For the purposes of this paper, such an extension is not needed as the truncation method that will be developed in section 8 below yields far more general results. However, this extra room can be of significant utility in extending the approach of this paper to random matrices that are not captured by the independent sum model (2.1), cf. [55].

6.3. **Resolvent moments.** The aim of this section is to prove the following universality principle for the moments of the resolvent. This result will form the basis for the proof of Theorem 2.5 (which is given in section 7 below).

Theorem 6.8 (Resolvent moments universality)**.** *We have*

$$\left|\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X|^{-2p}]^{\frac{1}{2p}} - \mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-G|^{-2p}]^{\frac{1}{2p}}\right| \lesssim \frac{R(X)\sigma(X)^2p^2 + R(X)^3p^3}{(\mathrm{Im}\, z)^4}$$

for any $p \in \mathbb{N}$ *and* $z \in \mathbb{C}$ *with* $\mathrm{Im}\, z > 0$.

The proof of Theorem 6.8 is very similar in spirit to that of Theorem 2.8. However, as the resolvent is not a polynomial (and does not have a globally convergent power series), we must truncate the cumulant expansion as in Corollary 6.2.

We begin by computing the derivatives of $M \mapsto \mathrm{tr}\,|z\mathbf{1}-M|^{-2p}$.

Lemma 6.9. *Let $z \in \mathbb{C}$ with $\mathrm{Im}\, z > 0$, $p \in \mathbb{N}$, and $M, B_1, \ldots, B_k \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$. Denote the resolvent of M as $R_M(z) := (z\mathbf{1} - M)^{-1}$. Then*

$$\partial_{B_1} \cdots \partial_{B_k} \operatorname{tr} |z\mathbf{1} - M|^{-2p} = \sum_{\sigma \in \mathrm{Sym}(k)} \sum_{\substack{l,m \geq 0 \\ l+m=k}} \sum_{\substack{r_1,\ldots,r_{l+1} \geq 1 \\ r_1+\cdots+r_{l+1}=p+l}} \sum_{\substack{s_1,\ldots,s_{m+1} \geq 1 \\ s_1+\cdots+s_{m+1}=p+m}} \operatorname{tr}[R_M(z)^{r_1} B_{\sigma(1)} \cdots R_M(z)^{r_l} B_{\sigma(l)} R_M(z)^{r_{l+1}} R_M(\bar{z})^{s_1} B_{\sigma(l+1)} \cdots R_M(\bar{z})^{s_m} B_{\sigma(k)} R_M(\bar{z})^{s_{m+1}}].$$

In particular,

$$\left|\partial_{B_1} \cdots \partial_{B_k} \operatorname{tr} |z\mathbf{1} - M|^{-2p}\right| \leq \frac{(2p-1+k)!}{(2p-1)!} \frac{\|B_1\|_k \cdots \|B_k\|_k}{(\mathrm{Im}\, z)^{2p+k}}.$$

Proof. The identity follows by applying the product rule k times to $\operatorname{tr}|z\mathbf{1}-M|^{-2p} = \operatorname{tr}[R_M(z)^p R_M(\bar{z})^p]$ and using that $\partial_B R_M(z) = R_M(z) B R_M(z)$. To prove the inequality, note that each summand is bounded by $(\mathrm{Im}\, z)^{-2p-k} \|B_1\|_k \cdots \|B_k\|_k$ by Hölder's inequality and $\|R_M(z)\| \leq |\mathrm{Im}\, z|^{-1}$, while the sums have $\frac{(2p-1+k)!}{(2p-1)!}$ terms (the latter is most easily seen by applying the first identity with $d = 1$.) ∎

We can now apply the cumulant expansion.

Proposition 6.10. *For any $z \in \mathbb{C}$ with $\mathrm{Im}\, z > 0$, $p \in \mathbb{N}$, and $t \in [0,1]$, we have*

$$\left|\frac{d}{dt}\mathbf{E}[\operatorname{tr}|z\mathbf{1} - X(t)|^{-2p}]\right| \lesssim \frac{p^3 R(X)\sigma(X)^2}{(\mathrm{Im}\, z)^4} \max\left\{\mathbf{E}[\operatorname{tr}|z\mathbf{1} - X(t)|^{-2p}]^{1-\frac{1}{2p}}, \frac{(32pR(X))^{6p-3}}{(\mathrm{Im}\, z)^{8p-4}}\right\}.$$

Proof. Combining Corollary 6.2 with Lemma 6.9 yields

$$\frac{d}{dt}\mathbf{E}[\operatorname{tr}|z\mathbf{1} - X(t)|^{-2p}] = \frac{1}{2}\sum_{k=3}^{6p-1} k t^{\frac{k}{2}-1} \sum_{\pi \in \mathrm{P}([k])} (-1)^{|\pi|-1}(|\pi|-1)! \times$$
$$\sum_{\substack{l,m \geq 0 \\ l+m=k}} \sum_{\substack{r_1,\ldots,r_{l+1} \geq 1 \\ r_1+\cdots+r_{l+1}=p+l}} \sum_{\substack{s_1,\ldots,s_{m+1} \geq 1 \\ s_1+\cdots+s_{m+1}=p+m}} \sum_{i=1}^n \mathbf{E}[\operatorname{tr} R_{X(t)}(z)^{r_1} Z_{i1|\pi} \cdots R_{X(t)}(z)^{r_l} Z_{il|\pi} \cdot R_{X(t)}(z)^{r_{l+1}} R_{X(t)}(\bar{z})^{s_1} Z_{i(l+1)|\pi} \cdots R_{X(t)}(\bar{z})^{s_m} Z_{ik|\pi} R_{X(t)}(\bar{z})^{s_{m+1}}] + \mathcal{R}$$

with

$$|\mathcal{R}| \lesssim \frac{(8p-1)!}{(2p-1)!} \frac{2^{6p}}{(\mathrm{Im}\, z)^{8p}} \sum_{i=1}^n \mathbf{E}[\operatorname{tr} Z_i^{6p}].$$

Here we eliminated the sum over permutations σ in the identity as in the proof of Proposition 6.5, and we used Lemma 6.4 and Hölder's inequality in the estimate of the remainder. To proceed, we apply Proposition 5.1 with $p_j = \frac{2p+k}{r_{j+1}}$ for $1 \leq j < l$, $p_l = \frac{2p+k}{r_{l+1}+s_1}$, $p_j = \frac{2p+k}{s_{j-l+1}}$ for $l < j < k$, $p_k = \frac{2p+k}{s_{m+1}+r_1}$, and $q = \infty$ to estimate

$$\left|\sum_{i=1}^n \mathbf{E}[\operatorname{tr} R_{X(t)}(z)^{r_1} Z_{i1|\pi} \cdots R_{X(t)}(z)^{r_l} Z_{il|\pi} R_{X(t)}(z)^{r_{l+1}} R_{X(t)}(\bar{z})^{s_1} Z_{i(l+1)|\pi} \cdots R_{X(t)}(\bar{z})^{s_m} Z_{ik|\pi} R_{X(t)}(\bar{z})^{s_{m+1}}]\right| \leq R(X)^{k-2}\sigma(X)^2 \mathbf{E}[\operatorname{tr}|z\mathbf{1} - X(t)|^{-2p-k}].$$

We can therefore estimate

$$\left|\frac{d}{dt}\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p}]\right| \le C\frac{(8p-1)!}{(2p-1)!}\frac{2^{6p}}{(\mathrm{Im}\,z)^{8p}}\sum_{i=1}^{n}\mathbf{E}[\mathrm{tr}\,Z_i^{6p}] + \frac{1}{2}\sum_{k=3}^{6p-1}\frac{(2p-1+k)!}{(2p-1)!}2^k R(X)^{k-2}\sigma(X)^2\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p-k}]$$

for a universal constant C, where we used Lemma 6.4 and that the sums over l, m, r_j, s_j contain a total of $\binom{2p-1+k}{2p-1}$ terms (cf. the proof of Lemma 6.9). Thus

$$\left|\frac{d}{dt}\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p}]\right| \le C\frac{(16p)^{6p}}{(\mathrm{Im}\,z)^{8p}}R(X)^{6p-2}\sigma(X)^2 + \frac{1}{2}\sum_{k=3}^{6p-1}(16p)^k R(X)^{k-2}\sigma(X)^2\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p-k}],$$

where we used $\frac{(2p-1+k)!}{(2p-1)!} \le (8p)^k$ for $k \le 6p$ and $\sum_{i=1}^n \mathbf{E}[\mathrm{tr}\,Z_i^{6p}] \le R(X)^{6p-2}\sigma(X)^2$.

We now proceed as in the proof of Proposition 6.5. As the terms inside the sum are convex as a function of k, we can estimate

$$\begin{aligned}&\frac{1}{2}\sum_{k=3}^{6p-1}(16p)^k R(X)^{k-2}\sigma(X)^2\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p-k}]\\ &\le \frac{1}{8}\max\big\{(32p)^3R(X)\sigma(X)^2\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p-3}],\\ &\qquad (32p)^{6p}R(X)^{6p-2}\sigma(X)^2\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-8p}]\big\}\\ &\le \frac{(32p)^3R(X)\sigma(X)^2}{8(\mathrm{Im}\,z)^4}\max\left\{\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p}]^{1-\frac{1}{2p}}, \frac{(32pR(X))^{6p-3}}{(\mathrm{Im}\,z)^{8p-4}}\right\},\end{aligned}$$

where we used that $\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p+1}] \le \mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-X(t)|^{-2p}]^{1-\frac{1}{2p}}$ by Jensen's inequality, and that $\|(z\mathbf{1}-X(t))^{-1}\| \le (\mathrm{Im}\,z)^{-1}$. The conclusion follows readily. ☐

The proof of Theorem 6.8 is now immediate.

Proof of Theorem 6.8. Combine Proposition 6.10 and Lemma 6.6. ☐

6.4. **Resolvent.** The aim of this section is to prove the resolvent universality principle of Theorem 2.9. In contrast to the universality principles for the moments and resolvent moments, the present result is much more classical in nature as its proof does not require the cumulant expansion; it could therefore also be approached by means of more traditional universality methods as in [29]. In particular, we will apply Corollary 6.2 with $p = 3$: in this special case, the proof of Corollary 6.2 uses only Taylor expansion and no cumulants appear.

Nonetheless, the present situation is somewhat different in nature than the previous universality results in that we bound the difference between the expected resolvents of X and G in norm (as opposed to their traces, see Remark 6.13 below). This introduces some additional subtleties that must be addressed in the proof.

We begin by applying Corollary 6.2 in the present setting.

Lemma 6.11. *We have*

$$\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}]-\mathbf{E}[(z\mathbf{1}-G)^{-1}]\big\| \lesssim \bigg(\frac{R(X)}{(\operatorname{Im} z)^4}+\frac{R(X)^3}{(\operatorname{Im} z)^6}\bigg)\mathbf{E}\bigg\|\sum_{i=1}^n Z_i^2\bigg\|.$$

Proof. We readily compute

$$\partial_{B_1}\partial_{B_2}\partial_{B_3}(z\mathbf{1}-M)^{-1} = \sum_{\sigma\in\mathrm{Sym}(3)}(z\mathbf{1}-M)^{-1}B_{\sigma(1)}(z\mathbf{1}-M)^{-1}B_{\sigma(2)}(z\mathbf{1}-M)^{-1}B_{\sigma(3)}(z\mathbf{1}-M)^{-1}.$$

Applying Corollary 6.2 to $f(M)=\langle v,(z\mathbf{1}-M)^{-1}w\rangle$ with $p=3$ yields

$$\begin{aligned}\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}]-\mathbf{E}[(z\mathbf{1}-G)^{-1}]\big\| &= \sup_{\|v\|=\|w\|=1}\bigg|\int_0^1 \frac{d}{dt}\mathbf{E}[\langle v,(z\mathbf{1}-X(t))^{-1}w\rangle]\,dt\bigg| \\ &\lesssim F(Z,Z,Z)+F(Z,Z',Z')+F(Z',Z,Z')+F(Z',Z',Z),\end{aligned}$$

where $Z'=(Z_i')_{1\le i\le n}$ is an independent copy of $Z=(Z_i)_{1\le i\le n}$,

$$F(Z^{(1)},Z^{(2)},Z^{(3)}) = \sup_{\|v\|=\|w\|=1}\sup_{s,t\in[0,1]}\bigg|\sum_{i=1}^n \mathbf{E}[\langle v, G_{tis}Z_i^{(1)}G_{tis}Z_i^{(2)}G_{tis}Z_i^{(3)}G_{tis}w\rangle]\bigg|,$$

and $G_{tis}=(z\mathbf{1}-X(t,i,s))^{-1}$ with $X(t,i,s)=X(t)-(1-s)\sqrt{t}Z_i$. (Note that the term with $|\pi|=2$ in the bound on $|\mathcal{R}|$ in Corollary 6.2 vanishes as $\mathbf{E}[Z_i]=0$.)

As $\|G_{tis}\|\le(\operatorname{Im} z)^{-1}$ and $\|Z_i\|\le R(X)$, we have

$$F(Z^{(1)},Z^{(2)},Z^{(3)}) \le \frac{R(X)}{(\operatorname{Im} z)^2}\sup_{\|v\|=\|w\|=1}\sup_{s,t\in[0,1]}\sum_{i=1}^n \mathbf{E}[\|Z_i^{(1)}G_{tis}^*v\|\|Z_i^{(3)}G_{tis}w\|].$$

Now note that as $A^{-1}-B^{-1}=A^{-1}(B-A)B^{-1}$, we have

$$\begin{aligned}G_{tis} &= G_t-(1-s)\sqrt{t}\,G_{tis}Z_iG_t,\\ G_{tis}^* &= G_t^*-(1-s)\sqrt{t}\,G_{tis}Z_iG_t^*\end{aligned}$$

with $G_t=(z\mathbf{1}-X(t))^{-1}$. Thus

$$\begin{aligned}\|Z_i^{(3)}G_{tis}w\| &\le \|Z_i^{(3)}G_tw\|+\frac{R(X)}{\operatorname{Im} z}\|Z_iG_tw\|,\\ \|Z_i^{(1)}G_{tis}^*v\| &\le \|Z_i^{(1)}G_t^*v\|+\frac{R(X)}{\operatorname{Im} z}\|Z_iG_t^*v\|\end{aligned}$$

for $s,t\in[0,1]$. Using $(a+b)(c+d)\le a^2+b^2+c^2+d^2$ yields

$$\begin{aligned}\mathbf{E}[\|Z_i^{(1)}G_{tis}^*v\|\|Z_i^{(3)}G_{tis}w\|] \le{}& \mathbf{E}[\langle G_tw,(Z_i^{(3)})^2G_tw\rangle]+\mathbf{E}[\langle G_t^*v,(Z_i^{(1)})^2G_t^*v\rangle]+\\ &\frac{R(X)^2}{(\operatorname{Im} z)^2}\big(\mathbf{E}[\langle G_tw,Z_i^2G_tw\rangle]+\mathbf{E}[\langle G_t^*v,Z_i^2G_t^*v\rangle]\big).\end{aligned}$$

We can therefore estimate

$$F(Z^{(1)},Z^{(2)},Z^{(3)}) \lesssim \bigg(\frac{R(X)}{(\operatorname{Im} z)^4}+\frac{R(X)^3}{(\operatorname{Im} z)^6}\bigg)\mathbf{E}\bigg\|\sum_{i=1}^n Z_i^2\bigg\|$$

whenever $Z^{(k)}$ are equidistributed with Z, concluding the proof. □

We also need the following simple lemma.

Lemma 6.12. *We have*

$$\mathbf{E}\left\|\sum_{i=1}^{n} Z_i^2\right\| \lesssim \sigma(X)^2 + R(X)^2 \log d.$$

Proof. We can estimate

$$\begin{aligned}\mathbf{E}\left\|\sum_{i=1}^{n} Z_i^2\right\| &\le \sigma(X)^2 + \mathbf{E}\left\|\sum_{i=1}^{n}(Z_i^2 - \mathbf{E}Z_i^2)\right\| \\ &\lesssim \sigma(X)^2 + \left\|\sum_{i=1}^{n}\mathbf{E}[(Z_i^2 - \mathbf{E}Z_i^2)^2]\right\|^{\frac{1}{2}}\sqrt{\log d} + R(X)^2\log d \\ &\lesssim \sigma(X)^2 + R(X)\sigma(X)\sqrt{\log d} + R(X)^2 \log d,\end{aligned}$$

where the second line follows from the matrix Bernstein inequality (1.3). The conclusion follows as $R(X)\sigma(X)\sqrt{\log d} \le \sigma(X)^2 + R(X)^2 \log d$. □

We can now complete the proof of Theorem 2.9.

Proof of Theorem 2.9. Lemmas 6.11 and 6.12 yield

$$\big\|\mathbf{E}[(z\mathbf{1} - X)^{-1}] - \mathbf{E}[(z\mathbf{1} - G)^{-1}]\big\| \lesssim \frac{\sigma(X)^2 + R(X)^2 \log d}{(\operatorname{Im} z)^3}\left(\frac{R(X)}{\operatorname{Im} z} + \frac{R(X)^3}{(\operatorname{Im} z)^3}\right).$$

This yields the first inequality of Theorem 2.9 when $\operatorname{Im} z \ge R(X)$. On the other hand, when $\operatorname{Im} z < R(X)$, we can crudely estimate

$$\begin{aligned}\big\|\mathbf{E}[(z\mathbf{1} - X)^{-1}] - \mathbf{E}[(z\mathbf{1} - G)^{-1}]\big\| &\le \big\|\mathbf{E}[(z\mathbf{1} - X)^{-1}]\big\| + \big\|\mathbf{E}[(z\mathbf{1} - G)^{-1}]\big\| \\ &\le \frac{2}{\operatorname{Im} z} \le \frac{2R(X)^3}{(\operatorname{Im} z)^4},\end{aligned}$$

concluding the proof of the first inequality.

To prove the second inequality, it suffices to consider real-valued $\varphi : \mathbb{R} \to \mathbb{R}$ (as otherwise we may apply the real-valued inequality separately to the real and imaginary parts of φ). Then $\varphi(X)$ and $\varphi(G)$ are self-adjoint, so we may express

$$\begin{aligned}\big\|\mathbf{E}[\varphi(X)] - \mathbf{E}[\varphi(G)]\big\| &= \sup_{\|v\|=1}\big|\langle v, \mathbf{E}[\varphi(X)]v\rangle - \langle v, \mathbf{E}[\varphi(G)]v\rangle\big| \\ &= \sup_{\|v\|=1}\left|\int \varphi\, d\mu_v - \int \varphi\, d\nu_v\right|\end{aligned}$$

where we defined the probability measures μ_v, ν_v for $\|v\| = 1$ so that $\int \varphi\, d\mu_v = \langle v, \mathbf{E}[\varphi(X)]v\rangle$ and $\int \varphi\, d\nu_v = \langle v, \mathbf{E}[\varphi(G)]v\rangle$ for all bounded continuous φ. As the first inequality of Theorem 2.9 implies that

$$\sup_{\|v\|=1}\left|\int \frac{1}{z-x}\,\mu_v(dx) - \int \frac{1}{z-x}\,\nu_v(dx)\right| \lesssim \frac{R(X)\sigma(X)^2 + R(X)^3 \log d}{(\operatorname{Im} z)^4},$$

the second inequality of Theorem 2.9 follows from [9, Lemma 5.11]. □

Remark 6.13 (An improved inequality for Stieltjes transforms). The main complication in the proof of Theorem 2.9 arises from the fact that we aim to achieve a norm estimate. If we are only interested in establishing universality of the trace of the resolvent (that is, of the Stieltjes transform of the empirical spectral distribution), the proof simplifies greatly and yields a better bound. Indeed, the first equation display of the proof of Lemma 6.11 and Hölder's inequality readily yield

$$|\mathbf{E}[\partial_{Y_1}\partial_{Y_2}\partial_{Y_3}\,\mathrm{tr}\,(z\mathbf{1}-X(t,i,s))^{-1}]| \le \frac{6\|Y_1\|_3\|Y_2\|_3\|Y_3\|_3}{(\mathrm{Im}\,z)^4}$$

for any random matrices Y_1, Y_2, Y_3. Combining this inequality with the $p = 3$ case of Corollary 6.2 immediately yields the estimate

$$|\mathbf{E}[\mathrm{tr}(z\mathbf{1}-X)^{-1}] - \mathbf{E}[\mathrm{tr}(z\mathbf{1}-G)^{-1}]| \lesssim \frac{1}{(\mathrm{Im}\,z)^4}\sum_{i=1}^n \mathbf{E}[\mathrm{tr}\,|Z_i|^3].$$

In particular, in this case the logarithmic dimension dependence is eliminated.

7. Universality of the spectrum

The aim of this section is to prove Theorem 2.5 and Corollary 2.6. The main idea behind the proof is that universality of the spectrum can be deduced from the bound on the moments of the resolvent in Theorem 6.8 using a technique that was developed for Gaussian random matrices in [9, §6.2]. The difficulty in the present setting is that the resolvent of the non-Gaussian random matrix X exhibits more complicated concentration properties than in the Gaussian case.

We first introduce some basic estimates in section 7.1. In sections 7.2 and 7.3, we bound the probability that $\mathrm{sp}(X) \subseteq \mathrm{sp}(G) + [-\varepsilon, \varepsilon]$ and $\mathrm{sp}(G) \subseteq \mathrm{sp}(X) + [-\varepsilon, \varepsilon]$, respectively. Combining these bounds yields the Hausdorff distance bound of Theorem 2.5. Finally, Corollary 2.6 will be proved in section 7.4.

7.1. Preliminaries.

The basic principle behind te proof is the following deterministic lemma, which is a trivial modification of [9, Lemma 6.4].

Lemma 7.1. *Let $C, K_1, K_2, K_3 \ge 0$, and let $A, B \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ satisfy*

$$\|(z\mathbf{1}-A)^{-1}\| \le C\|(z\mathbf{1}-B)^{-1}\| + \frac{K_1}{(\mathrm{Im}\,z)^2} + \frac{K_2}{(\mathrm{Im}\,z)^3} + \frac{K_3}{(\mathrm{Im}\,z)^4}$$

for all $z = \lambda + i\varepsilon$ with $\lambda \in \mathrm{sp}(A)$ and $\varepsilon = 6K_1 \vee (6K_2)^{\frac{1}{2}} \vee (6K_3)^{\frac{1}{3}}$. Then

$$\mathrm{sp}(A) \subseteq \mathrm{sp}(B) + 2C\varepsilon[-1, 1].$$

Proof. Fix $\lambda \in \mathrm{sp}(A)$ and $z = \lambda + i\varepsilon$, where ε is as defined in the statement. As $\|(z\mathbf{1}-A)^{-1}\| = (\mathrm{dist}(z, \mathrm{sp}(A)))^{-1}$, the assumption implies that

$$\frac{1}{\varepsilon} \le \frac{C}{\sqrt{\varepsilon^2 + \mathrm{dist}(\lambda, \mathrm{sp}(B))^2}} + \frac{K_1}{\varepsilon^2} + \frac{K_2}{\varepsilon^3} + \frac{K_3}{\varepsilon^4}.$$

If $\mathrm{dist}(\lambda, \mathrm{sp}(B)) > 2C\varepsilon$, we would have $\frac{1}{2} < \frac{K_1}{\varepsilon} + \frac{K_2}{\varepsilon^2} + \frac{K_3}{\varepsilon^3} \le \frac{1}{2}$ by the definition of ε, which is impossible. Thus $\mathrm{dist}(\lambda, \mathrm{sp}(B)) \le 2C\varepsilon$ for all $\lambda \in \mathrm{sp}(A)$. ☐

The main idea behind the proof of Theorem 2.5 is that we will engineer the assumption of Lemma 7.1 by using Theorem 6.8 and concentration of measure. Before we turn to the details of the argument, let us prove a crude *a priori* bound on the spectrum that will be needed below.

Lemma 7.2. *We have*

$$\mathbf{P}\big[\mathrm{sp}(X) \subseteq \mathrm{sp}(\mathbf{E}X) + C\{\sigma_*(X)\sqrt{d+t} + R(X)(d+t)\}[-1,1]\big] \ge 1 - e^{-t}$$

and

$$\mathbf{P}\big[\mathrm{sp}(G) \subseteq \mathrm{sp}(\mathbf{E}G) + C\sigma_*(X)\sqrt{d+t}\,[-1,1]\big] \ge 1 - e^{-t}$$

for all $t \ge 0$, where C is a universal constant.

Proof. Let $N \subset \mathbb{S}^{d-1}$ be a $\frac{1}{4}$-net of the unit sphere $\mathbb{S}^{d-1} := \{x \in \mathbb{C}^d : \|x\| = 1\}$, that is, $\mathrm{dist}(x, N) \le \frac{1}{4}$ for all $x \in \mathbb{S}^{d-1}$. A routine estimate [65, p. 110] yields

$$\begin{aligned}\mathbf{P}[\|X - \mathbf{E}X\| \ge x] &\le \mathbf{P}\bigg[\max_{v,w\in N} |\langle v, (X - \mathbf{E}X)w\rangle| \ge \frac{x}{4}\bigg] \\ &\le |N|^2 \sup_{v,w\in\mathbb{S}^{d-1}} \mathbf{P}\bigg[|\langle v, (X - \mathbf{E}X)w\rangle| \ge \frac{x}{4}\bigg].\end{aligned}$$

By viewing $\mathbb{C}^d$ as a $2d$-dimensional real vector space, we may use a standard volume argument [65, Lemma 2.3.4] to choose the net N so that $|N| \le C^d$ for a universal constant C. On the other hand, by Bernstein's inequality [21, Theorem 2.10]

$$\begin{aligned}&\mathbf{P}\big[|\langle v, (X - \mathbf{E}X)w\rangle| \ge 2\sigma_*(X)\sqrt{x} + \sqrt{2}R(X)x\big] \\ &\quad\le \mathbf{P}\Bigg[\bigg|\sum_{i=1}^n \mathrm{Re}\,\langle v, Z_i w\rangle\bigg| \ge \sigma_*(X)\sqrt{2x} + R(X)x\Bigg] \\ &\qquad + \mathbf{P}\Bigg[\bigg|\sum_{i=1}^n \mathrm{Im}\,\langle v, Z_i w\rangle\bigg| \ge \sigma_*(X)\sqrt{2x} + R(X)x\Bigg] \le 4e^{-x}\end{aligned}$$

for all $x \ge 0$ and $v, w \in \mathbb{S}^{d-1}$. Combining the above estimates yields

$$\mathbf{P}[\|X - \mathbf{E}X\| \ge 8\sigma_*(X)\sqrt{cd+t} + 4\sqrt{2}R(X)(cd+t)] \le C^{2d}e^{-cd-t} \le e^{-t}$$

for all $t \ge 0$, provided the universal constant c is chosen sufficiently large. The first inequality in the statement now follows by noting that

$$\mathrm{sp}(X) \subseteq \mathrm{sp}(\mathbf{E}X) + \|X - \mathbf{E}X\|[-1,1]$$

by Weyl's inequality $\max_i |\lambda_i(A) - \lambda_i(B)| \le \|A - B\|$ for self-adjoint matrices A, B [17, Corollary III.2.6] (here $\lambda_i(A)$ is the ith largest eigenvalue of A).

The inequality for the Gaussian matrix G follows in the identical fashion, except that we replace Bernstein's inequality by the Gaussian bound

$$\mathbf{P}\big[|\langle v, (G - \mathbf{E}G)w\rangle| \ge 2\sigma_*(X)\sqrt{x}\big] \le 4e^{-x}$$

(this follows from the Gaussian tail bound [21, p. 22] as the real and imaginary parts of $\langle v, (G - \mathbf{E}G)w\rangle$ are Gaussian variables with variance bounded by $\sigma_*(X)^2$). □

7.2. **Proof of Theorem 2.5: upper bound.** The aim of the present section is to prove that $\mathrm{sp}(X) \subseteq \mathrm{sp}(G) + [-\varepsilon, \varepsilon]$ with high probability for a suitable choice of ε. This will be accomplished by showing that the corresponding resolvent norm inequality of Lemma 7.1 holds with with high probability.

We begin by showing that this is the case for a single choice of z. Note that the joint distribution of X and G is irrelevant to the following proofs.

Lemma 7.3. *Let $z \in \mathbb{C}$ with $\mathrm{Im}\, z > 0$. Then*

$$\mathbf{P}\bigg[\|(z\mathbf{1}-X)^{-1}\| \ge C\bigg\{\|(z\mathbf{1}-G)^{-1}\| + \frac{\sigma_*(X)}{(\mathrm{Im}\, z)^2}\sqrt{x} + \frac{R(X)\sigma(X)^2x^2 + R(X)^3x^3}{(\mathrm{Im}\, z)^4}\bigg\}\bigg] \le 3e^{-x}$$

for all $x \ge \log d$, where C is a universal constant.

Proof. We begin by noting that Markov's inequality implies

$$\mathbf{P}\Big[\|(z\mathbf{1}-X)^{-1}\| \ge e\,\mathbf{E}[\|(z\mathbf{1}-X)^{-1}\|^{2p}]^{\frac{1}{2p}}\Big] \le e^{-2p}.$$

By Theorem 6.8, the expectation inside the probability satisfies

$$\mathbf{E}[\|(z\mathbf{1}-X)^{-1}\|^{2p}]^{\frac{1}{2p}} \le d^{\frac{1}{2p}}\mathbf{E}[\|(z\mathbf{1}-G)^{-1}\|^{2p}]^{\frac{1}{2p}} + Cd^{\frac{1}{2p}}\frac{R(X)\sigma(X)^2p^2 + R(X)^3p^3}{(\mathrm{Im}\, z)^4}$$

for $p \in \mathbb{N}$, where C is a universal constant. Here we used that $\frac{1}{d}\|A\| \le \mathrm{tr}\, A \le \|A\|$ for any positive semidefinite matrix $A \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$.

To proceed, note first that Lemma 5.5 implies

$$\mathbf{E}[\|(z\mathbf{1}-G)^{-1}\|^{2p}]^{\frac{1}{2p}} \le \mathbf{E}\|(z\mathbf{1}-G)^{-1}\| + C\sqrt{p}\frac{\sigma_*(X)}{(\mathrm{Im}\, z)^2}$$

for $p \in \mathbb{N}$, where C is a universal constant (this follows as the L^p-norm of a σ^2-subgaussian random variable is at most of order $\sigma\sqrt{p}$, cf. [21, Theorem 2.1]). Another application of Lemma 5.5 therefore yields

$$\mathbf{P}\bigg[\mathbf{E}[\|(z\mathbf{1}-G)^{-1}\|^{2p}]^{\frac{1}{2p}} \ge \|(z\mathbf{1}-G)^{-1}\| + C_1\sqrt{p}\frac{\sigma_*(X)}{(\mathrm{Im}\, z)^2}\bigg] \le 2e^{-C_2p}$$

for $p \in \mathbb{N}$ and universal constants C_1, C_2.

Combining the above bounds yields

$$\mathbf{P}\bigg[\|(z\mathbf{1}-X)^{-1}\| \ge ed^{\frac{1}{2p}}\|(z\mathbf{1}-G)^{-1}\| + C_1ed^{\frac{1}{2p}}\sqrt{p}\frac{\sigma_*(X)}{(\mathrm{Im}\, z)^2} + Ced^{\frac{1}{2p}}\frac{R(X)\sigma(X)^2p^2 + R(X)^3p^3}{(\mathrm{Im}\, z)^4}\bigg] \le e^{-2p} + 2e^{-C_2p}$$

for $p \in \mathbb{N}$. The conclusion follows readily using $d^{\frac{1}{2p}} \le e^{\frac{1}{2}}$ for $p \ge \log d$. ☐

We are now ready to prove one direction of Theorem 2.5.

Proposition 7.4. *For any $t \ge 0$, we have*

$$\mathbf{P}\big[\mathrm{sp}(X) \subseteq \mathrm{sp}(G) + C\varepsilon(t)[-1,1]\big] \ge 1 - de^{-t},$$

where C is a universal constant and $\varepsilon(t)$ is as defined in Theorem 2.5.

Proof. Define the set

$$\Omega_x := \mathrm{sp}(\mathbf{E}X) + C'\{\sigma_*(X)\sqrt{d+x} + R(X)(d+x)\}[-1,1],$$

where C' is the universal constant of Lemma 7.2. Then Ω_x is a union of d intervals of length $2C'\{\sigma_*(X)\sqrt{d+x} + R(X)(d+x)\}$. We can therefore find $\mathcal{N}_x \subset \Omega_x$ with $|\mathcal{N}_x| \le \frac{4C'd(d+x)}{x}$ such that each $\lambda \in \Omega_x$ satisfies $\mathrm{dist}(\lambda, \mathcal{N}_x) \le \sigma_*\sqrt{x} + R(X)x$. In particular, for every $\lambda \in \Omega_x$, there exists $\lambda' \in \mathcal{N}_x$ so that

$$\big|\|((\lambda+i\varepsilon)\mathbf{1}-X)^{-1}\| - \|((\lambda'+i\varepsilon)\mathbf{1}-X)^{-1}\|\big| \le \frac{\sigma_*\sqrt{x} + R(X)x}{\varepsilon^2}$$

as well as the analogous bound where X is replaced by G (here we used the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$). We can therefore estimate

$$\begin{aligned}
&\mathbf{P}\bigg[\|(z\mathbf{1} - X)^{-1}\| \ge C\bigg\{\|(z\mathbf{1} - G)^{-1}\| + \frac{3\sigma_*(X)\sqrt{x} + 2R(X)x}{\varepsilon^2} \\
&\qquad + \frac{R(X)\sigma(X)^2x^2 + R(X)^3x^3}{\varepsilon^4}\bigg\} \text{ for some } z \in \mathrm{sp}(X) + i\varepsilon\bigg] \\
&\le \mathbf{P}\bigg[\|(z\mathbf{1} - X)^{-1}\| \ge C\bigg\{\|(z\mathbf{1} - G)^{-1}\| + \frac{3\sigma_*(X)\sqrt{x} + 2R(X)x}{\varepsilon^2} \\
&\qquad + \frac{R(X)\sigma(X)^2x^2 + R(X)^3x^3}{\varepsilon^4}\bigg\} \text{ for some } z \in \Omega_x + i\varepsilon\bigg] + e^{-x} \\
&\le \mathbf{P}\bigg[\|(z\mathbf{1} - X)^{-1}\| \ge C\bigg\{\|(z\mathbf{1} - G)^{-1}\| + \frac{\sigma_*(X)}{\varepsilon^2}\sqrt{x} \\
&\qquad + \frac{R(X)\sigma(X)^2x^2 + R(X)^3x^3}{\varepsilon^4}\bigg\} \text{ for some } z \in \mathcal{N}_x + i\varepsilon\bigg] + e^{-x} \\
&\le (3|\mathcal{N}_x| + 1)e^{-x} \le \bigg(1 + \frac{12C'd(d + x)}{x}\bigg)e^{-x}
\end{aligned}$$

for $x \ge \log d$, where we used Lemma 7.2 in the first inequality and a union bound and Lemma 7.3 in the third inequality (here $C > 1$ is the constant of Lemma 7.3).

Now let $x = Lt$ for a universal constant L. Recalling the standing assumption $d \ge 2$, it is readily seen that we may choose $L > 1$ sufficiently large so that

$$\bigg(1 + \frac{12C'd(d + x)}{x}\bigg)e^{-x} \le de^{-t}$$

for all $t \ge \log d$. Then we have shown that

$$\begin{aligned}
&\mathbf{P}\bigg[\|(z\mathbf{1} - X)^{-1}\| \le 3L^3C\bigg\{\|(z\mathbf{1} - G)^{-1}\| + \frac{\sigma_*(X)\sqrt{t} + R(X)t}{\varepsilon^2} \\
&\qquad + \frac{R(X)\sigma(X)^2t^2 + R(X)^3t^3}{\varepsilon^4}\bigg\} \text{ for all } z \in \mathrm{sp}(X) + i\varepsilon\bigg] \ge 1 - de^{-t}
\end{aligned}$$

for all $t \ge \log d$. On the other hand, the same bound holds trivially for $t < \log d$ as then $1 - de^{-t} < 0$. The proof is concluded by applying Lemma 7.1. □

7.3. **Proof of Theorem 2.5: lower bound.** We now turn to the complementary inequality $\mathrm{sp}(G) \subseteq \mathrm{sp}(X) + [-\varepsilon, \varepsilon]$ with high probability. The proof is similar in spirit to that of the upper bound, but we must now work with the more complicated concentration inequality of Proposition 5.6. As before, we begin by establishing a resolvent norm inequality for a single choice of z.

Lemma 7.5. *Let $z \in \mathbb{C}$ with $\mathrm{Im}\, z > 0$. Then*

$$\begin{aligned}
&\mathbf{P}\bigg[\|(z\mathbf{1} - G)^{-1}\| \ge C\bigg\{\|(z\mathbf{1} - X)^{-1}\| + \frac{R(X)\sigma(X)x + R(X)^2x^{\frac{3}{2}}}{(\mathrm{Im}\, z)^3} + \\
&\frac{R(X)\sigma(X)^2x^2 + R(X)^3x^3}{(\mathrm{Im}\, z)^4} + \frac{\sigma_*(X)x^{\frac{1}{2}} + R(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}x^{\frac{3}{4}} + R(X)x}{(\mathrm{Im}\, z)^2}\bigg\}\bigg] \le 3e^{-x}
\end{aligned}$$

for all $x \ge \log d$, where C is a universal constant.

Proof. As in the proof of Lemma 7.3, we have

$$\mathbf{P}\Big[\|(z\mathbf{1}-G)^{-1}\| \ge e\,\mathbf{E}[\|(z\mathbf{1}-G)^{-1}\|^{2p}]^{\frac{1}{2p}}\Big] \le e^{-2p}$$

and

$$\mathbf{E}[\|(z\mathbf{1}-G)^{-1}\|^{2p}]^{\frac{1}{2p}} \le d^{\frac{1}{2p}}\mathbf{E}[\|(z\mathbf{1}-X)^{-1}\|^{2p}]^{\frac{1}{2p}} + Cd^{\frac{1}{2p}}\frac{R(X)\sigma(X)^2p^2+R(X)^3p^3}{(\mathrm{Im}\,z)^4}$$

for $p \in \mathbb{N}$ by Markov's inequality and Theorem 6.8.

To proceed, we use that Proposition 5.6 implies

$$\mathbf{E}[\|(z\mathbf{1}-X)^{-1}\|^{2p}]^{\frac{1}{2p}} \le \mathbf{E}\|(z\mathbf{1}-X)^{-1}\| + C\bigg\{\frac{R(X)}{(\mathrm{Im}\,z)^2}+\frac{R(X)^2}{(\mathrm{Im}\,z)^3}\bigg\}p$$
$$+\,C\bigg\{\frac{\sigma_*(X)+R(X)^{\frac12}(\mathbf{E}\|X-\mathbf{E}X\|)^{\frac12}}{(\mathrm{Im}\,z)^2}+\frac{R(X)(\mathbf{E}\|X-\mathbf{E}X\|^2)^{\frac12}}{(\mathrm{Im}\,z)^3}\bigg\}\sqrt{p}$$

for $p \in \mathbb{N}$ by [21, Theorem 2.3]. Another application of Proposition 5.6 yields

$$\mathbf{P}\bigg[\mathbf{E}[\|(z\mathbf{1}-X)^{-1}\|^{2p}]^{\frac{1}{2p}} \ge \|(z\mathbf{1}-X)^{-1}\| + C\bigg\{\frac{R(X)}{(\mathrm{Im}\,z)^2}+\frac{R(X)^2}{(\mathrm{Im}\,z)^3}\bigg\}p\,+$$
$$C\bigg\{\frac{\sigma_*(X)+R(X)^{\frac12}(\mathbf{E}\|X-\mathbf{E}X\|)^{\frac12}}{(\mathrm{Im}\,z)^2}+\frac{R(X)(\mathbf{E}\|X-\mathbf{E}X\|^2)^{\frac12}}{(\mathrm{Im}\,z)^3}\bigg\}\sqrt{p}\bigg] \le 2e^{-p}$$

for $p \in \mathbb{N}$, provided the universal constant C is chosen sufficiently large. Now recall that the matrix Bernstein inequality [67, eq. (6.1.4)] implies

$$(\mathbf{E}\|X-\mathbf{E}X\|^2)^{\frac12} \lesssim \sigma(X)\sqrt{\log d} + R(X)\log d \le \sigma(X)\sqrt{p} + R(X)p$$

for $p \ge \log d$. We can therefore further estimate

$$\mathbf{P}\bigg[\mathbf{E}[\|(z\mathbf{1}-X)^{-1}\|^{2p}]^{\frac{1}{2p}} \ge \|(z\mathbf{1}-X)^{-1}\| + C\bigg\{\frac{R(X)\sigma(X)p+R(X)^2p^{\frac32}}{(\mathrm{Im}\,z)^3}\,+$$
$$\frac{\sigma_*(X)\sqrt{p}+R(X)^{\frac12}\sigma(X)^{\frac12}p^{\frac34}+R(X)p}{(\mathrm{Im}\,z)^2}\bigg\}\bigg] \le 2e^{-p}$$

for $p \ge \log d$, provided C is chosen sufficiently large. The proof is now readily concluded by combining the above bounds and using $d^{\frac{1}{2p}} \le e^{\frac12}$ for $p \ge \log d$. □

Remark 7.6. We have emphasized in the introduction that the matrix Bernstein inequality may be viewed as a consequence of the universality principles of this paper. On the other hand, we have used the matrix Bernstein inequality in the proof of Lemma 7.5 to estimate the matrix norms that appear in Proposition 5.6. There is no circular reasoning here: the present section is only concerned with lower bounds on the spectrum of X, while the matrix Bernstein inequality already follows from the upper bound of Proposition 7.4 (or from Theorem 2.8 by choosing $p \asymp \log d$ and $q = \infty$) and the noncommutative Khintchine inequality.

The same remark applies to the application of the matrix Bernstein inequality in the proof of Theorem 2.9 (cf. Lemma 6.12 in section 6.4).

We are now ready to prove the converse direction of Theorem 2.5.

Proposition 7.7. *For any $t \ge 0$, we have*

$$\mathbf{P}\big[\mathrm{sp}(G) \subseteq \mathrm{sp}(X) + C\varepsilon(t)[-1,1]\big] \ge 1 - de^{-t},$$

where C is a universal constant and $\varepsilon(t)$ is as defined in Theorem 2.5.

Proof. By following exactly the same steps as in the proof of Proposition 7.4, we can deduce using Lemmas 7.2 and Lemma 7.5 the inequality

$$\mathbf{P}\Bigg[\|(z\mathbf{1}-G)^{-1}\| \le C\bigg\{\|(z\mathbf{1}-X)^{-1}\| + \frac{R(X)\sigma(X)t + R(X)^2 t^{\frac{3}{2}}}{\varepsilon^3} + \frac{R(X)\sigma(X)^2t^2 + R(X)^3t^3}{\varepsilon^4} + \frac{\sigma_*(X)t^{\frac{1}{2}} + R(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}t^{\frac{3}{4}} + R(X)t}{\varepsilon^2}\bigg\} \text{ for all } z \in \mathrm{sp}(G) + i\varepsilon\Bigg] \ge 1 - de^{-t}$$

for all $t, \varepsilon \ge 0$, where C is a universal constant. Thus Lemma 7.1 implies

$$\mathbf{P}\big[\mathrm{sp}(G) \subseteq \mathrm{sp}(X) + C\varepsilon'(t)[-1,1]\big] \ge 1 - de^{-t}$$

for all $t \ge 0$ and a universal constant C, where

$$\varepsilon'(t) = \sigma_*(X)t^{\frac{1}{2}} + R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}t^{\frac{2}{3}} + R(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}t^{\frac{3}{4}} + R(X)t.$$

It remains to note that

$$R(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}t^{\frac{3}{4}} \le \frac{3}{4}R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}t^{\frac{2}{3}} + \frac{1}{4}R(X)t$$

by Young's inequality, concluding the proof. □

We now conclude the proof of Theorem 2.5.

Proof of Theorem 2.5. Combining Propositions 7.4 and 7.7 yields

$$\mathbf{P}[\mathrm{d_H}(\mathrm{sp}(X), \mathrm{sp}(G)) > C\varepsilon(s)] \le 2de^{-s}$$

for all $s \ge 0$ by the union bound. Choosing $s = 2t$, we obtain

$$\mathbf{P}[\mathrm{d_H}(\mathrm{sp}(X), \mathrm{sp}(G)) > 2C\varepsilon(t)] \le 2de^{-2t} \le de^{-t}$$

for $t \ge \log d$, as the latter implies $2e^{-t} \le \frac{2}{d} \le 1$ by the standing assumption $d \ge 2$. But for $t < \log d$ the inequality is trivial as then $de^{-t} > 1$. The tail bound follows.

To deduce the expectation bound, we note that

$$\begin{aligned}\mathbf{E}[\mathrm{d_H}(\mathrm{sp}(X), \mathrm{sp}(G))] &\le C\varepsilon(\log d) + \int_{C\varepsilon(\log d)}^{\infty} \mathbf{P}[\mathrm{d_H}(\mathrm{sp}(X), \mathrm{sp}(G)) > x]\,dx \\ &= C\varepsilon(\log d) + C\int_{\log d}^{\infty} \mathbf{P}[\mathrm{d_H}(\mathrm{sp}(X), \mathrm{sp}(G)) > C\varepsilon(t)]\,\frac{d\varepsilon(t)}{dt}\,dt \\ &\le C\varepsilon(\log d) + 2dC\int_{\log d}^{\infty} e^{-t}\,\frac{d\varepsilon(t)}{dt}\,dt \lesssim \varepsilon(\log d)\end{aligned}$$

using $\int_a^\infty e^{-t}t^\beta\,dt \le C_\beta e^{-a}a^\beta$ for $a > \frac{1}{4}$, $\beta \in \mathbb{R}$, where C_β depends only on β. □

7.4. **Proof of Corollary 2.6.** Now that Theorem 2.5 has been established, the proof of Corollary 2.6 follows by routine manipulations.

Proof of Corollary 2.6. We first note that

$$\mathrm{sp}(A) \subseteq \mathrm{sp}(B) + [-\varepsilon, \varepsilon]$$

certainly implies

$$\lambda_{\max}(A) \le \lambda_{\max}(B) + \varepsilon$$

for any $A, B \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ and $\varepsilon > 0$. Thus

$$|\lambda_{\max}(A) - \lambda_{\max}(B)| \le \mathrm{d_H}(\mathrm{sp}(A), \mathrm{sp}(B)),$$

and the first and last bound of Corollary 2.6 follow immediately from Theorem 2.5.

To prove the middle bound, we note that a routine application of Gaussian concentration (see, e.g., [9, Corollary 4.14]) yields

$$\mathbf{P}\big[|\lambda_{\max}(G)-\mathbf{E}\lambda_{\max}(G)|\geq\sigma_*(X)\sqrt{2t}\big]\leq 2e^{-t}$$

for all $t\geq 0$. Combined with the first bound of Corollary 2.6, we obtain

$$\begin{aligned}\mathbf{P}\big[|\lambda_{\max}(X)-\mathbf{E}\lambda_{\max}(G)|>\sigma_*(X)\sqrt{2t}+C\varepsilon(t)\big]\\ \leq 2e^{-t}+\mathbf{P}\big[|\lambda_{\max}(X)-\lambda_{\max}(G)|>C\varepsilon(t)\big]\leq(d+2)e^{-t}\end{aligned}$$

for all $t\geq 0$. The second inequality of Corollary 2.6 follows for a suitable choice of the universal constant (as in the last step of the proof of Theorem 2.5).

The analogous bounds for $\|X\|,\|G\|$ are proved in an identical manner. □

8. Truncation

The aim of this section is to prove Theorems 2.7 and 2.10. The basic idea behind these results is the following truncation argument. Let X be as in (2.1), and let G be the associated Gaussian model. Define the truncated model

$$\tilde{X}:=Z_0+\sum_{i=1}^n 1_{\|Z_i\|\leq R}Z_i.$$

Then $X=\tilde{X}$ on the event $\{\max_i\|Z_i\|\leq R\}$, while $R(\tilde{X})\leq R$. We can therefore obtain universality principles for unbounded X by conditioning on the above event, and applying the results of the previous sections to $\tilde{X}$.

The problem with this approach is that it does not yield a comparison between the spectra of X and G, but rather between the spectra of X and $\tilde{G}$, where $\tilde{G}$ is the Gaussian model associated to $\tilde{X}$. The main difficulty in the implementation of the truncation argument is therefore to compare the spectra of the Gaussian models G and $\tilde{G}$. To this end, we will first prove general comparison principles for the spectra of Gaussian random matrices in section 8.1. In section 8.2, we will upper bound the relevant parameters in the specific case of G and $\tilde{G}$. Finally, we combine these estimates in section 8.3 to complete the proof of Theorems 2.7 and 2.10.

8.1. Gaussian comparison principles.

The aim of this section is to prove general comparison principles for the spectra of Gaussian random matrices. We begin by stating a comparison principle for the resolvent moments.

Lemma 8.1. *Let $H,\tilde{H}$ be self-adjoint Gaussian random matrices. Then we have*

$$\big|\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-H|^{-2p}]^{\frac{1}{2p}}-\mathbf{E}[\mathrm{tr}\,|z\mathbf{1}-\tilde{H}|^{-2p}]^{\frac{1}{2p}}\big|\leq\frac{\|\mathbf{E}H-\mathbf{E}\tilde{H}\|}{(\mathrm{Im}\,z)^2}+2p\,\frac{\Delta(H,\tilde{H})}{(\mathrm{Im}\,z)^3},$$

where

$$\Delta(H,\tilde{H}):=\sup_{\|M\|\leq 1}\big\|\mathbf{E}[(H-\mathbf{E}H)M(H-\mathbf{E}H)]-\mathbf{E}[(\tilde{H}-\mathbf{E}\tilde{H})M(\tilde{H}-\mathbf{E}\tilde{H})]\big\|.$$

Proof. First, note that as $A^{-1}-B^{-1}=A^{-1}(B-A)B^{-1}$, we have

$$\|(z\mathbf{1}-H)^{-1}-(z\mathbf{1}-(H-\mathbf{E}H+\mathbf{E}\tilde{H}))^{-1}\|\leq\frac{\|\mathbf{E}H-\mathbf{E}\tilde{H}\|}{(\mathrm{Im}\,z)^2}.$$

Thus we can assume in the sequel that $\mathbf{E}H=\mathbf{E}\tilde{H}$.

Assume without loss of generality that $H, \tilde{H}$ are independent and $\mathbf{E}H = \mathbf{E}\tilde{H}$, and let $Y, \tilde{Y}$ be independent copies of $H - \mathbf{E}H$ and $\tilde{H} - \mathbf{E}\tilde{H}$, respectively. Define

$$H(t) := \mathbf{E}H + \sqrt{t}\,(H - \mathbf{E}H) + \sqrt{1-t}\,(\tilde{H} - \mathbf{E}\tilde{H}).$$

By the Gaussian interpolation lemma [9, Lemma 4.11]

$$\frac{d}{dt}\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - H(t)|^{-2p}] = \frac{\mathbf{E}[\partial_Y^2 \,\mathrm{tr}\,|z\mathbf{1} - H(t)|^{-2p}] - \mathbf{E}[\partial_{\tilde{Y}}^2 \,\mathrm{tr}\,|z\mathbf{1} - H(t)|^{-2p}]}{2}.$$

By the product rule, we have

$$\begin{aligned}\partial_B^2 \,\mathrm{tr}\,|z\mathbf{1} - H|^{-2p} = 2p\sum_{k=0}^{p} \mathrm{Re}\,\mathrm{tr}[B(z\mathbf{1} - H)^{-k-1}B(z\mathbf{1} - H)^{-p-1+k}(\bar{z}\mathbf{1} - H)^{-p}] \\ + 2p\sum_{k=0}^{p-1} \mathrm{Re}\,\mathrm{tr}[B(z\mathbf{1} - H)^{-p-1}(\bar{z}\mathbf{1} - H)^{-k-1}B(\bar{z}\mathbf{1} - H)^{-p+k}].\end{aligned}$$

We can therefore bound

$$\left|\frac{d}{dt}\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - H(t)|^{-2p}]\right| \le p(2p+1)\,\Delta(H, \tilde{H})\,\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - H(t)|^{-2p-2}]$$

by applying Lemma 5.2 to $F(M, M') := \mathbf{E}[\mathrm{tr}\,YMYM'] - \mathbf{E}[\mathrm{tr}\,\tilde{Y}M\tilde{Y}M']$ as in the proof of Lemma 5.3, and using $\sup_{\|M\|\le 1, \|M'\|_1 \le 1} |F(M, M')| = \Delta(H, \tilde{H})$.

It remains to note that $\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - H(t)|^{-2p-2}] \le (\mathrm{Im}\,z)^{-3}\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - H(t)|^{-2p}]^{1-\frac{1}{2p}}$ and the chain rule readily yield the estimate

$$\left|\frac{d}{dt}\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - H(t)|^{-2p}]^{\frac{1}{2p}}\right| \le (p + \tfrac{1}{2})\,\frac{\Delta(H, \tilde{H})}{(\mathrm{Im}\,z)^3}.$$

The conclusion now follows by integrating over t (using $p + \frac{1}{2} \le 2p$). □

A bound on the Hausdorff distance now follows along familiar lines.

Proposition 8.2. *Let $H, \tilde{H}$ be self-adjoint Gaussian random matrices. Then*

$$\begin{aligned}\mathbf{P}\big[\mathrm{d_H}(\mathrm{sp}(H), \mathrm{sp}(\tilde{H})) > C\{\|\mathbf{E}H - \mathbf{E}\tilde{H}\| + \Delta(H, \tilde{H})^{\frac{1}{2}}\sqrt{\log d} \\ + (\sigma_*(H) + \sigma_*(\tilde{H}))x\}\big] \le de^{-x^2}\end{aligned}$$

for all $x \ge 0$, where C is a universal constant and $\Delta(H, \tilde{H})$ is as in Lemma 8.1.

Proof. Note first that

$$\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - H|^{-2p}]^{\frac{1}{2p}} \le \mathbf{E}[\|(z\mathbf{1} - H)^{-1}\|^{2p}]^{\frac{1}{2p}} \le \mathbf{E}\|(z\mathbf{1} - H)^{-1}\| + C\sqrt{p}\frac{\sigma_*(H)}{(\mathrm{Im}\,z)^2}$$

for a universal constant C as in the proof of Lemma 7.3. On the other hand,

$$\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - \tilde{H}|^{-2p}]^{\frac{1}{2p}} \ge d^{-\frac{1}{2p}}\mathbf{E}\|(z\mathbf{1} - \tilde{H})^{-1}\|.$$

Combining these bounds with Lemmas 5.5 and 8.1 yields

$$\begin{aligned}\mathbf{P}\Bigg[d^{-\frac{1}{2p}}\|(z\mathbf{1} - \tilde{H})^{-1}\| \ge \|(z\mathbf{1} - H)^{-1}\| + \frac{\|\mathbf{E}H - \mathbf{E}\tilde{H}\|}{(\mathrm{Im}\,z)^2} + 2p\,\frac{\Delta(H, \tilde{H})}{(\mathrm{Im}\,z)^3} \\ + C\sqrt{p}\frac{\sigma_*(H)}{(\mathrm{Im}\,z)^2} + \frac{\sigma_*(\tilde{H}) + \sigma_*(H)}{(\mathrm{Im}\,z)^2}\,x\Bigg] \le 4e^{-x^2/2}.\end{aligned}$$

Choosing $p = \lfloor \log d \rfloor$ and proceeding as in the proof of Proposition 7.4 yields

$$\mathbf{P}\Bigg[\|(z\mathbf{1}-\tilde H)^{-1}\| \ge L\bigg\{\|(z\mathbf{1}-H)^{-1}\| + \frac{\|\mathbf{E}H-\mathbf{E}\tilde H\|}{\varepsilon^2} + \frac{\Delta(H,\tilde H)\log d}{\varepsilon^3} + \frac{\sigma_*(\tilde H)+\sigma_*(H)}{\varepsilon^2}\,x\bigg\} \text{ for some } z\in \mathrm{sp}(\tilde H)+i\varepsilon\Bigg] \le de^{-x^2}$$

for all $x \ge 0$, where L is a universal constant. The same bound holds if we reverse the roles of H and $\tilde H$, and the conclusion follows readily from Lemma 7.1. □

We finally formulate a variant of Lemma 8.1 for the resolvent.

Lemma 8.3. *Let $H, \tilde H$ be self-adjoint Gaussian random matrices. Then we have*

$$\big\|\mathbf{E}[(z\mathbf{1}-H)^{-1}] - \mathbf{E}[(z\mathbf{1}-\tilde H)^{-1}]\big\| \le \frac{\|\mathbf{E}H-\mathbf{E}\tilde H\|}{(\mathrm{Im}\,z)^2} + \frac{\Delta(H,\tilde H)}{(\mathrm{Im}\,z)^3},$$

where $\Delta(H,\tilde H)$ is as in Lemma 8.1.

Proof. As in the proof of Lemma 8.1, it suffices to assume that $\mathbf{E}H = \mathbf{E}\tilde H$. Moreover, applying Gaussian interpolation as in the proof of Lemma 8.1 yields

$$\frac{d}{dt}\mathbf{E}[\langle v,(z\mathbf{1}-H(t))^{-1}w\rangle] = \mathbf{E}[\langle v,(z\mathbf{1}-H(t))^{-1}Y(z\mathbf{1}-H(t))^{-1}Y(z\mathbf{1}-H(t))^{-1}w\rangle] - \mathbf{E}[\langle v,(z\mathbf{1}-H(t))^{-1}\tilde Y(z\mathbf{1}-H(t))^{-1}\tilde Y(z\mathbf{1}-H(t))^{-1}w\rangle].$$

Integrating this identity and taking the supremum over v, w with $\|v\| = \|w\| = 1$ readily yields the conclusion, where we use that $Y, \tilde Y$ are independent of $H(t)$. □

8.2. **The truncation error.** In order to apply the above comparison principles to $G, \tilde G$, we must estimate the relevant parameters in this case.

Lemma 8.4. $\|\mathbf{E}G-\mathbf{E}\tilde G\| \le \sqrt{2}\,\sigma_*(X)$ *for* $R \ge \sqrt{2}\,\bar R(X)$.

Proof. We first note that by the independence of $Z_1,\ldots,Z_n$

$$\mathbf{E}\tilde G - \mathbf{E}G = \sum_{i=1}^n \mathbf{E}[1_{\|Z_i\|\le R}Z_i] = \sum_{i=1}^n b_i^{-1}\mathbf{E}[1_{\max_j\|Z_j\|\le R}Z_i],$$

where $b_i := \mathbf{P}[\max_{j\ne i}\|Z_j\| \le R] \ge \mathbf{P}[\max_j\|Z_j\|\le R]$. We therefore obtain

$$\|\mathbf{E}\tilde G-\mathbf{E}G\| = \sup_{\|v\|=\|w\|=1}\left|\mathbf{E}\left[1_{\max_j\|Z_j\|\le R}\sum_{i=1}^n b_i^{-1}\langle v, Z_i w\rangle\right]\right| \le \frac{\sigma_*(X)}{\mathbf{P}[\max_j\|Z_j\|\le R]^{\frac12}}$$

by Cauchy-Schwarz. It remains to note that $\mathbf{P}[\max_j\|Z_j\| \le R] \ge \frac12$ whenever $R \ge \sqrt{2}\,\mathbf{E}[\max_j\|Z_j\|^2]^{\frac12} =: \sqrt{2}\,\bar R(X)$ by Markov's inequality. □

Lemma 8.5. $\Delta(G,\tilde G) \le 24\,\bar R(X)\sigma(X)$ *for* $R \ge \bar R(X)^{\frac12}\sigma(X)^{\frac12}$.

Proof. Suppose first that $M \ge 0$ with $\|M\| \le 1$. We begin by writing

$$\mathbf{E}[(G-\mathbf{E}G)M(G-\mathbf{E}G)] - \mathbf{E}[(\tilde G-\mathbf{E}\tilde G)M(\tilde G-\mathbf{E}\tilde G)] = \sum_{i=1}^n\big\{\mathbf{E}[1_{\|Z_i\|>R}Z_iMZ_i] + \mathbf{E}[1_{\|Z_i\|\le R}Z_i]\,M\,\mathbf{E}[1_{\|Z_i\|\le R}Z_i]\big\} \ge 0.$$

Now note that as $\mathbf{E}[Z_i]=0$, we have $\mathbf{E}[1_{\|Z_i\|\le R}Z_i] = -\mathbf{E}[1_{\|Z_i\|>R}Z_i]$. Moreover, for any self-adjoint random matrix Y, we have $\mathbf{E}[Y]\,M\,\mathbf{E}[Y] \le \mathbf{E}[Y]^2 \le \mathbf{E}[Y^2]$ using $\|M\|\le 1$ and Jensen's inequality. We therefore obtain

$$\|\mathbf{E}[(G-\mathbf{E}G)M(G-\mathbf{E}G)]-\mathbf{E}[(\tilde{G}-\mathbf{E}\tilde{G})M(\tilde{G}-\mathbf{E}\tilde{G})]\| \le 2\sup_{\|v\|=1}\sum_{i=1}^n \mathbf{E}[1_{\|Z_i\|>R}\|Z_iv\|^2].$$

To proceed, let Z_i' be independent copies of Z_i, and note that

$$\begin{aligned}\sum_{i=1}^n \mathbf{E}[1_{\|Z_i\|>R}\|Z_iv\|^2] &\le \sum_{i=1}^n \mathbf{E}[1_{\max_j\|Z_j\|>R}\|Z_iv\|^2]\\ &\le \sum_{i=1}^n \mathbf{E}[1_{\max_j\|Z_j\|>R}(\|Z_iv\|^2-\|Z_i'v\|^2)] + \mathbf{P}[\max_j\|Z_j\|>R]\,\sigma(X)^2\end{aligned}$$

for $\|v\|=1$. Moreover, we have

$$\sum_{i=1}^n \mathbf{E}[1_{\max_j\|Z_j\|>R}(\|Z_iv\|^2-\|Z_i'v\|^2)] \le \mathbf{E}\bigg[\bigg|\sum_{i=1}^n(\|Z_iv\|^2-\|Z_i'v\|^2)\bigg|\bigg] =$$
$$\mathbf{E}\bigg[\bigg|\sum_{i=1}^n\varepsilon_i(\|Z_iv\|^2-\|Z_i'v\|^2)\bigg|\bigg] \le 2\,\mathbf{E}\bigg[\bigg|\sum_{i=1}^n\varepsilon_i\|Z_iv\|^2\bigg|\bigg] \le 2\,\mathbf{E}\bigg[\bigg(\sum_{i=1}^n\|Z_iv\|^4\bigg)^{\frac12}\bigg],$$

where ε_i are i.i.d. random signs that are independent of the other variables, we used that the distribution of (Z_i, Z_i') is invariant under exchanging Z_i and Z_i' for any i, and we applied Jensen's inequality conditionally on (Z_i) in the last inequality. Bounding $\|Z_iv\|^4 \le (\max_j\|Z_j\|^2)\|Z_iv\|^2$ and applying Cauchy-Schwarz yields

$$\sum_{i=1}^n \mathbf{E}[1_{\max_j\|Z_j\|>R}(\|Z_iv\|^2-\|Z_i'v\|^2)] \le 2\,\mathbf{E}[\max_j\|Z_j\|^2]^{\frac12}\sigma(X)$$

for $\|v\|=1$. Putting together all the above estimates yields

$$\begin{aligned}&\|\mathbf{E}[(G-\mathbf{E}G)M(G-\mathbf{E}G)]-\mathbf{E}[(\tilde{G}-\mathbf{E}\tilde{G})M(\tilde{G}-\mathbf{E}\tilde{G})]\| \le\\ &2\,\mathbf{P}[\max_j\|Z_j\|>R]\,\sigma(X)^2 + 4\,\mathbf{E}[\max_j\|Z_j\|^2]^{\frac12}\sigma(X) \le 6\,\mathbf{E}[\max_j\|Z_j\|^2]^{\frac12}\sigma(X)\end{aligned}$$

for $R\ge \mathbf{E}[\max_j\|Z_j\|^2]^{\frac14}\sigma(X)^{\frac12}$ using Markov's inequality in the last step.

Finally, note that any matrix M with $\|M\|\le 1$ can be written as $M = \operatorname{Re}M + i\operatorname{Im}M$ with $\|\operatorname{Re}M\| = \frac12\|M+M^*\|\le 1$ and $\|\operatorname{Im}M\| = \frac12\|M-M^*\|\le 1$. As any self-adjoint matrix is the difference of its positive and negative parts, we can write $M = M_1 - M_2 + iM_3 - iM_4$ with $M_i\ge 0$ with $\|M_i\|\le 1$. Applying the above estimate to each M_i and using the triangle inequality concludes the proof. ☐

Finally, we must bound the matrix parameters of $\tilde{X}$.

Lemma 8.6. *We have $R(\tilde{X})\le 2R$, $\sigma_*(\tilde{X})\le\sigma_*(X)$, and $\sigma(\tilde{X})\le\sigma(X)$.*

Proof. The first inequality follows immediately from

$$R(\tilde{X}) = \Big\|\max_{1\le i\le n}\big\|1_{\|Z_i\|\le R}Z_i - \mathbf{E}[1_{\|Z_i\|\le R}Z_i]\big\|\Big\|_\infty$$

and the triangle inequality. Next, note that for any (complex) random variable Y and event A, we have $\mathbf{E}[|1_A Y - \mathbf{E}[1_A Y]|^2] \le \mathbf{E}[1_A|Y|^2] \le \mathbf{E}[|Y|^2]$. Thus

$$\begin{aligned}\mathbf{E}[|\langle v,(\tilde X-\mathbf{E}\tilde X)w\rangle|^2] &= \sum_{i=1}^n \mathbf{E}[|1_{\|Z_i\|\le R}\langle v,Z_iw\rangle - \mathbf{E}[1_{\|Z_i\|\le R}\langle v,Z_iw\rangle]|^2]\\ &\le \sum_{i=1}^n \mathbf{E}[|\langle v,Z_iw\rangle|^2] = \mathbf{E}[|\langle v,(X-\mathbf{E}X)w\rangle|^2]\end{aligned}$$

for any nonrandom vectors v,w. The remaining bounds follow, respectively, by taking the supremum over $\|v\|=\|w\|=1$, or by summing over the coordinate basis $w=e_k$ and then taking the supremum over $\|v\|=1$. □

8.3. **Proof of Theorems 2.7 and 2.10.** We now put everything together.

Proof of Theorem 2.7. As $X=\tilde X$ on the event $\{\max_i\|Z_i\|\le R\}$, we obtain

$$\begin{aligned}&\mathbf{P}\Big[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(\tilde G)) > C\tilde\varepsilon(t),\ \max_{1\le i\le n}\|Z_i\|\le R\Big]\\ &\qquad\le \mathbf{P}\big[\mathrm{d_H}(\mathrm{sp}(\tilde X),\mathrm{sp}(\tilde G)) > C\tilde\varepsilon(t)\big] \le de^{-t}\end{aligned}$$

with

$$\tilde\varepsilon(t) := \sigma_*(\tilde X)\,t^{\frac12} + R(\tilde X)^{\frac13}\sigma(\tilde X)^{\frac23}t^{\frac23} + R(\tilde X)\,t.$$

from Theorem 2.5. Moreover, we can replace $\tilde\varepsilon(t)$ by $\varepsilon_R(t)$ on the left-hand side of the above inequality as $\tilde\varepsilon(t)\le\varepsilon_R(t)$ by Lemma 8.6.

On the other hand, Proposition 8.2 and Lemmas 8.4, 8.5 and 8.6 imply

$$\mathbf{P}\big[\mathrm{d_H}(\mathrm{sp}(\tilde G),\mathrm{sp}(G)) > C\{R+\sigma_*(X)\}t^{\frac12}\big] \le de^{-t}$$

for all $t\ge 0$ and $R\ge R_0 := \bar R(X)^{\frac12}\sigma(X)^{\frac12} + \sqrt2\,\bar R(X)$. Here we used that we may assume without loss of generality that $t\ge\log d$ in the above estimate (as otherwise the right-hand side exceeds one and the bound is trivial), and that $\Delta(G,\tilde G)\le 24R^2$ by the assumption on R and Lemma 8.5. We may once again replace $\{R+\sigma_*(X)\}t^{\frac12}$ by $\varepsilon_R(t)$ on the left-hand side as $t^{\frac12}\lesssim t$ for $t\ge\log d$.

Combining the above bounds, we obtain

$$\begin{aligned}&\mathbf{P}\Big[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(G)) > 2C\varepsilon_R(t),\ \max_{1\le i\le n}\|Z_i\|\le R\Big]\\ &\le \mathbf{P}\Big[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(\tilde G)) + \mathrm{d_H}(\mathrm{sp}(\tilde G),\mathrm{sp}(G)) > 2C\varepsilon_R(t),\ \max_{1\le i\le n}\|Z_i\|\le R\Big]\\ &\le \mathbf{P}\Big[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(\tilde G)) > C\varepsilon_R(t),\ \max_{1\le i\le n}\|Z_i\|\le R\Big]\\ &\qquad + \mathbf{P}\big[\mathrm{d_H}(\mathrm{sp}(\tilde G),\mathrm{sp}(G)) > C\varepsilon_R(t)\big] \le 2de^{-t}.\end{aligned}$$

As in the proof of Theorem 2.5, the upper bound can be replaced by de^{-t} if we increase the value of the universal constant on the left-hand side.

This concludes the proof of the tail bound. To prove the expectation bound, note that choosing $R=R_0t$ in the tail bound yields

$$\mathbf{P}[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(G)) > C\delta(t)] \le e^{-t/2} + \mathbf{P}\Big[\max_{1\le i\le n}\|Z_i\| > R_0t\Big]$$

for $t\ge 2\log d$, where

$$\delta(t) := \sigma_*(X)\,t^{\frac12} + R_0^{\frac13}\sigma(X)^{\frac23}t + R_0t^2$$

and we used $de^{-t} \le e^{-t/2}$ for $t \ge 2\log d$. We now compute

$$\begin{aligned}
\mathbf{E}[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(G))] &\le C\delta(2\log d) + \int_{C\delta(2\log 2)}^{\infty} \mathbf{P}[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(G)) > x]\,dx \\
&= C\delta(2\log d) + C\int_{2\log d}^{\infty} \mathbf{P}[\mathrm{d_H}(\mathrm{sp}(X),\mathrm{sp}(G)) > C\delta(t)]\,\frac{d\delta(t)}{dt}\,dt \\
&\le C\delta(2\log d) + C\int_0^\infty e^{-t/2}\frac{d\delta(t)}{dt}\,dt + C\int_0^\infty \mathbf{P}\bigg[\max_{1\le i\le n}\|Z_i\| > R_0 t\bigg]\frac{d\delta(t)}{dt}\,dt \\
&\lesssim \delta(2\log d) + \mathbf{E}[\delta(\max_i \|Z_i\|/R_0)] \lesssim \delta(2\log d).
\end{aligned}$$

The conclusion follows readily using the assumption $\bar{R}(X)\,(\log d)^3 \lesssim \sigma(X)$. □

Proof of Theorem 2.10. As $X = \tilde{X}$ on the event $\{\max_i \|Z_i\| \le R\}$, we have

$$\mathbf{E}[(z\mathbf{1}-X)^{-1}] = \mathbf{E}[(z\mathbf{1}-\tilde{X})^{-1}] + \mathbf{E}[\{(z\mathbf{1}-X)^{-1} - (z\mathbf{1}-\tilde{X})^{-1}\}1_{\max_i \|Z_i\|>R}].$$

Thus Markov's inequality yields

$$\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}] - \mathbf{E}[(z\mathbf{1}-\tilde{X})^{-1}]\big\| \le \frac{2\,\mathbf{P}[\max_i \|Z_i\| > R]}{\mathrm{Im}\,z} \le \frac{2\bar{R}(X)^2}{R^2\,\mathrm{Im}\,z}.$$

Applying Theorem 2.9 and Lemmas 8.3, 8.4, 8.5 and 8.6 yields

$$\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}] - \mathbf{E}[(z\mathbf{1}-G)^{-1}]\big\| \lesssim \frac{\bar{R}(X)^2}{R^2\,\mathrm{Im}\,z} + \frac{\sigma_*(X)}{(\mathrm{Im}\,z)^2} + \frac{R^2}{(\mathrm{Im}\,z)^3} + \frac{R\sigma(X)^2 + R^3\log d}{(\mathrm{Im}\,z)^4}$$

for $R \ge R_0 := \bar{R}(X)^{\frac12}\sigma(X)^{\frac12} + \sqrt{2}\,\bar{R}(X)$.

Now assume first that $\mathrm{Im}\,z \ge R_0$ and choose $R = R_0^{\frac12}(\mathrm{Im}\,z)^{\frac12}$. Then we obtain

$$\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}] - \mathbf{E}[(z\mathbf{1}-G)^{-1}]\big\| \lesssim \frac{\sigma_*(X) + R_0}{(\mathrm{Im}\,z)^2} + \frac{R_0^{\frac32}\log d}{(\mathrm{Im}\,z)^{\frac52}} + \frac{R_0^{\frac12}\sigma(X)^2}{(\mathrm{Im}\,z)^{\frac72}}.$$

In particular, if $\mathrm{Im}\,z \ge R_0^{\frac15}\sigma(X)^{\frac45} + R_0(\log d)^{\frac23}$, we obtain

$$\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}] - \mathbf{E}[(z\mathbf{1}-G)^{-1}]\big\| \lesssim \frac{\sigma_*(X) + R_0^{\frac15}\sigma(X)^{\frac45} + R_0(\log d)^{\frac23}}{(\mathrm{Im}\,z)^2}.$$

On the other hand, for $\mathrm{Im}\,z < R_0^{\frac15}\sigma(X)^{\frac45} + R_0(\log d)^{\frac23}$ we can estimate

$$\big\|\mathbf{E}[(z\mathbf{1}-X)^{-1}] - \mathbf{E}[(z\mathbf{1}-G)^{-1}]\big\| \le \frac{2}{\mathrm{Im}\,z} \lesssim \frac{R_0^{\frac15}\sigma(X)^{\frac45} + R_0(\log d)^{\frac23}}{(\mathrm{Im}\,z)^2}.$$

If $\bar{R}(X)(\log d)^{\frac53} \lesssim \sigma(X)$, then $R_0(\log d)^{\frac23} \lesssim R_0^{\frac15}\sigma(X)^{\frac45} \asymp \bar{R}(X)^{\frac{1}{10}}\sigma(X)^{\frac{9}{10}}$, and the first part of the theorem follows. The second part of the theorem now follows from [9, Lemma 5.11] as in the proof of Theorem 2.9. □

9. Applications: Proofs

9.1. **Random lifts.** The aim of this section is to prove Theorem 3.9. We will first prove a more general result, and then specialize to the case of lifts.

9.1.1. *Strong convergence.* In this section, we let $\Pi_1, \ldots, \Pi_k$ be i.i.d. uniformly distributed random $n \times n$ permutation matrices, and we fix $A_1, \ldots, A_k \in \mathrm{M}_d(\mathbb{C})$. We consider the random matrix

$$X = \sum_{i=1}^{k} (A_i \otimes \Pi_i + A_i^* \otimes \Pi_i^*),$$

and let $X^\perp$ be its restriction to $\mathbb{C}^d \otimes 1^\perp$. Recall that if $s_1, \ldots, s_{2k}$ is a free semicircular family, then $c_1, \ldots, c_k$ defined by $c_j = \frac{s_j + i s_{k+j}}{\sqrt{2}}$ is a free circular family.

Proposition 9.1. *Let $c_1, \ldots, c_k$ be a free circular family, and define*

$$X_{\mathrm{F}} = \sum_{i=1}^{k} \big\{ \big((1-\varepsilon^2)^{\frac{1}{2}} A_i + \varepsilon A_i^*\big) \otimes c_i + \big((1-\varepsilon^2)^{\frac{1}{2}} A_i + \varepsilon A_i^*\big)^* \otimes c_i^* \big\}$$

with $\varepsilon = \frac{1}{\sqrt{n-1}(\sqrt{n}+\sqrt{n-2})}$. Then

$$\mathbf{P}\big[\|X^\perp\| \geq \|X_{\mathrm{F}}\| + C\big\{ v^{\frac{1}{2}} \sigma^{\frac{1}{2}} (\log nd)^{\frac{3}{4}} + v t^{\frac{1}{2}} + R^{\frac{1}{3}} \sigma^{\frac{2}{3}} t^{\frac{2}{3}} + Rt \big\}\big] \leq 2nde^{-t}$$

for all $t \geq 0$, where C is a universal constant and

$$\sigma = \left\| \sum_{i=1}^{k} \left(A_i A_i^* + A_i^* A_i + \frac{A_i^2 + A_i^{*2}}{n-1} \right) \right\|^{\frac{1}{2}}, \qquad R = 2 \max_{1 \leq i \leq k} \|A_i\|,$$

and

$$v = \frac{2}{\sqrt{n-1}} \left\| \mathrm{Cov}\left(\sum_{i=1}^{k} A_i g_i \right) \right\|^{\frac{1}{2}},$$

where $g_1, \ldots, g_k$ are i.i.d. standard real Gaussians.

Proposition 9.1 is an immediate consequence of Theorem 2.14 once we prove that $\sigma(X^\perp) = \sigma$, $R(X^\perp) \leq R$, $\sigma_*(X^\perp) \leq v(X^\perp) \leq v$, and $\|X^\perp_{\mathrm{free}}\| = \|X_{\mathrm{F}}\|$. These facts will be established in the following lemmas, concluding the proof.

Lemma 9.2. *We have $\sigma(X^\perp) = \sigma$ and $R(X^\perp) \leq R$.*

Proof. The bound on $R(X^\perp)$ follows immediately from $\|\Pi_i\| = 1$. To compute $\sigma(X^\perp)$, we note that the restriction of Π_i to $1^\perp$ a random matrix as in Lemma 3.1 with $d = n-1$ and $s = 1$ (as this is an $(n-1)$-dimensional real representation of the symmetric group; alternatively, the conclusions of Lemma 3.1 can be verified in this case by a direct computation). We can therefore compute

$$\mathbf{E}[X^\perp] = 0, \qquad \mathbf{E}[(X^\perp)^2] = \sum_{i=1}^{k} \left(A_i A_i^* + A_i^* A_i + \frac{A_i^2 + A_i^{*2}}{n-1} \right) \otimes \mathbf{1},$$

and the conclusion follows immediately. □

Lemma 9.3. *We have $\|X^\perp_{\mathrm{free}}\| = \|X_{\mathrm{F}}\|$.*

Proof. Let $\bar{c}_i := (1-\varepsilon^2)^{\frac{1}{2}} c_i + \varepsilon c_i^*$. Then we write

$$X^\perp_{\mathrm{free}} = \sum_{i=1}^{k} (A_i \otimes \Pi^\perp_{i,\mathrm{free}} + A_i^* \otimes \bar{\Pi}^{\perp *}_{i,\mathrm{free}}), \qquad X_{\mathrm{F}} = \sum_{i=1}^{k} (A_i \otimes \bar{c}_i + A_i^* \otimes \bar{c}_i^*).$$

Now note that

$$\tau(\bar{c}_i) = 0, \qquad \tau(\bar{c}_i \bar{c}_i^*) = \tau(\bar{c}_i^* \bar{c}_i) = 1, \qquad \tau(\bar{c}_i^2) = \tau((\bar{c}_i^*)^2) = \frac{1}{n-1},$$

while applying Lemma 3.1 as in the proof of Lemma 9.2 yields

$$\mathbf{E}[\Pi_i^\perp] = 0, \quad \mathbf{E}[\Pi_i^\perp \Pi_i^{\perp *}] = \mathbf{E}[\Pi_i^{\perp *}\Pi_i^\perp] = \mathbf{1}, \quad \mathbf{E}[(\Pi_i^\perp)^2] = \mathbf{E}[(\Pi_i^{\perp *})^2] = \frac{1}{n-1}\mathbf{1}.$$

The conclusion follows as in the proof of [9, Lemma 7.9]. □

Lemma 9.4. *We have $v(X^\perp) \le v$.*

Proof. By Lemma 3.1, we have

$$\mathrm{Cov}(A_i \otimes \Pi_i^\perp) = \iota(A_i)\iota(A_i)^* \otimes \mathrm{Cov}(\Pi_i^\perp) = \iota(A_i)\iota(A_i)^* \otimes \frac{1}{n-1}\mathbf{1},$$

where $\iota : \mathrm{M}_d(\mathbb{C}) \to \mathbb{C}^{d^2}$ maps a matrix to its vector of entries. Therefore, as $\Pi_1, \ldots, \Pi_k$ are independent, we have

$$\mathrm{Cov}\bigg(\sum_{i=1}^k A_i \otimes \Pi_i^\perp\bigg) = \sum_{i=1}^k \mathrm{Cov}(A_i \otimes \Pi_i^\perp) = \frac{1}{n-1}\mathrm{Cov}\bigg(\sum_{i=1}^k A_i g_i\bigg) \otimes \mathbf{1}.$$

The conclusion follows from the triangle inequality $v(A+B) \le v(A) + v(B)$. □

9.1.2. *Free generators and circular variables.* The operator X_{F} in Proposition 9.1 is defined by a free circular family. In the study of random lifts, however, we are interested in the analogous operator where the circular variables c_i are replaced by the left-regular representation $\lambda(g_i)$ of the free generators of F_k. We presently establish a comparison principle between these objects.

Proposition 9.5. *Let $c_1, \ldots, c_k$ be a free circular family and let $g_1, \ldots, g_k$ be free generators of F_k. Then for any $A_1, \ldots, A_k \in \mathrm{M}_d(\mathbb{C})$, we have*

$$\bigg\|\sum_{i=1}^k (A_i \otimes c_i + A_i^* \otimes c_i^*)\bigg\| \le \bigg\|\sum_{i=1}^k (A_i \otimes \lambda(g_i) + A_i^* \otimes \lambda(g_i)^*)\bigg\| + 2\tilde\sigma^{\frac12}\tilde R^{\frac12},$$

where $\tilde\sigma^2 = \|\sum_{i=1}^k (A_iA_i^* + A_i^*A_i)\|$ and $\tilde R = \max_{1\le i\le k}\|A_i\|$.

Proof. We proceed in several steps.

Step 1. We begin by noting that if $c_1, \ldots, c_k$ is a circular family, $c_1^*, \ldots, c_k^*$ is also a circular family. We can therefore write

$$\bigg\|\sum_{i=1}^k (A_i \otimes c_i + A_i^* \otimes c_i^*)\bigg\| = \bigg\|\sum_{i=1}^k (A_i^* \otimes c_i + A_i \otimes c_i^*)\bigg\| = \bigg\|\sum_{i=1}^k (\tilde A_i \otimes c_i + \tilde B_i \otimes c_i^*)\bigg\|,$$

where we defined the self-adjoint matrices

$$\tilde A_i = \begin{bmatrix} 0 & A_i \\ A_i^* & 0 \end{bmatrix}, \qquad \tilde B_i = \begin{bmatrix} 0 & A_i^* \\ A_i & 0 \end{bmatrix}.$$

Similarly, if $g_1, \ldots, g_k$ are free generators of F_k, then $g_1^{-1}, \ldots, g_k^{-1}$ are as well, and thus the analogous identities holds when c_i is replaced by $\lambda(g_i)$.

Step 2. For any $A, M \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ with $M > 0$, define

$$\mathrm{R}_A(M) = M^{\frac12}\big((\mathbf{1} + (M^{-\frac12}AM^{-\frac12})^2)^{\frac12} - \mathbf{1}\big)M^{\frac12}.$$

Then by [46, Theorem 1.1 and p. 454], we have

$$\bigg\|\sum_{i=1}^k (\tilde A_i \otimes \lambda(g_i) + \tilde B_i \otimes \lambda(g_i)^*)\bigg\| = \inf_{M>0}\bigg\|2M + \sum_{i=1}^k (\mathrm{R}_{\tilde A_i}(M) + \mathrm{R}_{\tilde B_i}(M))\bigg\|.$$

On the other hand, as the circular family $c_1,\ldots,c_k$ can be realized by setting $c_i = l_i + l_{k+i}^*$ where $l_1,\ldots,l_{2k}$ are canoncial creation operators on the free Fock space (cf. [46, §1.2]), we obtain by [46, Corollary 1.4]

$$\bigg\|\sum_{i=1}^k(\tilde A_i\otimes c_i+\tilde B_i\otimes c_i^*)\bigg\| \le \inf_{M>0}\bigg\|2M+\sum_{i=1}^k \frac{\tilde A_iM^{-1}\tilde A_i+\tilde B_iM^{-1}\tilde B_i}{2}\bigg\|.$$

Step 3. We claim that for any $\delta>0$, there exists $M\ge\frac{\delta}{2}\mathbf{1}$ so that

$$2M+\sum_{i=1}^k(\mathrm{R}_{\tilde A_i}(M)+\mathrm{R}_{\tilde B_i}(M)) = \bigg(\bigg\|\sum_{i=1}^k(\tilde A_i\otimes\lambda(g_i)+\tilde B_i\otimes\lambda(g_i)^*)\bigg\|+\delta\bigg)\mathbf{1}.$$

Indeed, let $\tilde g_1,\ldots,\tilde g_{2k}$ be the free generators of the free product $\mathbb{Z}_2*\cdots*\mathbb{Z}_2$ of $2k$ copies of $\mathbb{Z}_2$, and define the operator $\tilde X=\sum_{i=1}^k(\tilde A_i\otimes\lambda(\tilde g_i)+\tilde B_i\otimes\lambda(\tilde g_{i+k}))$. Then [46, Theorem 1.1, Lemma 2.3 and Proposition 3.1] show that the above identity is satisfied if we choose the matrix M so that $(2M)^{-1}=(\mathrm{id}\otimes\tau)\big[\big((\|\tilde X\|+\delta)\mathbf{1}-\tilde X\big)^{-1}\big]$. As $\|\tilde X\|\mathbf{1}-\tilde X\ge 0$, we clearly have $(2M)^{-1}\le\delta^{-1}\mathbf{1}$, establishing the claim.

Step 4. Define the function

$$h(x)=\frac{x^2}{2}-((1+x^2)^{\frac12}-1)=\frac{x^4}{2((1+x^2)^{\frac12}+1)^2}.$$

Then clearly $h(x)\le\frac18x^4$. Thus we have for any self-adjoint A,M with $M>0$

$$\frac{AM^{-1}A}{2}\le \mathrm{R}_A(M)+\frac{AM^{-1}AM^{-1}AM^{-1}A}{8},$$

where we used that $\frac12AM^{-1}A-\mathrm{R}_A(M)=M^{\frac12}h(M^{-\frac12}AM^{-\frac12})M^{\frac12}$.

Step 5. We now put everything together. Let $\delta>0$ and choose M as in Step 3. Then we can estimate

$$\begin{aligned}
\bigg\|\sum_{i=1}^k(A_i\otimes c_i+A_i^*\otimes c_i^*)\bigg\| &\le \bigg\|2M+\sum_{i=1}^k\frac{\tilde A_iM^{-1}\tilde A_i+\tilde B_iM^{-1}\tilde B_i}{2}\bigg\| \\
&\le \bigg\|\sum_{i=1}^k(A_i\otimes\lambda(g_i)+A_i^*\otimes\lambda(g_i)^*)\bigg\|+\delta \\
&\qquad+\bigg\|\sum_{i=1}^k\frac{\tilde A_iM^{-1}\tilde A_iM^{-1}\tilde A_iM^{-1}\tilde A_i+\tilde B_iM^{-1}\tilde B_iM^{-1}\tilde B_iM^{-1}\tilde B_i}{8}\bigg\| \\
&\le \bigg\|\sum_{i=1}^k(A_i\otimes\lambda(g_i)+A_i^*\otimes\lambda(g_i)^*)\bigg\|+\delta+\frac{\max_i(\|\tilde A_i\|^2\vee\|\tilde B_i\|^2)}{\delta^3}\bigg\|\sum_{i=1}^k(\tilde A_i^2+\tilde B_i^2)\bigg\|,
\end{aligned}$$

where we used $\langle v,\tilde A_iM^{-1}\tilde A_iM^{-1}\tilde A_iM^{-1}\tilde A_iv\rangle\le\|\tilde A_iv\|^2\|M^{-1}\|^3\|\tilde A_i\|^2$ (and analogously for $\tilde B_i$) and $M\ge\frac{\delta}{2}$ in the last line. As $\max_i(\|\tilde A_i\|^2\vee\|\tilde B_i\|^2)=\tilde R^2$ and $\|\sum_i(\tilde A_i^2+\tilde B_i^2)\|=\tilde\sigma^2$, the conclusion follows by choosing $\delta=\tilde\sigma^{\frac12}\tilde R^{\frac12}$. $\square$

9.1.3. *Random n-lifts.* We now specialize the above results to the situation of random n-lifts. That is, we fix a base graph $H = ([d], E_H)$, set $k = |E_H|$, and $A_e = e_i e_j^*$ for $e = (i,j) \in E_H$, $i \le j$. Then the random matrix $X = X^{(n)}$ is the adjacency matrix of the random n-lift of H. Let us begin by estimating the parameters that appear in Propositions 9.1 and 9.5 in this case.

Lemma 9.6. *Denote by* $\mathrm{D}(H)$ *the maximal degree of a vertex of* H *and by* $\mathrm{M}(H)$ *the maximal multiplicity of an edge of* H*. Then we have*

$$\tilde{\sigma}^2 \le \sigma^2 \le 2\,\mathrm{D}(H), \qquad v^2 = \frac{4\,\mathrm{M}(H)}{n-1}, \qquad \tilde{R} \le R \le 2.$$

Proof. That $\tilde{R} \le R \le 2$ is immediate. Now let $A_e = e_i e_j^*$ for $e = (i,j) \in E_H$. Then $A_e A_e^* + A_e^* A_e = e_i e_i^* + e_j e_j^*$ and $A_e^2 + A_e^{*2} = 2 e_i e_i^* 1_{i=j}$. Therefore

$$\Bigg\| \sum_{e\in E_H} (A_e A_e^* + A_e^* A_e) \Bigg\| = \mathrm{D}(H), \qquad \Bigg\| \sum_{e\in E_H} (A_e^2 + A_e^{*2}) \Bigg\| = 2\,\mathrm{L}(H) \le \mathrm{D}(H),$$

where $\mathrm{L}(H)$ denotes the maximal number of self-loops attached to a vertex of H. The bounds on $\tilde{\sigma}, \sigma$ follows. Finally, note that $\sum_{e\in E_H} A_e g_e$, where $(g_e)_{e\in E_H}$ are i.i.d. standard real Gaussians, is a matrix with independent entries such that the variance of its (i,j) entry for $i \le j$ is the number of edges in H between vertices i and j. The computation of v follows immediately. □

Combining Proposition 9.1 and Lemma 9.6 yields an analogue of Theorem 3.9, in which $\varrho(\hat{H})$ is replaced by $\|X_{\mathrm{F}}\|$. Moreover, Proposition 9.5 and (3.2) readily imply that $\|X_{\mathrm{F}}\| \le (1 + Cn^{-1} + C\mathrm{D}(H)^{-\frac{1}{4}})\varrho(\hat{H})$. Thus the conclusion of Theorem 3.9 holds even when H has self-loops, but with an extra $O(n^{-1})$ error term. However, when H has no self-loops, the special structure of the coefficients A_e enables us to eliminate the $O(n^{-1})$ term, so that the bound can be sharp even when $n \not\to \infty$.

Lemma 9.7. *If* H *has no self-loops, then*

$$\|X_{\mathrm{F}}\| \le \left(1 + \frac{C}{\mathrm{D}(H)^{\frac{1}{4}}}\right) \varrho(\hat{H})$$

for a universal constant C*.*

Proof. Let ε be as in Proposition 9.1. As H is loopless, all A_e are of the form $A = e_i e_j^*$ with $i < j$. If we define $A_\varepsilon = (1-\varepsilon^2)^{\frac{1}{2}} A + \varepsilon A^*$, then we can compute

$$A_\varepsilon M A_\varepsilon^* + A_\varepsilon^* M A_\varepsilon = M_{jj} e_i e_i^* + M_{ii} e_j e_j^* + \frac{M_{ji} e_i e_j^* + M_{ij} e_j e_i^*}{n-1}.$$

The circular family $(c_e)_{e\in E_H}$ can be realized as $c_e = l_e + \tilde{l}_e^*$ where $l_e, \tilde{l}_e$ are canonical creation operators on a free Fock space [46, §1.2], we have [46, Theorem 1.3]

$$\|X_{\mathrm{F}}\| = \inf_{M>0} \Bigg\| M^{-1} + \sum_{e\in E_H} (A_{e\varepsilon} M A_{e\varepsilon}^* + A_{e\varepsilon}^* M A_{e\varepsilon}) \Bigg\|,$$

and moreover the infimum is attained by an M so that the quantity inside the norm on the right-hand side is proportional to the identity. We can now reason precisely as in the proof of [9, Lemma 3.2] that the infimum in the above expression can be taken over diagonal matrices only. Therefore

$$\|X_{\rm F}\| = \inf_{x\in\mathbb{R}^d:x>0} \max_{i\in[d]} \left\{\frac{1}{x_i} + \sum_{j\in[d]:j\sim i} x_j\right\},$$

where $j\sim i$ denotes that there is an edge between i,j in H. As the latter expression does not depend on n, we can conclude that when H has no self-loops, $\|X_F\|$ is unchanged if we set $\varepsilon=0$. Then Proposition 9.5, (3.2), and Lemma 9.6 yield

$$\|X_{\rm F}\| \le \varrho(\hat H) + C\mathrm{D}(H)^{\frac14}.$$

It remains to note that $\varrho(\hat H)\ge \mathrm{D}(H)^{\frac12}$ by (3.2) and [58, eq. (9.7.2)]. □

We now conclude the proof of Theorem 3.9.

Proof of Theorem 3.9. Applying Proposition 9.1 with $t=(a+2)\log nd$ yields

$$\mathbf{P}\left[\|X^{(n)\perp}\| \ge \left(1+C\frac{\mathrm{M}(H)^{\frac14}}{n^{\frac14}}\frac{(\log nd)^{\frac34}}{\mathrm{D}(H)^{\frac14}}+C\frac{(\log nd)^{\frac23}}{\mathrm{D}(H)^{\frac16}}+C\frac{\log nd}{\mathrm{D}(H)^{\frac12}}\right)\varrho(\hat H)\right] \le (nd)^{-a}$$

using Lemmas 9.6 and 9.7, that $\varrho(\hat H)\ge \mathrm{D}(H)^{\frac12}$ as in the proof of Lemma 9.7, and that $\mathrm{M}(H)\le\mathrm{D}(H)$. Here C is a constant that depends on a only. □

9.2. **Smallest singular value.** The aim of this section is to prove Theorem 3.12. The proof is based on the following linearization lemma, which we state in a slightly more general form than is needed here as it will be used again in section 9.3.

Lemma 9.8. *Let Y be a $d\times m$ random matrix, and let $B\ge 0$ be a nonrandom $m\times m$ positive semidefinite matrix. Define the $(d+2m)\times(d+2m)$ random matrix*

$$\hat Y_\varepsilon := \begin{bmatrix} 0 & Y^* & (B+4\varepsilon^2\mathbf{1})^{\frac12} \\ Y & 0 & 0 \\ (B+4\varepsilon^2\mathbf{1})^{\frac12} & 0 & 0\end{bmatrix},$$

and let $\hat Y_{\varepsilon,\rm free}$ be its noncommutative model. Then

$$\mathrm{sp}(\hat Y_\varepsilon)\subseteq \mathrm{sp}(\hat Y_{\varepsilon,\rm free})+[-\varepsilon,\varepsilon] \qquad\Longrightarrow$$
$$\begin{cases}\lambda_{\max}(Y^*Y+B+4\varepsilon^2\mathbf{1})^{\frac12}\le \lambda_{\max}(Y_{\rm free}^*Y_{\rm free}+B\otimes\mathbf{1}+4\varepsilon^2\mathbf{1})^{\frac12}+\varepsilon,\\ \lambda_{\min}(Y^*Y+B+4\varepsilon^2\mathbf{1})^{\frac12}\ge \lambda_{\min}(Y_{\rm free}^*Y_{\rm free}+B\otimes\mathbf{1}+4\varepsilon^2\mathbf{1})^{\frac12}-\varepsilon\end{cases}$$

for any $\varepsilon\ge0$, where $\lambda_{\max}(X):=\sup\mathrm{sp}(X)$ and $\lambda_{\min}(X):=\inf\mathrm{sp}(X)$.

Proof. The proof is identical to that of [9, Lemma 3.13]. □

We can now complete the proof of Theorem 3.12.

Proof of Theorem 3.12. We readily compute $\sigma_*(\hat Y_\varepsilon)=\sigma_*(Y)$, $\sigma(\hat Y_\varepsilon)=\sigma(Y)$, and $v(\hat Y_\varepsilon)\le\sqrt{2}\,v(Y)$ by [9, Lemma 4.10], while clearly $R(\hat Y_\varepsilon)=R(Y)$ by Remark 2.1. Applying Theorem 2.14 and Lemma 9.8 with $B=0$ therefore yields

$$\mathbf{P}\big[\lambda_{\min}(Y^*Y+4\delta(t)^2\mathbf{1})^{\frac12}\le\lambda_{\min}(Y_{\rm free}^*Y_{\rm free}+4\delta(t)^2\mathbf{1})^{\frac12}-\delta(t)\big]\le 6de^{-t},$$

where

$$\delta(t)=C\big\{v(Y)^{\frac12}\sigma(Y)^{\frac12}(\log d)^{\frac34}+\sigma_*(Y)\,t^{\frac12}+R(Y)^{\frac13}\sigma(Y)^{\frac23}t^{\frac23}+R(Y)\,t\big\}$$

for a universal constant C. Using that $\lambda_{\min}(Y_{\rm free}^*Y_{\rm free}+4\delta(t)^2\mathbf{1})^{\frac12}\geq \mathrm{s}_{\min}(Y_{\rm free})$ and $\lambda_{\min}(Y^*Y+4\delta(t)^2\mathbf{1})^{\frac12}\leq \mathrm{s}_{\min}(Y)+2\delta(t)$, we obtain

$$\mathbf{P}\big[\mathrm{s}_{\min}(Y)\leq \mathrm{s}_{\min}(Y_{\rm free})-3\delta(t)\big]\leq 6de^{-t}.$$

It remains to note that we can replace $6d$ by d on the right-hand side if we increase the universal constant C (as in the last step of the proof of Theorem 2.5). □

9.3. **Sample covariance matrices.**

9.3.1. *Proof of Theorem 3.17.* Gaussian random matrices are unbounded but possess moments of all orders. Therefore, a result along the lines of Theorem 3.17 can be proved either using Theorem 2.7 or using Theorem 2.8. These two approaches yield similar conclusions; we have chosen the latter approach here as it yields a slightly cleaner bound in the present setting. In preparation for the proof, let us estimate the relevant matrix parameters of the random matrix S of (3.4).

Lemma 9.9. *We have*

$$\sigma(S)=\bigg\|\sum_{i=1}^n\big(\mathrm{Tr}[\Sigma_i]\Sigma_i+\Sigma_i^2\big)\bigg\|^{\frac12},\qquad v(S)\leq\sqrt{2}\,\bigg\|\sum_{i=1}^n\Sigma_i^2\bigg\|^{\frac12}.$$

Proof. The identity for $\sigma(S)$ follows readily using

$$\mathbf{E}[(S-\mathbf{E}S)^2]=\sum_{i=1}^n\mathbf{E}[(Y_iY_i^*-\Sigma_i)^2]=\sum_{i=1}^n\big(\mathbf{E}[Y_iY_i^*\|Y_i\|^2]-\Sigma_i^2\big)$$

and that $\mathbf{E}[Y_iY_i^*\|Y_i\|^2]=\mathrm{Tr}[\Sigma_i]\Sigma_i+2\Sigma_i^2$ by the Wick formula (4.1).

To bound $v(S)$, we reason analogously. We first note that

$$\mathbf{E}[|\mathrm{Tr}[M(S-\mathbf{E}S)]|^2]=\sum_{i=1}^n\big(\mathbf{E}[|\langle Y_i,MY_i\rangle|^2]-|\mathrm{Tr}[M\Sigma_i]|^2\big)\leq 2\sum_{i=1}^n\mathrm{Tr}[M\Sigma_iM^*\Sigma_i],$$

using $\mathbf{E}[|\langle Y_i,MY_i\rangle|^2]\leq|\mathrm{Tr}[M\Sigma_i]|^2+2\,\mathrm{Tr}[M\Sigma_iM^*\Sigma_i]$ by (4.1) and Cauchy-Schwarz. As $2\,\mathrm{Tr}[M\Sigma_iM^*\Sigma_i]\leq 2\,\mathrm{Tr}[M\Sigma_i^2M^*]^{\frac12}\,\mathrm{Tr}[\Sigma_iMM^*\Sigma_i]^{\frac12}\leq\mathrm{Tr}[(M^*M+MM^*)\Sigma_i^2]$ by Cauchy-Schwarz and Young's inequality, we have

$$v(S)^2=\sup_{\mathrm{Tr}\,|M|^2\leq1}\mathbf{E}[|\mathrm{Tr}[M(S-\mathbf{E}S)]|^2]\leq\sup_{\mathrm{Tr}\,|M|^2\leq1}\mathrm{Tr}\bigg[(M^*M+MM^*)\sum_{i=1}^n\Sigma_i^2\bigg],$$

and the conclusion follows readily. □

We must further estimate the parameter $R_q(S)$ in Theorem 2.8.

Lemma 9.10. *For $q\geq1$, we have $R_q(S)\lesssim n^{\frac1q}\max_{i\leq n}\{\mathrm{Tr}\,\Sigma_i+q\|\Sigma_i\|\}$.*

Proof. It follows directly from the definition of $R_q(S)$ that

$$R_q(S)\leq n^{\frac1q}\max_{i\leq n}\mathbf{E}[\|Z_i\|^q]^{\frac1q}\leq 2n^{\frac1q}\max_{i\leq n}\mathbf{E}[\|Y_i\|^{2q}]^{\frac1q}$$

where we used $\|Z_i\|=\|Y_iY_i^*-\mathbf{E}Y_iY_i^*\|\leq\|Y_i\|^2+\mathbf{E}\|Y_i\|^2$. It remains to note that

$$\mathbf{E}[\|Y_i\|^{2q}]^{\frac{1}{2q}}\leq\mathbf{E}\|Y_i\|+\mathbf{E}[(\|Y_i\|-\mathbf{E}\|Y_i\|)^{2q}]^{\frac{1}{2q}}\lesssim(\mathrm{Tr}\,\Sigma_i)^{\frac12}+\|\Sigma_i\|^{\frac12}\sqrt{q},$$

where we used that $\mathbf{E}\|Y_i\|\leq(\mathrm{Tr}\,\Sigma_i)^{\frac12}$ by Cauchy-Schwarz and that $\|Y_i\|$ is $\|\Sigma_i\|$-subgaussian by Gaussian concentration [21, Theorems 5.6 and 2.1]. □

We can now complete the proof of Theorem 3.17.

Proof of Theorem 3.17. Theorem 2.8 and [9, Theorem 2.7 and Lemma 2.5] yield

$$d^{-\frac{1}{2p}}\mathbf{E}\|S-\mathbf{E}S\| \le \mathbf{E}[\mathrm{tr}(S-\mathbf{E}S)^{2p}]^{\frac{1}{2p}} \le 2\sigma(S) + Cv(S)^{\frac{1}{2}}\sigma(S)^{\frac{1}{2}}p^{\frac{3}{4}} + CR_{2p}(S)p^2$$

for a universal constant C. Now let $p=\lceil\frac{2}{\varepsilon}\log(d+n)\rceil$, so that $\max(d^{\frac{1}{2p}},n^{\frac{1}{2p}})\le e^{\frac{\varepsilon}{4}}$. Moreover, $Cv(S)^{\frac{1}{2}}\sigma(S)^{\frac{1}{2}}p^{\frac{3}{4}} \le (e^{\frac{\varepsilon}{4}}-1)2\sigma(S)+\varepsilon^{-1}C^2v(S)p^{\frac{3}{2}}$ by Young's inequality and $e^x\ge 1+x$. We therefore obtain for any $\varepsilon\in(0,1]$

$$\mathbf{E}\|S-\mathbf{E}S\| \le (1+\varepsilon)\,2\sigma(S) + \frac{K}{\varepsilon^3}\Big(v(S)+\max_{i\le n}\mathrm{Tr}\,\Sigma_i\Big)\log^3(d+n),$$

where K is a universal constant and we used Lemma 9.10 and $e^{\frac{\varepsilon}{2}}\le 1+\varepsilon$ for $\varepsilon\le 1$. The conclusion follows readily using Lemma 9.9. □

9.3.2. *A simple lower bound.* The aim of this short section is to show that the leading terms of the upper bounds of Theorems 3.16 and 3.17 are also lower bounds up to a universal constant. These results therefore capture the correct "user-friendly" quantity in the present setting. In general, it is not the case these these terms are optimal to leading order, that is up to a factor $1+o(1)$; if such a sharp bound is desired, the proofs of Theorems 3.16 and 3.17 may be adapted to obtain bounds in terms of $\|Y_{\rm free}Y_{\rm free}^*-\mathbf{E}S\otimes\mathbf{1}\|$ and $\|S_{\rm free}-\mathbf{E}S\otimes\mathbf{1}\|$, respectively.

Lemma 9.11. *In the setting of section 3.3.1, we have*

$$\mathbf{E}\|S-\mathbf{E}S\| \gtrsim \Bigg\|\sum_{i=1}^n \mathrm{Tr}[\Sigma_i]\Sigma_i\Bigg\|^{\frac{1}{2}} + \max_{i\le n}\mathrm{Tr}\,\Sigma_i.$$

Proof. We begin by noting that

$$\mathbf{E}\|S-\mathbf{E}S\| \ge \sup_{\|v\|=1}\mathbf{E}\|(S-\mathbf{E}S)v\| \gtrsim \sup_{\|v\|=1}\mathbf{E}[\|(S-\mathbf{E}S)v\|^2]^{\frac{1}{2}},$$

where the last inequality follows by hypercontractivity [43, Theorem 3.50] using that $\|(S-\mathbf{E}S)v\|^2$ is a polynomial of degree 4 of the Gaussian variables Y_{ij}. The first part of the proof of Lemma 9.9 therefore yields

$$\mathbf{E}\|S-\mathbf{E}S\| \gtrsim \Bigg\|\sum_{i=1}^n\big(\mathrm{Tr}[\Sigma_i]\Sigma_i+\Sigma_i^2\big)\Bigg\|^{\frac{1}{2}} \ge \Bigg\|\sum_{i=1}^n \mathrm{Tr}[\Sigma_i]\Sigma_i\Bigg\|^{\frac{1}{2}} \ge \max_{i\le n}\|\Sigma_i\|.$$

On the other hand, we can readily estimate by Jensen's inequality

$$\mathbf{E}\|S-\mathbf{E}S\| \ge \max_{i\le n}\mathbf{E}\|Y_iY_i^*-\Sigma_i\| \ge \max_{i\le n}\mathrm{Tr}\,\Sigma_i - \max_{i\le n}\|\Sigma_i\|,$$

where we used that $S-\mathbf{E}S=\sum_{i=1}^n(Y_iY_i^*-\Sigma_i)$ is a sum of independent centered random matrices. We can therefore estimate

$$\Bigg\|\sum_{i=1}^n \mathrm{Tr}[\Sigma_i]\Sigma_i\Bigg\|^{\frac{1}{2}} + \max_{i\le n}\mathrm{Tr}\,\Sigma_i \le \Bigg\|\sum_{i=1}^n \mathrm{Tr}[\Sigma_i]\Sigma_i\Bigg\|^{\frac{1}{2}} + \max_{i\le n}\|\Sigma_i\| + \mathbf{E}\|S-\mathbf{E}S\|,$$

and the proof is readily concluded. □

9.3.3. *Proof of Theorem 3.20.* The proof of Theorem 3.20 combines our universality principles with a linearization argument as in Lemma 9.8.

Proof of Theorem 3.20. Given $B = \|\mathbf{E}YY^*\|\mathbf{1} - \mathbf{E}YY^*$, define

$$\breve{Y}_\varepsilon := \begin{bmatrix} 0 & Y & (B+4\varepsilon^2\mathbf{1})^{\frac{1}{2}} \\ Y^* & 0 & 0 \\ (B+4\varepsilon^2\mathbf{1})^{\frac{1}{2}} & 0 & 0 \end{bmatrix},$$

and let $\breve{H}_\varepsilon$ be its Gaussian model. Then a completely analogous argument to the one used in the proof of Lemma 9.8 yields

$$\mathrm{d_H}(\mathrm{sp}(\breve{Y}_\varepsilon), \mathrm{sp}(\breve{H}_\varepsilon)) \le \varepsilon \quad\implies$$
$$\begin{cases} \big|\lambda_{\max}(YY^* + B + 4\varepsilon^2\mathbf{1})^{\frac{1}{2}} - \lambda_{\max}(HH^* + B + 4\varepsilon^2\mathbf{1})^{\frac{1}{2}}\big| \le \varepsilon, \\ \big|\lambda_{\min}(YY^* + B + 4\varepsilon^2\mathbf{1})^{\frac{1}{2}} - \lambda_{\min}(HH^* + B + 4\varepsilon^2\mathbf{1})^{\frac{1}{2}}\big| \le \varepsilon \end{cases}$$

for any $\varepsilon \ge 0$. Using $|a^{\frac{1}{2}} - b^{\frac{1}{2}}|(a^{\frac{1}{2}} + b^{\frac{1}{2}}) = |a-b|$ for $a, b \ge 0$, we obtain

$$\mathrm{d_H}(\mathrm{sp}(\breve{Y}_\varepsilon), \mathrm{sp}(\breve{H}_\varepsilon)) \le \varepsilon \quad\implies$$
$$\big|\|YY^* - \mathbf{E}YY^*\| - \|HH^* - \mathbf{E}HH^*\|\big| \le \big(\|Y\| + \|H\| + 2\|\mathbf{E}YY^*\|^{\frac{1}{2}} + 4\varepsilon\big)\varepsilon,$$

where we used that $|\|M\| - \|N\|| \le |\lambda_{\max}(M) - \lambda_{\max}(N)| \vee |\lambda_{\min}(M) - \lambda_{\min}(N)|$ and $\mathbf{E}HH^* = \mathbf{E}YY^*$. Furthermore, we have $\sigma_*(\breve{Y}_\varepsilon) = \sigma_*(Y)$, $\sigma(\breve{Y}_\varepsilon) = \sigma(Y)$, and $R(\breve{Y}_\varepsilon) = R(Y)$ by [9, Remark 2.6]. Thus Theorem 2.5 yields

$$\mathbf{P}\big[\big|\|YY^* - \mathbf{E}YY^*\| - \|HH^* - \mathbf{E}HH^*\|\big| > C\big(\|Y\| + \|H\| + \|\mathbf{E}YY^*\|^{\frac{1}{2}} + \varepsilon(t)\big)\varepsilon(t)\big] \le (2d+m)e^{-t}$$

for all $t > 0$, where C is a universal constant and $\varepsilon(t)$ is as in Theorem 2.5.

To proceed, note that

$$\mathbf{P}[\|H\| > \mathbf{E}\|H\| + C\varepsilon(t)] \le e^{-t}$$

by Gaussian concentration as in [9, Corollary 4.14], while

$$\mathbf{P}[\|Y\| > \mathbf{E}\|H\| + C\varepsilon(t)] \le (d+m)e^{-t}$$

by Corollary 2.6 and Remark 2.1. Combining the above bounds yields

$$\mathbf{P}\big[\big|\|YY^* - \mathbf{E}YY^*\| - \|HH^* - \mathbf{E}HH^*\|\big| > C\varepsilon(t)\,\mathbf{E}\|H\| + C\varepsilon(t)^2\big] \le C(d+m)e^{-t}$$

for a universal constant C, where we used that

$$\|\mathbf{E}YY^*\| = \|\mathbf{E}HH^*\| = \sup_{\|v\|=1} \mathbf{E}\|H^*v\|^2 \lesssim \sup_{\|v\|=1} (\mathbf{E}\|H^*v\|)^2 \le (\mathbf{E}\|H\|)^2$$

by hypercontractivity [43, Theorem 3.50]. The conclusion follows by integrating this tail bound as in the proof of Theorem 2.5. □

9.3.4. *Proof of Theorem 3.23.* Throughout this section we adopt the setting and notation of Theorem 3.23. Its proof combines two distinct universality principles. Let us begin by applying universality to Y.

Proposition 9.12. *We have*

$$\mathbf{E}\|S - \mathbf{E}S\| \le 2\|B\|_{\mathrm{HS}}\|B\|\sqrt{n} + \|B\|_{\mathrm{HS}}^2 + C\big\{\alpha^{\frac{1}{3}}\|B\|^{\frac{1}{3}}(\|B\|\sqrt{n} + \|B\|_{\mathrm{HS}})^{\frac{5}{3}} + \alpha^2\|B\|^2\big\}\log^2(d+n).$$

Proof. We will write Y in the form (2.1) as

$$Y = \sum_{i=1}^{N}\sum_{j=1}^{n} Z_{ij}, \qquad Z_{ij} = A_{ij}\, Be_i e_j^*.$$

We readily compute

$$\mathbf{E}[YY^*] = n\,BB^*, \qquad \mathbf{E}[Y^*Y] = \|B\|_{\mathrm{HS}}^2\mathbf{1}, \qquad v(Y) = \|B\|, \qquad R(Y) \le \alpha\|B\|$$

(here we used that Y has independent columns $Y_1,\ldots,Y_n$ with $\mathrm{Cov}(Y_i) = BB^*$, so that $\|\mathrm{Cov}(Y)\| = \max_i \|\mathrm{Cov}(Y_i)\| = \|B\|^2$). Applying Theorem 3.20 yields

$$\big|\mathbf{E}\|S - \mathbf{E}S\| - \mathbf{E}\|HH^* - \mathbf{E}HH^*\|\big| \lesssim \delta\,(\|B\|\sqrt{n} + \|B\|_{\mathrm{HS}}) + \delta^2,$$

where

$$\delta \lesssim \alpha^{\frac{1}{3}}\|B\|^{\frac{1}{3}}(\|B\|\sqrt{n} + \|B\|_{\mathrm{HS}})^{\frac{2}{3}}\log^{\frac{2}{3}}(d+n) + \alpha\|B\|\log(d+n)$$

and we used that $\mathbf{E}\|H\| \lesssim \|B\|\sqrt{n} + \|B\|_{\mathrm{HS}}$ (see, e.g., [70, Lemma 5.4]) and $\alpha \ge 1$. On the other hand, $\mathbf{E}\|HH^* - \mathbf{E}HH^*\|$ can be estimated by Theorem 3.16 with $\Sigma_i = BB^*$. Combining all the above bounds yields the conclusion. □

We now apply universality to S. In preparation for the following computations, we begin by estimating $\sigma(S)$ and $v(S)$. Recall that Y_i denotes the ith column of Y.

Lemma 9.13. *We have*

$$\sigma(S) \le (\|B\|_{\mathrm{HS}}\|B\| + 2\alpha\|B\|^2)\sqrt{n}, \qquad v(S) \le 4\alpha\|B\|^2\sqrt{n}.$$

Proof. By Lemma 4.1, if $W_1,\ldots,W_m$ are independent centered random variables with unit variance and $g_1,\ldots,g_m$ are independent standard Gaussians, then

$$\mathbf{E}[W_iW_jW_kW_l] = \mathbf{E}[g_ig_jg_kg_l] + (\mathbf{E}[W_i^4] - 3)1_{i=j=k=l}.$$

We will apply this identity in the case that W_i are entries of the random matrix A. In particular, arguing as in the first part of the proof of Lemma 9.9, we obtain

$$\mathbf{E}[(S-\mathbf{E}S)^2] = n\|B\|_{\mathrm{HS}}^2 BB^* + n(BB^*)^2 + \sum_{i=1}^{N}\sum_{j=1}^{n}(\mathbf{E}[A_{ij}^4] - 3)(B^*B)_{ii}Be_ie_i^*B^*.$$

As $\mathbf{E}[A_{ij}^4] \le \alpha^2$, we readily obtain

$$\sigma(S)^2 \le n\|B\|_{\mathrm{HS}}^2\|B\|^2 + (1+\alpha^2)n\|B\|^4,$$

and the bound on $\sigma(S)$ follows using $\alpha \ge 1$.

The parameter $v(S)$ can be estimated analogously, but an adequate bound also follows from a standard concentration argument. Indeed, note that

$$v(S)^2 = \sup_{\mathrm{Tr}\,|M|^2=1}\mathrm{Var}(\mathrm{Tr}\,MS) = \sup_{\mathrm{Tr}\,|M|^2=1}\sum_{i=1}^{n}\mathrm{Var}(\langle Y_i, MY_i\rangle).$$

By the convex Poincaré inequality [21, Theorem 3.17] and Cauchy-Schwarz, we obtain $\mathrm{Var}(\langle Y_i, MY_i\rangle) \le 16\alpha^2\|B\|^4$ for $\mathrm{Tr}\,|M|^2 = 1$, concluding the proof. □

Next, we estimate the parameter $R_q(S)$ in Theorem 2.8.

Lemma 9.14. *For $q \ge 1$, we have $R_q(S) \lesssim n^{\frac{1}{q}}\{\|B\|_{\mathrm{HS}}^2 + \alpha^2 q\|B\|^2\}$.*

Proof. The convex concentration inequality [21, Theorems 6.10 and 2.1] yields that $\mathbf{E}[\|Y_i\|^{2q}]^{\frac{1}{2q}} \le \|B\|_{\mathrm{HS}} + C\alpha\sqrt{q}\|B\|$, where we used $\mathbf{E}\|Y_i\| \le \mathbf{E}[\|Y_i\|^2]^{\frac{1}{2}} = \|B\|_{\mathrm{HS}}$. The conclusion follows directly as in the proof of Lemma 9.10. □

We can now proceed as in the proof of Theorem 3.17.

Proposition 9.15. *We have for* $\varepsilon \in (0,1]$

$$\mathbf{E}\|S - \mathbf{E}S\| \le (1+\varepsilon)\, 2\|B\|_{\mathrm{HS}}\|B\|\sqrt{n} + \frac{C}{\varepsilon^3}\big(\|B\|_{\mathrm{HS}}^2 + (\alpha\sqrt{n}+\alpha^2)\|B\|^2\big)\log^3(d+n).$$

Proof. Apply Lemmas 9.13 and 9.14 precisely as in the proof of Theorem 3.17. □

We now complete the proof of Theorem 3.23.

Proof of Theorem 3.23. It is convenient to define $\gamma = \frac{\|B\|_{\mathrm{HS}}}{\|B\|\sqrt{n}}$ and $\delta = \frac{\alpha}{\sqrt{n}}$. In terms of these dimensionless parameters, Proposition 9.12 can be expressed as

$$\frac{\mathbf{E}\|S - \mathbf{E}S\|}{n\|B\|^2} \le 2\gamma + \gamma^2 + C\big\{\delta^{\frac{1}{3}}(1+\gamma)^{\frac{5}{3}} + \delta^2\big\}\log^2(d+n)$$

while Proposition 9.15 yields

$$\frac{\mathbf{E}\|S - \mathbf{E}S\|}{n\|B\|^2} \le 2\gamma + C\big\{(\gamma^2+\delta+\delta^2)^{\frac{1}{4}}\gamma^{\frac{3}{4}} + \gamma^2 + \delta + \delta^2\big\}\log^3(d+n),$$

where in the last equation we used $\inf_{\varepsilon\le 1}\{2\varepsilon\gamma + \frac{K}{\varepsilon^3}\} \le 3K^{\frac{1}{4}}\gamma^{\frac{3}{4}} + 3K$.

Now note that the assumptions of the theorem imply $\delta \le 1$ and $\gamma \ge \delta$. Using $\delta^2 \le \delta^{\frac{1}{3}}(1+\gamma)^{\frac{5}{3}}$, $\frac{\delta^{\frac{1}{3}}(1+\gamma)^{\frac{5}{3}}}{2\gamma+\gamma^2} \lesssim \frac{\delta^{\frac{1}{3}}}{\gamma} + (\frac{\delta}{\gamma})^{\frac{1}{3}}$ and $\frac{\delta}{\gamma} \le 1$, we can rearrange and combine the above inequalities to estimate

$$\frac{\mathbf{E}\|S - \mathbf{E}S\|}{n\|B\|^2} \le \left(1 + C\left\{\min\left(\frac{\delta^{\frac{1}{3}}}{\gamma}, \gamma^{\frac{1}{4}} + \gamma\right) + \frac{\delta^{\frac{1}{4}}}{\gamma^{\frac{1}{4}}}\right\}\log^3(d+n)\right)(2\gamma+\gamma^2).$$

We conclude with $\min(\frac{a}{\gamma}, \gamma^{\frac{1}{4}}+\gamma) \le \min(\frac{a}{\gamma}, \gamma^{\frac{1}{4}}) + \min(\frac{a}{\gamma}, \gamma) \le a^{\frac{1}{5}} + a^{\frac{1}{2}}$ and $\delta < 1$. □

Remark 9.16 (Unbounded entries). The formulation of Theorem 3.23 for bounded A_{ij} is not a fundamental restriction of our approach: results for unbounded entries can be obtained analogously by using our universality principles for unbounded random matrices. We have restricted to the bounded case largely for simplicity and brevity of exposition. However, in order to illustrate some features of the unbounded case, let us briefly discuss these here in the context of the slightly simpler problem of estimating $\|Y\|$ (as opposed to $\|YY^* - \mathbf{E}YY^*\|$).

Let $Y = BA$ with A, B as in Theorem 3.23, except that we now assume only that $\|A_{ij}\|_s \le \alpha$ for some $4 < s < \infty$. We write Y in the form (2.1) as in the proof of Proposition 9.12. Applying Theorem 2.7 as in the proof of Corollary 2.15 yields

$$\mathbf{E}\|Y\| \le \|\mathbf{E}YY^*\|^{\frac{1}{2}} + \|\mathbf{E}Y^*Y\|^{\frac{1}{2}} + C\big\{v(Y)^{\frac{1}{2}}\sigma(Y)^{\frac{1}{2}} + \bar{R}(Y)^{\frac{1}{6}}\sigma(Y)^{\frac{5}{6}}\big\}\log(d+n)$$

provided that $\bar{R}(Y)\log^3(d+n) \le \sigma(Y)$. All parameters in this bound were already computed in the proof of Proposition 9.12 except $\bar{R}(Y)$, which we estimate as

$$\bar{R}(Y) = \mathbf{E}\bigg[\max_{i\le N}\max_{j\le n} A_{ij}^2\|Be_i\|^2\bigg]^{\frac{1}{2}} \le \mathbf{E}\bigg[\sum_{i=1}^N\sum_{j=1}^n |A_{ij}|^s\|Be_i\|^s\bigg]^{\frac{1}{s}}$$
$$\le \alpha n^{\frac{1}{s}}\bigg(\sum_{i=1}^N \|Be_i\|^s\bigg)^{\frac{1}{s}} \le \alpha n^{\frac{1}{s}}\|B\|^{1-\frac{2}{s}}\|B\|_{\mathrm{HS}}^{\frac{2}{s}}.$$

Combining the above estimates, we obtain the bound

$$\mathbf{E}\|Y\| \le \bigg(1+\frac{C\log(d+n)}{(n\vee r)^{\frac{1}{4}}}+\frac{C\alpha^{\frac{1}{6}}\log(d+n)}{(n\vee r)^{\frac{1}{12}-\frac{1}{3s}}}\bigg)(\|B\|\sqrt{n}+\|B\|_{\mathrm{HS}})$$

for $\alpha\log^3(d+n) \le (n\vee r)^{\frac{1}{2}-\frac{2}{s}}$, where $r = \|B\|_{\mathrm{HS}}^2\|B\|^{-2}$ is the effective rank of B and we used that $\|B\|\sqrt{n}+\|B\|_{\mathrm{HS}} \ge (\|B\|\sqrt{n})^{1-t}\|B\|_{\mathrm{HS}}^t$ for any $t\in[0,1]$.

This result should be compared to the best previous bound in this setting due to Vershynin [72], which states that $\mathbf{E}\|Y\| \le C(s)\alpha(\|B\|\sqrt{n}+\|B\|_{\mathrm{HS}})$ where $C(s)<\infty$ for $s>4$. In contrast, our bound yields $\mathbf{E}\|Y\| \le (1+o(1))(\|B\|\sqrt{n}+\|B\|_{\mathrm{HS}})$ as soon as $n\vee r \gg (\alpha^{\frac{1}{6}}\log(d+n))^{\beta(s)}$, which not only yields the best possible constant for a bound of this kind (in view of the Bai-Yin law) but also allows α to diverge without affecting the estimate to leading order. We emphasize a key feature of both bounds that was highlighted in [72]: they do not depend on the inner dimension N, despite that $\bar{R}(Y)$ is defined as a maximum over nN random variables.

9.4. **Strong asymptotic freeness.** The main aim of this section is to prove Theorem 3.25. In section 9.4.1, we first develop the special case of bounded random matrices by means of a linearization argument. The result is then extended to the general setting in section 9.4.2 by employing a truncation argument as in section 8. Finally, Corollary 3.28 will be proved in section 9.4.3.

9.4.1. *Linearization.* The aim of this section is to prove the following special case of Theorem 3.25 for bounded random matrices. The general case will be deduced from this result in the next section by a truncation argument.

Theorem 9.17. *Let $s_1,\ldots,s_m$ be a free semicircular family, and let $H_1^N,\ldots,H_m^N$ be independent self-adjoint random matrices as in Theorem 3.25. Suppose*

$$\lim_{N\to\infty}\|\mathbf{E}[H_k^N]\| = \lim_{N\to\infty}\|\mathbf{E}[(H_k^N)^2]-\mathbf{1}\| = 0$$

and

$$\lim_{N\to\infty}(\log d_N)^{\frac{3}{2}}v(H_k^N) = \lim_{N\to\infty}(\log d_N)^2R(H_k^N) = 0$$

for every $1\le k\le m$. Then

$$\lim_{N\to\infty}\mathrm{tr}\,p(H_1^N,\ldots,H_m^N) = \tau(p(s_1,\ldots,s_m)) \quad a.s.,$$
$$\lim_{N\to\infty}\|p(H_1^N,\ldots,H_m^N)\| = \|p(s_1,\ldots,s_m)\| \quad a.s.$$

for every noncommutative polynomial p.

The difficulty here is that we are interested in general noncommutative polynomials of random matrices, while our universality principles apply only to the linear situation of (2.1). To reduce the former to the latter, we will use classical linearization arguments that we presently recall.

Proposition 9.18 (Linearization). *Let $H_1^N,\ldots,H_m^N$ be self-adjoint random matrices and let $s_1,\ldots,s_m$ be a free semicircular family.*

a. Suppose that for every $d'\in\mathbb{N}$ and $A_0,\ldots,A_m\in \mathrm{M}_{d'}(\mathbb{C})_{\mathrm{sa}}$

$$\mathrm{sp}\big(A_0\otimes \mathbf{1}+\textstyle\sum_{k=1}^m A_k\otimes H_k^N\big)\subseteq \mathrm{sp}\big(A_0\otimes\mathbf{1}+\sum_{k=1}^m A_k\otimes s_k\big)+[-\varepsilon,\varepsilon]$$

eventually as $N\to\infty$ a.s. for all $\varepsilon>0$. Then

$$\limsup_{N\to\infty}\|p(H_1^N,\ldots,H_m^N)\|\le\|p(s_1,\ldots,s_m)\|\quad a.s.$$

for every noncommutative polynomial p.

b. Suppose that for every $d'\in\mathbb{N}$ and $A_0,\ldots,A_m\in \mathrm{M}_{d'}(\mathbb{C})_{\mathrm{sa}}$

$$\lim_{N\to\infty}\mathrm{tr}\,\Big[\big(A_0\otimes\mathbf{1}+\textstyle\sum_{k=1}^m A_k\otimes H_k^N\big)^{2r}\Big]=(\mathrm{tr}\otimes\tau)\Big[\big(A_0\otimes\mathbf{1}+\sum_{k=1}^m A_k\otimes s_k\big)^{2r}\Big]$$

a.s. for all $r\in\mathbb{N}$. Then

$$\lim_{N\to\infty}\mathrm{tr}\,p(H_1^N,\ldots,H_k^N)=\tau(p(s_1,\ldots,s_k))\quad a.s.$$

for every noncommutative polynomial p.

Proof. The first part is proved in [41, Lemma 1 and pp. 758–760]. The second part follows directly from the proof of [33, Lemma 1.1]. □

To apply the linearization argument in the present setting, we use the following.

Lemma 9.19. *Let $H_1^N,\ldots,H_m^N$ be random matrices as in Theorem 9.17, and fix $d'\in\mathbb{N}$ and $A_0,\ldots,A_m\in\mathrm{M}_{d'}(\mathbb{C})_{\mathrm{sa}}$. Define the random matrix*

$$\Xi^N=A_0\otimes\mathbf{1}+\sum_{k=1}^m A_k\otimes H_k^N,$$

and let Ξ^N_{free} be the associated noncommutative model. Then

$$\mathrm{sp}(\Xi^N)\subseteq\mathrm{sp}(\Xi^N_{\mathrm{free}})+[-\varepsilon,\varepsilon]$$

eventually as $N\to\infty$ a.s. for every $\varepsilon>0$, and for every $r\in\mathbb{N}$

$$\lim_{N\to\infty}\big|\mathrm{tr}[(\Xi^N)^{2r}]^{\frac{1}{2r}}-(\mathrm{tr}\otimes\tau)[(\Xi^N_{\mathrm{free}})^{2r}]^{\frac{1}{2r}}\big|=0\quad a.s.$$

Proof. It follows as in the proof of [9, Lemma 7.8] that $\sigma(\Xi^N)=O(1)$ and $\sigma_*(\Xi^N)\le v(\Xi^N)=o((\log d_N)^{-\frac{3}{2}})$. Moreover, it is clear from the definition of Ξ^N that

$$R(\Xi^N)=\max_{k\le m}\|A_k\|R(H_k^N)=o((\log d_N)^{-2}).$$

Applying Theorem 2.14 with $t=3\log d_N$ yields

$$\mathbf{P}\big[\mathrm{sp}(\Xi^N)\subseteq\mathrm{sp}(\Xi^N_{\mathrm{free}})+o(1)[-1,1]\big]\ge 1-\frac{2d'}{d_N^2}\ge 1-\frac{2d'}{N^2}.$$

The first conclusion follows by the Borel-Cantelli lemma.

On the other hand, Theorem 2.8 and [9, Theorem 2.7] yield for all $r\in\mathbb{N}$

$$\lim_{N\to\infty}\big|\mathbf{E}[\mathrm{tr}(\Xi^N)^{2r}]^{\frac{1}{2r}}-(\mathrm{tr}\otimes\tau)[(\Xi^N_{\mathrm{free}})^{2r}]^{\frac{1}{2r}}\big|=0.$$

In particular, $\mathbf{E}[\mathrm{tr}(\Xi^N)^{2r}]^{\frac{1}{2r}} = O(1)$ as $(\mathrm{tr}\otimes\tau)[(\Xi^N_{\mathrm{free}})^{2r}]^{\frac{1}{2r}} \le \|\Xi^N_{\mathrm{free}}\| \le 2\sigma(\Xi^N)$. To conclude, we need to show that $\mathrm{tr}[(\Xi^N)^{2r}]^{\frac{1}{2r}}$ concentrates around $\mathbf{E}[\mathrm{tr}(\Xi^N)^{2r}]^{\frac{1}{2r}}$ a.s. To this end, we apply Lemma 9.20 below with $t = 2\log d_N$ to estimate

$$\mathbf{P}\big[\big|\mathrm{tr}[(\Xi^N)^{2r}]^{\frac{1}{2r}} - \mathbf{E}[\mathrm{tr}(\Xi^N)^{2r}]^{\frac{1}{2r}}\big| \ge o(1)\big] \le \frac{1}{d_N^2} \le \frac{1}{N^2},$$

and the conclusion follows by Borel-Cantelli. □

Above we used the following concentration inequality.

Lemma 9.20. *For any random matrix as in* (2.1)*, we have*

$$\mathbf{P}\big[\big|(\mathrm{tr}\,X^{2r})^{\frac{1}{2r}} - \mathbf{E}[\mathrm{tr}\,X^{2r}]^{\frac{1}{2r}}\big| \ge C\big(\sigma_*(X) + R(X)^{\frac{1}{2}}\mathbf{E}[\mathrm{tr}\,X^{2r}]^{\frac{1}{4r}}\big)\sqrt{t} + R(X)t\big] \le e^{-t}$$

for all $t \ge r$.

Proof. We begin by writing $(\mathrm{tr}\,X^{2r})^{\frac{1}{2r}} = \sup_{f\in\mathcal{F}}\big|\sum_{i=0}^n f(Z_i)\big|$ where

$$\mathcal{F} = \{Z \mapsto x\,\mathrm{Re}\,\mathrm{tr}[MZ] + y\,\mathrm{Im}\,\mathrm{tr}[MZ] : x^2+y^2 \le 1, \|M\|_{\frac{2r}{2r-1}} \le 1\}.$$

Then

$$\sup_{f\in\mathcal{F}}\sum_{i=0}^n \mathrm{Var}(f(Z_i)) \le \sup_{\|M\|_1\le 1}\sum_{i=1}^n \mathbf{E}|\mathrm{tr}\,MZ_i|^2 = \sigma_*(X)^2$$

as $\|M\|_1 \le \|M\|_{\frac{2r}{2r-1}}$ and as the extreme points of S_1^d are rank one matrices, and

$$\sup_{f\in\mathcal{F}}\max_{0\le i\le n}\|f(Z_i) - f(Z_i')\|_\infty \le R(X)$$

where Z_i' is an independent copy of Z_i. We now apply[3] [49, Theorem 3] to estimate

$$\mathbf{P}\big[\big|(\mathrm{tr}\,X^{2r})^{\frac{1}{2r}} - \mathbf{E}[(\mathrm{tr}\,X^{2r})^{\frac{1}{2r}}]\big| \ge C\big(\sigma_*(X) + R(X)^{\frac{1}{2}}\mathbf{E}[(\mathrm{tr}\,X^{2r})^{\frac{1}{2r}}]^{\frac{1}{2}}\big)\sqrt{t} + R(X)t\big] \le e^{-t}$$

for all $t \ge 0$. Consequently

$$\big|\mathbf{E}[\mathrm{tr}\,X^{2r}]^{\frac{1}{2r}} - \mathbf{E}[(\mathrm{tr}\,X^{2r})^{\frac{1}{2r}}]\big| \lesssim \big(\sigma_*(X) + R(X)^{\frac{1}{2}}\mathbf{E}[(\mathrm{tr}\,X^{2r})^{\frac{1}{2r}}]^{\frac{1}{2}}\big)\sqrt{r} + R(X)r$$

by [21, Theorem 2.3], and we conclude by combining the above inequalities. □

We can now conclude the proof of Theorem 9.17.

Proof of Theorem 9.17. Let Ξ^N and Ξ^N_{free} be as in Lemma 9.19. It follows from the proofs of [9, Lemmas 7.9 and 7.10] that

$$\mathrm{sp}(\Xi^N_{\mathrm{free}}) \subseteq \mathrm{sp}(A_0\otimes\mathbf{1} + \textstyle\sum_{k=1}^m A_k\otimes s_k) + [-\varepsilon,\varepsilon]$$

eventually as $N\to\infty$ for every $\varepsilon > 0$, and that

$$\lim_{N\to\infty}(\mathrm{tr}\otimes\tau)[(\Xi^N_{\mathrm{free}})^{2r}] = (\mathrm{tr}\otimes\tau)[(A_0\otimes\mathbf{1} + \textstyle\sum_{k=1}^m A_k\otimes s_k)^{2r}]$$

for all $r\in\mathbb{N}$. Thus Lemma 9.19 and Proposition 9.18 yield

$$\lim_{N\to\infty}\mathrm{tr}\,p(H_1^N,\ldots,H_m^N) = \tau(p(s_1,\ldots,s_m)) \quad \text{a.s.},$$
$$\limsup_{N\to\infty}\|p(H_1^N,\ldots,H_m^N)\| \le \|p(s_1,\ldots,s_m)\| \quad \text{a.s.}$$

[3]While the statement of [49, Theorem 3] assumes that $\|f\|_\infty \le b$ for all $f\in\mathcal{F}$, only the weaker assumption $\|f(\xi_i) - f(\xi_i')\|_\infty \le 2b$ is used in the proof. We also optimized the conclusion over ε.

for every noncommutative polynomial p. To conclude, note that

$$\liminf_{N\to\infty} \|p(H_1^N,\dots,H_m^N)\| \ge \liminf_{N\to\infty} (\mathrm{tr}\,|p(H_1^N,\dots,H_m^N)|^{2r})^{\frac{1}{2r}} = \tau(|p(s_1,\dots,s_m)|^{2r})^{\frac{1}{2r}} \quad \text{a.s.}$$

for any $r \in \mathbb{N}$, where we used that $|p|^{2r}$ is also a polynomial. As

$$\lim_{r\to\infty} \tau(|p(s_1,\dots,s_m)|^{2r})^{\frac{1}{2r}} = \|p(s_1,\dots,s_m)\|$$

(here we use that τ is faithful), the conclusion follows. □

9.4.2. *Proof of Theorem 3.25.* To prove Theorem 3.25 in the general setting, we will combine Theorem 9.17 with the truncation arguments of section 8. Before we proceed to the proof, we state an elementary lemma that will be needed below.

Lemma 9.21. *Let $(Y_n)_{n\ge 1}$ be a sequence of real-valued random variables such that $|Y_n| \to 0$ a.s. as $n\to\infty$. Then there is a nonrandom sequence $(a_n)_{n\ge1}$ with $a_n \to 0$ as $n\to\infty$, such that $|Y_n| \le a_n$ eventually as $n\to\infty$ a.s.*

Proof. Let $Y_n^* := \sup_{m\ge n} |Y_m|$, and let $n_k := \inf\{n : \mathbf{P}[Y_n^* > 2^{-k}] \le 2^{-k}]\}$. Then clearly n_k is nondecreasing, and $n_k < \infty$ as we assumed $|Y_n| \to 0$ a.s. Moreover, we may assume without loss of generality that $n_k \to \infty$, as otherwise $Y_n^* = 0$ a.s. for some n and the conclusion is trivial. We may therefore define $(a_n)_{n\ge1}$ by setting $a_n = 2^{-k}$ for $n_k \le n < n_{k+1}$, $k \ge 0$. As by construction

$$\mathbf{P}[|Y_n| > a_n \text{ for some } n_k \le n < n_{k+1}] \le \mathbf{P}[Y_{n_k}^* > 2^{-k}] \le 2^{-k},$$

the conclusion follows by the Borel-Cantelli lemma. □

We can now complete the proof of Theorem 3.25.

Proof of Theorem 3.25. We first note that it suffices to prove the a.s. version of the theorem, as the in probability version follows immediately from the a.s. version using the classical fact that a sequence of random variables converges in probability if and only if every subsequence has an a.s. convergent subsequence.

We therefore assume from now on that the assumptions of the theorem hold in the a.s. sense. By Lemma 9.21, the assumptions imply that there exists a nonrandom sequence (a_N) with $a_N \to 0$ as $N \to \infty$ such that

$$\max_{1\le k\le m} \max_{1\le i\le M_N} \|Z_{ki}^N\| \le (\log d_N)^{-2} a_N$$

eventually as $N\to\infty$ a.s. Now define the truncated random matrices

$$\tilde{H}_k^N := Z_{k0}^N + \sum_{i=1}^{M_N} 1_{\|Z_{ki}^N\| \le (\log d_N)^{-2} a_N} Z_{ki}^N$$

as in section 8. Then $\tilde{H}_k^N = H_k^N$ eventually as $N\to\infty$ a.s. for all k. To complete the proof, it therefore suffices to show that $\tilde{H}_1^N,\dots,\tilde{H}_m^N$ satisfy the assumptions of Theorem 9.17. To this end, note first that by Lemma 8.6

$$(\log d_N)^2 R(\tilde{H}_k^N) \le 2a_N \xrightarrow{N\to\infty} 0.$$

Using $\mathbf{E}[|1_A Y - \mathbf{E}[1_A Y]|^2] \le \mathbf{E}[|Y|^2]$ as in the proof of Lemma 8.6, we also have

$$(\log d_N)^{\frac{3}{2}} v(\tilde{H}_k^N) \le (\log d_N)^{\frac{3}{2}} v(H_k^N) \xrightarrow{N\to\infty} 0.$$

It remains to estimate $\mathbf{E}[\tilde{H}_k^N]$ and $\mathbf{E}[(\tilde{H}_k^N)^2]$.

To bound $\mathbf{E}[\tilde{H}_k^N]$, note that it was shown in the proof of Lemma 8.4 that

$$\|\mathbf{E}[\tilde{H}_k^N] - \mathbf{E}[H_k^N]\| \le \frac{\sigma_*(H_k^N)}{\mathbf{P}[\max_j \|Z_{kj}^N\| \le (\log d_N)^{-2} a_N]^{\frac{1}{2}}}.$$

But as $\max_j \|Z_{kj}^N\| \le (\log d_N)^{-2} a_N$ eventually a.s., the denominator on the right-hand side converges to one and the numerator satisfies $\sigma_*(H_k^N) \le v(H_k^N) \to 0$. As by assumption $\|\mathbf{E}[H_k^N]\| \to 0$, we conclude that $\|\mathbf{E}[\tilde{H}_k^N]\| \to 0$ as well.

To bound $\mathbf{E}[(\tilde{H}_k^N)^2]$, note that setting $M = \mathbf{1}$ in the proof of Lemma 8.5 yields

$$\begin{aligned}&\|\mathbf{E}[(\tilde{H}_k^N - \mathbf{E}\tilde{H}_k^N)^2] - \mathbf{E}[(H_k^N - \mathbf{E}H_k^N)^2]\| \\ &\qquad \le 2\,\mathbf{P}[\max_j \|Z_{kj}^N\| > (\log d_N)^{-2} a_N]\, \sigma(H_k^N)^2 + 4\,\bar{R}(H_k^N)\sigma(H_k^N).\end{aligned}$$

The right-hand side converges to zero as $\max_j \|Z_{kj}^N\| \le (\log d_N)^{-2} a_N$ eventually a.s. and as $\bar{R}(H_k^N) \to 0$ and $\sigma(H_k^N) = O(1)$ by assumption. As $\|\mathbf{E}[\tilde{H}_k^N]\| \to 0$, $\|\mathbf{E}[H_k^N]\| \to 0$, and $\|\mathbf{E}[(H_k^N)^2] - \mathbf{1}\| \to 0$, we conclude that $\|\mathbf{E}[(\tilde{H}_k^N)^2] - \mathbf{1}\| \to 0$. □

9.4.3. *Proof of Corollary 3.28.* Before we proceed to the proof of Corollary 3.28, we first state another elementary probabilistic lemma.

Lemma 9.22. *Let $(Y_n)_{n\ge 1}$ be a sequence of i.i.d. random variables with $\mathbf{E}[|Y_n|^p] < \infty$ for some $p > 0$. Then $\lim_{n\to\infty} n^{-\frac{1}{p}} \max_{m\le n} |Y_m| = 0$ a.s.*

Proof. By the union bound and as $\sum_{k\ge 0} 2^k 1_{2^k \le x} \le 2x$, we can estimate

$$\sum_{k\ge 0} \mathbf{P}\left[2^{-\frac{k}{p}} \max_{m\le 2^k} |Y_m| \ge \varepsilon\right] \le \sum_{k\ge 0} 2^k\, \mathbf{P}\left[|Y_1|^p \ge 2^k \varepsilon^p\right] \le \frac{2\,\mathbf{E}[|Y_1|^p]}{\varepsilon^p} < \infty$$

for any $\varepsilon > 0$. Thus

$$\lim_{k\to\infty} \max_{2^{k-1}\le n < 2^k} n^{-\frac{1}{p}} \max_{m\le n} |Y_m| \le 2^{\frac{1}{p}} \lim_{k\to\infty} 2^{-\frac{k}{p}} \max_{m\le 2^k} |Y_m| = 0 \quad \text{a.s.}$$

by the Borel-Cantelli lemma. □

We can now complete the proof of Corollary 3.28.

Proof of Corollary 3.28. We prove both parts separately.

Part a. It suffices to verify that the assumptions of Theorem 3.25 are satisfied. Let $\mathrm{G}_N = ([d_N], E_N)$ be k_N-regular, and write

$$H_k^N = \sum_{i<j:\{i,j\}\in E_N} \frac{\eta_{kij}}{\sqrt{k_N}} (e_i e_j^* + e_j e_i^*).$$

Then $\mathbf{E}[H_k^N] = 0$ and $\mathbf{E}[(H_k^N)^2] = \mathbf{1}$ by construction. Furthermore, we have

$$\lim_{N\to\infty} (\log d_N)^{\frac{3}{2}} v(H_k^N) = \sqrt{2} \lim_{N\to\infty} (\log d_N)^{\frac{3}{2}} k_N^{-\frac{1}{2}} = 0$$

by the assumption of part a. On the other hand, note that

$$\mathbf{E}\left[\max_{1\le i\le M_N} \|Z_{ki}^N\|^2\right] \le \frac{1}{k_N} \mathbf{E}\left[\max_{i<j:\{i,j\}\in E_N} |\eta_{kij}|^p\right]^{\frac{2}{p}} \le \frac{(k_N d_N)^{\frac{2}{p}}}{k_N} \mathbf{E}[|\eta_{kij}|^p]^{\frac{2}{p}}.$$

As the assumption of part a. implies $(k_N d_N)^{\frac{2}{p}} k_N^{-1} \ll (\log d_N)^{-4}$, we have shown that $(\log d_N)^2 \bar{R}(H_k^N) \to 0$ as $N \to \infty$. This simultaneously verifies both remaining assumptions of the in probability version of Theorem 3.25.

If in addition E_N is increasing, we can use Lemma 9.22 to obtain

$$\limsup_{N\to\infty} (\log d_N)^2 \max_{1\le i\le M_N} \|Z_{ki}^N\| \lesssim \lim_{N\to\infty} (k_N d_N)^{-\frac{1}{p}} \max_{i<j:\{i,j\}\in E_N} |\eta_{kij}| = 0 \quad \text{a.s.},$$

where we used that $(\log d_N)^2 k_N^{-\frac{1}{2}} \lesssim (k_N d_N)^{-\frac{1}{p}}$ by the assumption of part a. The remaining conclusion of part a. then follows from the a.s. version of Theorem 3.25.

Part b. Fix any $p > 2$. Then we may choose a distribution of the entries η_{kij} such that $\mathbf{E}[\eta_{kij}] = 0$, $\mathrm{Var}(\eta_{kij}) = 1$, and $\mathbf{P}[|\eta_{kij}| > x] = (x \log x)^{-p}$ for all $x \ge x_0$ (here $x_0 > 0$ is a sufficiently large constant). As $\mathbf{E}[|\eta_{kij}|^p] < \infty$, the assumptions of Corollary 3.28 are satisfied. Now note that as $\|M\| \ge \max_{i,j} |M_{ij}|$, we have

$$\mathbf{P}\big[\|H_1^N\| > k_N^{-\frac{1}{2}} x\big] \ge \mathbf{P}\bigg[\max_{i<j:\{i,j\}\in E_N} |(H_1^N)_{ij}| > k_N^{-\frac{1}{2}} x\bigg] = 1 - \big(1 - (x\log x)^{-p}\big)^{\frac{k_N d_N}{2}}$$

for $x \ge x_0$. Choosing $x = (k_N d_N)^{\frac{1}{p}} (\log d_N)^{-1}$ yields $(x \log x)^{-p} \ge (k_N d_N)^{-1}$ for all sufficiently large N, where we used $k_N \le d_N$. We have therefore shown that

$$\mathbf{P}\big[\|H_1^N\| > k_N^{\frac{1}{p}-\frac{1}{2}} d_N^{\frac{1}{p}} (\log d_N)^{-1}\big] \ge 1 - e^{-\frac{1}{2}}$$

for all large N. But the assumption of part b. implies that $k_N^{\frac{1}{p}-\frac{1}{2}} d_N^{\frac{1}{p}} (\log d_N)^{-1} \to \infty$. This stands in contradiction to the conclusion of part a., which would imply in particular that $\|H_1^N\| \to \|s_1\| = 2$ in probability. □

9.5. **Phase transitions in spiked models.** Here we prove Theorem 3.31. We first prove convergence of the outlier eigenvalues (as in part a. of Theorem 3.30), and then consider the eigenvectors (as in part b. of Theorem 3.30).

In the following, it will be convenient to introduce the notation

$$\mathrm{B}(\theta) := \begin{cases} \theta + \frac{1}{\theta} & \text{for } \theta > 1, \\ 2 & \text{for } \theta \le 1, \end{cases}$$

and to write $\theta_1 \ge \cdots \ge \theta_s > 1 \ge \theta_{s+1} \ge \cdots \ge \theta_r > 0 =: \theta_{r+1}$. The assumptions of Theorem 3.31 will be assumed to hold without further comment.

9.5.1. *Outlier eigenvalues.* We begin by a direct application of universality.

Lemma 9.23. $\mathrm{d_H}(\mathrm{sp}(A_d + H_d), \mathrm{sp}(A_d + G_d)) \to 0$ *as* $d \to \infty$ *a.s.*

Proof. We readily compute $\sigma(A_d + H_d) = O(1)$ and $\sigma_*(A_d + H_d) = O(d^{-\frac{1}{2}})$, while by assumption $R(A_d + H_d) = o((\log d)^{-2})$. Theorem 2.5 with $t = 3 \log d$ yields

$$\mathbf{P}[\mathrm{d_H}(\mathrm{sp}(A_d + H_d), \mathrm{sp}(A_d + G_d)) > o(1)] \le \frac{1}{d^2},$$

and the conclusion follows by the Borel-Cantelli lemma. □

When combined with Theorem 3.30, Lemma 9.23 suffices to detect the presence and locations of any outlier eigenvalues. However, Hausdorff convergence is not sufficiently strong to establish convergence of individual eigenvalues. In the present setting, this stronger conclusion can however be achieved by combining universality with the min-max principle. To this end we will use the following lemma.

Lemma 9.24. *Fix any orthonormal eigenvectors* $v_{d,1}, \ldots, v_{d,r}$ *of* A_d *with eigenvalues* $\theta_1, \ldots, \theta_r$*, respectively, and let* $Q_{d,i}$ *be the projection onto* $\{v_{d,1}, \ldots, v_{d,i}\}^\perp$*. Then* $\lambda_1(Q_{d,i}(A_d + H_d)Q_{d,i}^*) \to \mathrm{B}(\theta_{i+1})$ *as* $d \to \infty$ *a.s. for* $i = 1, \ldots, r$.

Proof. The identical argument as in Lemma 9.23 yields that

$$|\lambda_1(Q_{d,i}(A_d+H_d)Q_{d,i}^*) - \lambda_1(Q_{d,i}(A_d+G_d)Q_{d,i}^*)| \xrightarrow{d\to\infty} 0 \quad \text{a.s.}$$

But $Q_{d,i}G_dQ_{d,i}^*$ is GOE of dimension $d-i$ (scaled by a factor $(\frac{d-i}{d})^{\frac{1}{2}} = 1+o(1)$). Thus applying Theorem 3.30 to $Q_{d,i}(A_d+G_d)Q_{d,i}^*$ yields the conclusion. □

We can now deduce the convergence of the eigenvalues.

Corollary 9.25. *$\lambda_i(A_d+H_d) \to \mathrm{B}(\theta_i)$ as $d\to\infty$ a.s. for $1\le i\le r+1$.*

Proof. Note first that $\lambda_i(A_d+H_d) \le \lambda_1(Q_{d,i-1}(A_d+H_d)Q_{d,i-1}^*)$ by the min-max principle. Thus $\limsup_{d\to\infty}\lambda_i(A_d+H_d) \le \mathrm{B}(\theta_i)$ a.s. by Lemma 9.24.

Next, note that the empirical spectral distribution of H_d converges a.s. to the standard semicircle distribution by Theorem 3.25. This implies in particular that $\liminf_{d\to\infty}\lambda_i(A_d+H_d) \ge \liminf_{d\to\infty}\lambda_i(H_d) \ge 2$ a.s. for all $1\le i\le r+1$. Thus we obtain $\lambda_i(A_d+H_d)\to\mathrm{B}(\theta_i)=2$ as $d\to\infty$ a.s. for $s+1\le i\le r+1$.

On the other hand, by Lemma 9.23 and Theorem 3.30, each $\mathrm{B}(\theta_i)$ with $1\le i\le s$ is a limit point of the spectrum of A_d+H_d as $d\to\infty$. If $\theta_1>\theta_2>\cdots>\theta_s$ are all distinct, this immediately implies $\liminf_{d\to\infty}\lambda_i(A_d+H_d)\ge\mathrm{B}(\theta_i)$ a.s., concluding the proof. If $\theta_1,\ldots,\theta_s$ are not distinct, we can choose $0\le A_d'\le A_d$ of rank r with distinct nonzero eigenvalues so that $\lambda_i(A_d')\ge\theta_i-\varepsilon$ for all $1\le i\le r$. Then

$$\liminf_{d\to\infty}\lambda_i(A_d+H_d) \ge \liminf_{d\to\infty}\lambda_i(A_d'+H_d) \ge \mathrm{B}(\theta_i-\varepsilon) \quad \text{a.s.},$$

and the conclusion follows as $\varepsilon>0$ is arbitrary. □

9.5.2. *Outlier eigenvectors.* A universality statement for eigenvectors of A_d+H_d can be obtained using Theorem 2.9. This yields the following conclusion.

Lemma 9.26. *Let $\varepsilon>0$ be sufficiently small that $I_i := \theta_i+\frac{1}{\theta_i}+[-\varepsilon,\varepsilon]$ are disjoint for distinct values of θ_i. Then for all $1\le i\le s$ and nonrandom $v_d\in\mathbb{C}^d$, $\|v_d\|=1$*

$$|\langle v_d, 1_{I_i}(A_d+H_d)v_d\rangle - \langle v_d, 1_{I_i}(A_d+G_d)v_d\rangle| \xrightarrow{d\to\infty} 0 \quad \textit{a.s.}$$

Proof. Let $\varphi_i:\mathbb{R}\to[0,1]$ be a smooth function so that $\varphi_i(x)=0$ for $x\notin I_i$ and $\varphi_i(x)=1$ for $x\in\theta_i+\frac{1}{\theta_i}+[-\frac{\varepsilon}{2},\frac{\varepsilon}{2}]$. Then the second part of Theorem 2.9 yields

$$|\mathbf{E}[\langle v_d,\varphi_i(A_d+H_d)v_d\rangle] - \mathbf{E}[\langle v_d,\varphi_i(A_d+G_d)v_d\rangle]| \xrightarrow{d\to\infty} 0.$$

On the other hand, applying Proposition 5.10 with $x=2C^{-1}\log d$ yields

$$\mathbf{P}\big[\big|\langle v_d,\varphi_i(A_d+H_d)v_d\rangle - \mathbf{E}[\langle v_d,\varphi_i(A_d+H_d)v_d\rangle]\big| \ge o(1)\big] \le \frac{4}{d^2},$$

and analogously for A_d+G_d. Thus the Borel-Cantelli lemma yields

$$|\langle v_d,\varphi_i(A_d+H_d)v_d\rangle - \langle v_d,\varphi_i(A_d+G_d)v_d\rangle| \xrightarrow{d\to\infty} 0 \quad \text{a.s.}$$

But $\varphi_i(A_d+H_d)=1_{I_i}(A_d+H_d)$ and $\varphi_i(A_d+G_d)=1_{I_i}(A_d+G_d)$ eventually as $d\to\infty$ a.s. by Corollary 9.25 and by part a. of Theorem 3.30, respectively. □

The desired properties of the eigenvectors now follow directly in the case that $\theta_1>\cdots>\theta_s$ are all distinct. The main difficulty is to remove the latter requirement by means of a suitable perturbation argument.

Corollary 9.27. *For any $1 \le i \le s$ and $1 \le j \le r$, we have*

$$\|P_j(A_d)v_i(A_d+H_d)\|^2 \xrightarrow[a.s.]{d\to\infty} \left(1-\frac{1}{\theta_i^2}\right)1_{\theta_j=\theta_i}.$$

Proof. If $\theta_1 > \cdots > \theta_s$ are all distinct, then $1_{I_i}(A_d+H_d) = v_i(A_d+H_d)v_i(A_d+H_d)^*$ and $1_{I_i}(A_d+G_d) = v_i(A_d+G_d)v_i(A_d+G_d)^*$ eventually as $d\to\infty$ a.s. for $1 \le i \le s$ by Corollary 9.25 and part a. of Theorem 3.30. The conclusion then follows readily by applying Lemma 9.26 and part b. of Theorem 3.30.

In the general case, let $A_d' = \sum_{m=1}^r \theta_m' \, v_m(A_d)v_m(A_d)^*$ be a perturbation of A_d (for a choice of orthonormal eigenvectors $v_m(A_d)$) with distinct $\theta_1' > \cdots > \theta_r'$ and such that $|\theta_m - \theta_m'| \le \varepsilon$ for all m. Let $J_i := \{k : \theta_k = \theta_i\}$, and let $Q_i(A_d'+H_d)$ be the projection on the linear span of $\{v_k(A_d'+H_d) : k \in J_i\}$. We claim that

$$\limsup_{d\to\infty} \left| \|P_j(A_d)\tilde v_{d,i}\|^2 - \left(1-\frac{1}{\theta_i^2}\right)1_{\theta_j=\theta_i}\right| \lesssim \varepsilon \quad \text{a.s.}$$

for any (possibly random) choice of unit vector $\tilde v_{d,i} \in \mathrm{ran}(Q_i(A_d'+H_d))$. Indeed, writing $\tilde v_{d,i} = \sum_{k\in J_i} c_{d,k} v_k(A_d'+H_d)$ with $\sum_{k\in J_i} c_{d,k}^2 = 1$, we can compute

$$\begin{aligned}\|P_j(A_d)\tilde v_{d,i}\|^2 &= \sum_{m\in J_j}\sum_{k,l\in J_i} c_{d,k}c_{d,l}\langle v_k(A_d'+H_d), v_m(A_d)\rangle\langle v_m(A_d), v_l(A_d'+H_d)\rangle \\ &= \sum_{k\in J_i} c_{d,k}^2\left(1-\frac{1}{(\theta_k')^2}\right)1_{\theta_j=\theta_i} + o(1) \quad \text{a.s.}\end{aligned}$$

as $d\to\infty$, where we used that $|\langle v_m(A_d), v_l(A_d'+H_d)\rangle|^2 \to (1-\frac{1}{(\theta_l')^2})1_{m=l}$ a.s. as A_d' has distinct eigenvalues. The claim follows as $\theta \mapsto 1-\frac{1}{\theta^2}$ is Lipschitz on $[1,\infty)$.

For any projection matrices $P, Q \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ and $c \in \mathbb{R}$, we can write

$$\sup_{v\in \mathrm{ran}\, Q, \|v\|=1} \big|\|Pv\|^2 - c\big| = \|Q(P-c\mathbf{1})Q\|.$$

Using this identity, the above claim may be rewritten as

$$\limsup_{d\to\infty} \big\|Q_i(A_d'+H_d)\big(P_j(A_d) - (1-\tfrac{1}{\theta_i^2})1_{\theta_j=\theta_i}\mathbf{1}\big)Q_i(A_d'+H_d)\big\| \lesssim \varepsilon \quad \text{a.s.}$$

On the other hand, when ε is sufficiently small, Corollary 9.25 ensures that the eigenvalues $\{\lambda_k(A_d'+H_d) : k \in J_i\}$ are separated from the rest of the spectrum of $A_d'+H_d$ by a positive gap as $d\to\infty$. A routine application of the Davis-Kahan theorem [17, Theorem VII.3.1 and Exercise VII.1.11] yields

$$\|Q_i(A_d'+H_d) - Q_i(A_d+H_d)\| \lesssim \|A_d' - A_d\| \le \varepsilon$$

eventually as $d\to\infty$ a.s. As we may choose $\varepsilon > 0$ arbitrarily small, it follows that

$$\big\|Q_i(A_d+H_d)\big(P_j(A_d) - (1-\tfrac{1}{\theta_i^2})1_{\theta_j=\theta_i}\mathbf{1}\big)Q_i(A_d+H_d)\big\| \xrightarrow{d\to\infty} 0 \quad \text{a.s.}$$

The conclusion follows as $v_i(A_d+H_d) \in \mathrm{ran}(Q_i(A_d+H_d))$. □

Combining Corollaries 9.25 and 9.27 concludes the proof of Theorem 3.31.

Acknowledgments. This work was supported in part by the NSF grants DMS-1856221 and DMS-2054565. The authors thank Noga Alon, Afonso Bandeira, March Boedihardjo, Ioana Dumitriu, Mark Rudelson, Sasha Sodin, Joel Tropp, Pierre Youssef, and Yizhe Zhu for helpful discussions on the topic of this paper.

References

[1] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures Algorithms*, 13(3-4):457–466, 1998.

[2] N. Alon and Y. Roichman. Random Cayley graphs and expanders. *Random Structures Algorithms*, 5(2):271–284, 1994.

[3] A. Amit and N. Linial. Random graph coverings. I. General theory and graph connectivity. *Combinatorica*, 22(1):1–18, 2002.

[4] G. W. Anderson. Convergence of the largest singular value of a polynomial in independent Wigner matrices. *Ann. Probab.*, 41(3B):2103–2181, 2013.

[5] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[6] B. Au. BBP phenomena for deformed random band matrices, 2023. Preprint arxiv:2304.13047.

[7] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[8] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005.

[9] A. S. Bandeira, M. T. Boedihardjo, and R. van Handel. Matrix concentration inequalities and free probability. *Invent. Math.*, 2023. To appear.

[10] A. S. Bandeira, G. Cipolloni, D. Schröder, and R. Van Handel, 2023. In preparation.

[11] A. S. Bandeira and Y. Ding. The spectral norm of random lifts of matrices. *Electron. Commun. Probab.*, 26:Paper No. 48, 10, 2021.

[12] A. S. Bandeira, D. Kunisky, D. G. Mixon, and X. Zeng. On the concentration of Gaussian Cayley matrices, 2022. Preprint arxiv:2212.00066.

[13] A. S. Bandeira and R. van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 2016.

[14] A. D. Barbour. Asymptotic expansions based on smooth functions in the central limit theorem. *Probab. Theory Relat. Fields*, 72(2):289–303, 1986.

[15] R. Bauerschmidt, A. Knowles, and H.-T. Yau. Local semicircle law for random regular graphs. *Comm. Pure Appl. Math.*, 70(10):1898–1960, 2017.

[16] F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.*, 227(1):494–521, 2011.

[17] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[18] J. Bigot and C. Male. Freeness over the diagonal and outliers detection in deformed random matrices with a variance profile. *Inf. Inference*, 10(3):863–919, 2021.

[19] C. Bordenave and B. Collins. Eigenvalues of random lifts and polynomials of random permutation matrices. *Ann. of Math. (2)*, 190(3):811–875, 2019.

[20] C. Bordenave and B. Collins. Norm of matrix-valued polynomials in random unitaries and permutations, 2023. Preprint arxiv:2304.05714v1.

[21] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[22] T. T. Cai, R. Han, and A. R. Zhang. On the non-asymptotic concentration of heteroskedastic Wishart-type matrix. *Electron. J. Probab.*, 27:Paper No. 29, 40, 2022.

[23] A.-P. Calderón. Intermediate spaces and interpolation, the complex method. *Studia Math.*, 24:113–190, 1964.

[24] M. Capitaine and C. Donati-Martin. Spectrum of deformed random matrices and free probability. In *Advanced topics in random matrices*, volume 53 of *Panor. Synthèses*, pages 151–190. Soc. Math. France, Paris, 2017.

[25] M. Capitaine, C. Donati-Martin, and D. Féral. The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *Ann. Probab.*, 37(1):1–47, 2009.

[26] E. Carlen. Trace inequalities and quantum entropy: an introductory course. In *Entropy and the quantum*, volume 529 of *Contemp. Math.*, pages 73–140. Amer. Math. Soc., Providence, RI, 2010.

[27] M. Caspers, S. Montgomery-Smith, D. Potapov, and F. Sukochev. The best constants for operator Lipschitz functions on Schatten classes. *J. Funct. Anal.*, 267(10):3557–3579, 2014.

[28] D. Chafaï and K. Tikhomirov. On the convergence of the extremal eigenvalues of empirical covariance matrices with dependence. *Probab. Theory Related Fields*, 170(3-4):847–889, 2018.
[29] S. Chatterjee. A short survey of Stein's method. In *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. IV*, pages 1–24. Kyung Moon Sa, Seoul, 2014.
[30] N. Cook. Lower bounds for the smallest singular value of structured random matrices. *Ann. Probab.*, 46(6):3442–3500, 2018.
[31] N. Cook, L. Goldstein, and T. Johnson. Size biased couplings and the spectral gap for random regular graphs. *Ann. Probab.*, 46(1):72–125, 2018.
[32] E. B. Davies. *Spectral theory and differential operators*, volume 42 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995.
[33] M. de la Salle. Complete isometries between subspaces of noncommutative L_p-spaces. *J. Operator Theory*, 64(2):265–298, 2010.
[34] I. Dumitriu, T. Johnson, S. Pal, and E. Paquette. Functional limit theorems for random regular graphs. *Probab. Theory Related Fields*, 156(3-4):921–975, 2013.
[35] I. Dumitriu and Y. Zhu. Extreme singular values of inhomogeneous sparse random rectangular matrices, 2022. Preprint arxiv:2209.12271.
[36] L. Erdős, T. Krüger, and D. Schröder. Random matrices with slow correlation decay. *Forum Math. Sigma*, 7:Paper No. e8, 89, 2019.
[37] J. Friedman. Relative expanders or weakly relatively Ramanujan graphs. *Duke Math. J.*, 118(1):19–35, 2003.
[38] J. Friedman. A proof of Alon's second eigenvalue conjecture and related problems. *Mem. Amer. Math. Soc.*, 195(910):viii+100, 2008.
[39] J. Friedman, J. Kahn, and E. Szemerédi. On the second eigenvalue of random regular graphs. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, STOC '89, page 587–598, New York, NY, USA, 1989. Association for Computing Machinery.
[40] W. T. Gowers. Quasirandom groups. *Combin. Probab. Comput.*, 17(3):363–387, 2008.
[41] U. Haagerup and S. Thorbjørnsen. A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group. *Ann. of Math. (2)*, 162(2):711–775, 2005.
[42] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bull. Amer. Math. Soc. (N.S.)*, 43(4):439–561, 2006.
[43] S. Janson. *Gaussian Hilbert spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997.
[44] E. Kowalski. *An introduction to the representation theory of groups*, volume 155 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2014.
[45] S. M. Lee and J. O. Lee. Phase transition in the generalized stochastic block model, 2022. Preprint arxiv:2206.09566.
[46] F. Lehner. Computing norms of free operators with matrix coefficients. *Amer. J. Math.*, 121(3):453–486, 1999.
[47] A. Lytova and L. Pastur. Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *Ann. Probab.*, 37(5):1778–1840, 2009.
[48] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *Ann. Probab.*, 42(3):906–945, 2014.
[49] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000.
[50] J. A. Mingo and R. Speicher. *Free probability and random matrices*, volume 35 of *Fields Institute Monographs*. Springer, New York; Fields Institute for Research in Mathematical Sciences, Toronto, ON, 2017.
[51] A. Naor. On the Banach-space-valued Azuma inequality and small-set isoperimetry of Alon-Roichman graphs. *Combin. Probab. Comput.*, 21(4):623–634, 2012.
[52] A. Nica and R. Speicher. *Lectures on the combinatorics of free probability*, volume 335 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2006.
[53] A. Nilli. On the second eigenvalue of a graph. *Discrete Math.*, 91(2):207–210, 1991.
[54] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges, 2009. Preprint arxiv:0911.0600.
[55] G. Paouris and R. Van Handel, 2023. In preparation.
[56] E. Parmaksiz and R. Van Handel, 2023. In preparation.

[57] G. Peccati and M. S. Taqqu. *Wiener chaos: moments, cumulants and diagrams*, volume 1 of *Bocconi & Springer Series*. Springer, Milan; Bocconi University Press, Milan, 2011. A survey with computer implementation, Supplementary material available online.
[58] G. Pisier. *Introduction to operator space theory*, volume 294 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2003.
[59] G. Pisier and Q. Xu. Non-commutative L^p-spaces. In *Handbook of the geometry of Banach spaces, Vol. 2*, pages 1459–1517. North-Holland, Amsterdam, 2003.
[60] M. Raab and A. Steger. "Balls into bins"—a simple and tight analysis. In *Randomization and approximation techniques in computer science (Barcelona, 1998)*, volume 1518 of *Lecture Notes in Comput. Sci.*, pages 159–170. Springer, Berlin, 1998.
[61] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.
[62] M. Rudelson and O. Zeitouni. Singular values of Gaussian matrices and permanent estimators. *Random Structures Algorithms*, 48(1):183–212, 2016.
[63] S. Sodin. The spectral edge of some random band matrices. *Ann. of Math. (2)*, 172(3):2223–2251, 2010.
[64] L. Stephan and L. Massoulié. Non-backtracking spectra of weighted inhomogeneous random graphs. *Math. Stat. Learn.*, 5(3-4):201–271, 2022.
[65] T. Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.
[66] K. Tikhomirov. On pseudospectrum of inhomogeneous non-Hermitian random matrices, 2023. Preprint arxiv:2307.08211.
[67] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8:1–230, 2015.
[68] J. A. Tropp. The expected norm of a sum of independent random matrices: an elementary approach. In *High dimensional probability VII*, volume 71 of *Progr. Probab.*, pages 173–202. Springer, [Cham], 2016.
[69] J. A. Tropp. Second-order matrix concentration inequalities. *Appl. Comput. Harmon. Anal.*, 44(3):700–736, 2018.
[70] R. van Handel. Structured random matrices. In *Convexity and concentration*, volume 161 of *IMA Vol. Math. Appl.*, pages 107–156. Springer, New York, 2017.
[71] A. Van Werde and J. Sanders. Matrix concentration inequalities with dependent summands and sharp leading-order terms, 2023. Preprint arxiv:2307.11632.
[72] R. Vershynin. Spectral norm of products of random and deterministic matrices. *Probab. Theory Related Fields*, 150(3-4):471–509, 2011.
[73] D. Voiculescu. Limit laws for random matrices and free products. *Invent. Math.*, 104(1):201–220, 1991.
[74] V. Vu. Random discrete matrices. In *Horizons of combinatorics*, volume 17 of *Bolyai Soc. Math. Stud.*, pages 257–280. Springer, Berlin, 2008.
[75] N. Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle, 2022. Preprint arxiv:2108.08198.

Fine Hall 218, Princeton University, Princeton, NJ 08544, USA
Email address: `tatianab@princeton.edu`

Fine Hall 207, Princeton University, Princeton, NJ 08544, USA
Email address: `rvan@math.princeton.edu`