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 both explains the phenomenon behind classical matrix concentration inequalities such as the matrix Bernstein inequality, and yields new *sharp* matrix concentration inequalities for general sums of independent random matrices when combined with the recent Gaussian theory of Bandeira, Boedihardjo, and Van Handel. As an application of our main results, we prove strong asymptotic freeness of a very general class of random matrix models with non-Gaussian, nonhomogeneous, and dependent entries.

1. Introduction

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 $\mathbb{C}^d$.

More general summands Z_i arise naturally in a diverse range of pure and applied mathematical problems; cf. [28] and the references therein.

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 primarily concerned with the asymptotic behavior of highly symmetric models such as matrices with i.i.d. entries or invariant ensembles [2, 27].

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

---

2010 *Mathematics Subject Classification.* 60B20; 60E15; 46L53; 46L54; 15B52.

Key words and phrases. Random matrices; matrix concentration; universality; free probability.

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 [23, §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 [21, 28] 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, the reader should observe that $(\mathbf{E}\|X\|^2)^{\frac{1}{2}} \ge \|\mathbf{E}X^2\|^{\frac{1}{2}}$ by Jensen's inequality. These 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 [4], 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, [4, 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 [4] yields much more, however: it shows that both the support and the empirical distribution of the spectrum of X is close to that of X_{free}, and that 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 [4] 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 [21, 28], or deduce (1.3) from (1.2) by a symmetrization argument as in [26, 29]. Unfortunately, neither of these approaches appears to give rise to a satisfactory extension of the theory of [4]. The methods of [4] 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 will be explained in Remark 2.14 below.

In this paper, we take an entirely different viewpoint on such problems. To motivate the form of our main results, we begin by noting that if the last term

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

in the inequality (1.3) is negligible, then (1.3) has exactly the same form as the Gaussian inequality (1.2). The central insight of this paper is that this phenomenon has nothing to do with matrix concentration inequalities themselves, but is rather a consequence of the following 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, the universality principle simultaneously explains the phenomenon behind (1.2)—(1.3), and enables us to fully extend the sharp matrix concentration theory of [4] to the model (1.1).

The universality principle was stated above in an informal manner. A detailed formulation of this principle 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 provide a detailed formulation of the resulting sharp matrix concentration inequalities that arise from [4].

As will be explained in section 3, it appears that standard methods for proving universality, such as the Lindeberg and Stein methods [11], cannot adequately capture noncommutativity of the summands in (1.1), which leads to very poor bounds. Instead, our proofs will be based on an exact expansion for the normal approximation error in terms of cumulants, combined with careful use of trace inequalities and concentration of measure. The univariate form of the cumulant expansion dates back to Barbour [5], and was first applied to random matrices with independent entries by Lytova and Pastur [17]. To the best of our knowledge, however, the only extension to dependent models appears in a recent work of Erdős et al. [13], which involves a rather complicated pre-cumulant expansion. In section 3, we formulate a direct extension of Barbour's method to the multivariate setting. While this extension itself contains no new idea as compared to [5, 17], the key innovation of our approach is that we are able to combine it with trace inequalities and concentration to prove strong universality principles for general models of the form (1.1) that admit an essentially arbitrary structure. In contrast, prior universality results such as those in [17, 13] are limited to the "classical random matrix regime" where the entry variances are of order $d^{-\frac{1}{2}}$ and exhibit decay of correlations.

Sections 4–6 are devoted to the proofs of our main results. In section 4, we develop some basic tools that are needed in the proofs. Universality of spectral statistics and of the spectrum are proved in sections 5 and 6, respectively.

In section 7 we combine our universality theorems with the theory of [4] to develop an application to free probability theory. Recall that a family of random $N\times N$ matrices $X_1^N,\ldots,X_m^N$ is said to be strongly asymptotically free [14] if

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

for every noncommutative polynomial p, where $s_1,\ldots,s_m$ is a free semicircular family (i.e., the limiting object associated to independent Wigner matrices that arises in free probablity theory). The results of this paper establish that independent random matrices of the form (1.1) are strongly asymptotically free under surprisingly mild assumptions. This will imply, for example, that $O(N \log^5 N)$ random bits in dimension N suffice to achieve strong asymptotic freeness.

Finally, we discuss in section 8 the manner in which our universality theorems may be applied to the study of sample covariance matrices. This discussion will highlight a phenomenon that is somewhat hidden in the classical matrix concentration theory: there may be different ways to express the same random matrix in terms of a model of the form (1.1), which give rise to distinct universality principles. The theorems of this paper must therefore be applied to the "correct" universal model in order to obtain meaningful results. We conclude the paper by discussing some limitations of the model (1.1) in the setting of sample covariance matrices.

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 [4, Remark 2.6] and Remark 2.14 below.

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,

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

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_{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 [20] 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 state τ. The following may be viewed as a noncommutative analogue of jointly Gaussian variables with mean μ and covariance C, cf. [20, 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

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

for all $p \geq 1$ 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. [19, Chapter 9].[2] 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 $X_{\mathrm{free}} \in \mathrm{M}_d(\mathcal{A})_{\mathrm{sa}}$ as follows:

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

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

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

[2] Throughout this paper, we denote by $\mathrm{Tr}(M)$ the unnormalized trace of a matrix $M \in \mathrm{M}_d(\mathbb{C})$, and we denote by $\mathrm{tr}(M) := \frac{1}{d}\mathrm{Tr}(M)$ the normalized trace.

for some deterministic matrices $A_1, \ldots, A_N \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ 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_{free} as

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

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_{free} agrees with the one in [4].

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:

$$\sigma(X) := \big\|\mathbf{E}[(X - \mathbf{E}X)^2]\big\|^{\frac{1}{2}}, \tag{2.2}$$

$$\sigma_*(X) := \sup_{\|v\|=\|w\|=1} \mathbf{E}\big[|\langle v, (X - \mathbf{E}X)w\rangle|^2\big]^{\frac{1}{2}}, \tag{2.3}$$

$$v(X) := \|\mathrm{Cov}(X)\|^{\frac{1}{2}}, \tag{2.4}$$

$$R(X) := \bigg\| \max_{1\le i\le n} \|Z_i\| \bigg\|_{\infty}. \tag{2.5}$$

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.
- The parameters $\sigma(X), \sigma_*(X), v(X)$ 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, $R(X)$ is specific to the non-Gaussian model.
- Recall the basic inequalities $\sigma_*(X) \le \sigma(X)$ and $\sigma_*(X) \le v(X)$, cf. [4, §2.1].

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 4.2 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 [4]. Finally, the universality principle 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.

2.2. **Universality.** We now proceed to provide a precise formulation of the universality principle. We will in fact state several universality theorems that capture different aspects of the spectrum. Their proofs will be given in sections 5–6.

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

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

Our main result is the following.

Theorem 2.4 (Spectrum universality). *For any $t \ge 0$, we have*

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

where C is a universal constant and

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

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 2.4 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 2.4 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 2.5 (Edge universality). *For any $t \ge 0$, we have*

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

as well as

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

where C is a universal constant and $\varepsilon(t)$ is as in Theorem 2.4. 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.

As a first illustration, let us revisit the matrix Bernstein inequality.

Example 2.6 (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 2.5 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 universality. Note, however, that this way of using Corollary 2.5 is very crude: in many random matrix models, the right-hand side of the bound of Corollary 2.5 is $\ll \sigma(X)$, in which case the universality principle opens the door to obtaining sharp inequalities (as in section 2.3 below).

2.2.2. *Universality of spectral statistics.* We now complement the above results by formulating universality statements for various spectral statistics. Beside their independent interest, some of these bounds admit improved scale parameters.

We begin by establishing universality of even moments. See also Theorem 5.8 and Theorem 5.10 for some further variants of the following result.

Theorem 2.7 (Moment universality). *For any $p \in \mathbb{N}$ and $2p \le q \le \infty$, we have*

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

Here we defined

$$\sigma_q(X) := \big(\mathrm{tr}\,\mathbf{E}[(X-\mathbf{E}X)^2]^{\frac{q}{2}}\big)^{\frac{1}{q}}, \qquad R_q(X) := \big(\textstyle\sum_{i=1}^n \mathbf{E}[\mathrm{tr}\,|Z_i|^q]\big)^{\frac{1}{q}}$$

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

Let us illustrate this result by another classical example.

Example 2.8 (Matrix Rosenthal). Suppose that $\mathbf{E}[X]=0$. Then the noncommutative Khintchine inequality [23, §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 2.7 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 [18, Corollary 7.4] may therefore be viewed as another consequence of the universality principle.

Moment bounds provide limited information on the spectrum of a random matrix. Much more information can be extracted from its resolvent. The following theorem establishes a universality principle for the moments of the resolvent.

Theorem 2.9 (Resolvent universality)**.** *We have*

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

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

The case of Theorem 2.9 with $p\sim\log d$ will play a central role in the proof of Theorem 2.4. On the other hand, the special case $p=1$—which yields universality of the imaginary part of the Stieltjes transform—is much simpler and admits a slightly improved matrix parameter $\sum_{i=1}^n \mathbf{E}[\operatorname{tr}|Z_i|^3] \le R(X)\sigma(X)^2$. We spell out this case separately, as it yields a simple bound for smooth spectral statistics.

Theorem 2.10 (Stieltjes transform universality)**.** *We have*

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

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

$$|\mathbf{E}[\operatorname{tr}\varphi(X)] - \mathbf{E}[\operatorname{tr}\varphi(G)]| \lesssim \|\varphi\|_{W^{5,1}(\mathbb{R})}\sum_{i=1}^n \mathbf{E}[\operatorname{tr}|Z_i|^3]$$

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

We conclude our formulation of the universality principles with two remarks.

Remark 2.11 (On sharpness of the universality principles). It is readily seen that the above universality theorems are nearly the best of their kind, as this is already the case even in the scalar setting $d=1$. For example, if $Z_0=0$, $Z_1,\ldots,Z_n$ are i.i.d. symmetric Bernoulli variables, and $p=2n$, then

$$\mathbf{E}[|X|^{2p}]^{\frac{1}{2p}} \le \|X\|_\infty = n, \qquad \mathbf{E}[|G|^{2p}]^{\frac{1}{2p}} = (1+o(1))\sqrt{\tfrac{2pn}{e}} \approx 1.21\,n$$

as $n\to\infty$, while $R(X)=1$ and $\sigma(X)=\sqrt{n}$. Thus Theorem 2.7 is optimal in this case up to the value of the universal constant. On the other hand, if $d=1$, $Z_0=0$, and $Z_1,\ldots,Z_n$ are centered Bernoulli variables with parameter $\frac{1}{n}$, then the distribution of X converges as $n\to\infty$ to the centered Poisson distribution with unit variance. As the L^p-norms of the Poisson and Gaussian distributions are of order $\frac{p}{\log p}$ and $\sqrt{p}$, respectively, this shows that the linear growth in p of Theorem 2.7

cannot be improved except possibly by a logarithmic factor. Analogous conclusions may be drawn in the setting of Corollary 2.5 by choosing Z_i to be independent diagonal matrices with i.i.d. Bernoulli variables on the diagonal.

Remark 2.12 (Unbounded models). With the exception of Theorems 2.7 and 2.10, our universality results yield nontrivial information only when $R(X) < \infty$, that is, when the summands Z_i are uniformly bounded. However, unbounded situations may often be reduced to the bounded case by routine truncation arguments, as will be illustrated in the proof of the results in section 2.4 below.

2.3. **Sharp matrix concentration inequalities.** Combining the universality theorems of the previous section with the theory of [4] yields a powerful family of sharp matrix concentration inequalities for sums of independent random matrices. The proofs of the results in this section will be given in section 6.5 below.

We begin by formulating concentration of the spectrum.

Theorem 2.13 (Sharp matrix concentration). *For any $t \geq 0$, we have*

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

where C is a universal constant and $\varepsilon(t)$ is as in Theorem 2.4. In particular,

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

and

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

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

Theorem 2.13 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 [16, Corollary 1.5] (cf. [4, §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).

Remark 2.14 ("User-friendly" bounds). One may readily deduce from Theorem 2.13 "user-friendly" matrix concentration inequalities in the spirit of [28] without any reference to free probability. For example, let Y be the $d \times d$ random matrix

$$Y = \sum_{i=1}^{n} Z_i,$$

where $Z_1, \ldots, Z_n$ are independent (not necessarily self-adjoint) $d \times d$ random matrices with $\mathbf{E}[Z_i] = 0$. Then Theorem 2.13, Remark 2.1, and [4, Lemma 2.5] yield

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

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.

It is instructive to compare these bounds with the analogous bounds that were derived from the Gaussian theory in [4, Theorem 2.12] by a symmetrization argument. As compared to the inequalities that arise from the universality principle, the symmetrization approach is unsatisfactory in several ways:

- Symmetrization loses a universal constant in the leading term of the inequalities. Consequently, the symmetrized inequalities can no longer capture the sharp behavior that is predicted by free probability.
- In the symmetrized inequalities, $\max_i \|Z_i\|$ is replaced by the typically much larger quantity $\max_i \|Z_i\|_{\mathrm{HS}}$, where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm. This inefficiency considerably limits the applicability of the resulting bounds.
- Symmetrization is only applicable to convex functionals such as $\|Y\|$, and does not provide analogous inequalities for the full spectrum as in Theorem 2.13.

For these reasons, universality principles appear to be essential in order to extend the Gaussian theory of [4] to non-Gaussian models in its strongest form.

We now complement Theorem 2.13 by formulating two-sided inequalities for various spectral statistics of the random matrix and its noncommutative model.

Theorem 2.15 (Spectral statistics)**.** *For every $p \in \mathbb{N}$, we have*

$$\big|\mathbf{E}[\mathrm{tr}\, X^{2p}]^{\frac{1}{2p}} - (\mathrm{tr} \otimes \tau)(X_{\mathrm{free}}^{2p})^{\frac{1}{2p}}\big| \lesssim v(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}p^{\frac{3}{4}} + R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}p^{\frac{2}{3}} + R(X)\,p$$

and

$$\big|\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - X|^{-2p}]^{\frac{1}{2p}} - (\mathrm{tr} \otimes \tau)(|z\mathbf{1} - X_{\mathrm{free}}|^{-2p})^{\frac{1}{2p}}\big| \lesssim \frac{v(X)^2\sigma(X)^2p^3}{(\mathrm{Im}\, z)^5} + \frac{R(X)\sigma(X)^2p^2 + R(X)^3p^3}{(\mathrm{Im}\, z)^4}$$

for $z \in \mathbb{C}$, $\mathrm{Im}\, z > 0$. Moreover,

$$|\mathbf{E}[\mathrm{tr}\, f(X)] - (\mathrm{tr} \otimes \tau)[f(X_{\mathrm{free}})]| \lesssim \{v(X)^2 + R(X)\}\sigma(X)^2\|f\|_{W^{6,1}(\mathbb{R})}$$

for every $f \in W^{6,1}(\mathbb{R})$.

2.4. **Strong asymptotic freeness.** An important consequence of the theory of [4] is strong asymptotic freeness (that is, convergence of the norm of polynomials) of very general Gaussian random matrix models. We will presently extend these results to an even more general family of non-Gaussian random matrices. Beside its independent interest, this illustrates the applicability of the results of this paper to models where the summands Z_i need not be uniformly bounded.

The proof of the following result is given in section 7; see also Corollary 7.9 for a variant that uses convergence in probability rather than a.s. Recall that a free semicircular family is a semicircular family with zero mean and unit covariance.

Theorem 2.16 (Strong asymptotic freeness). *Let $s_1, \ldots, s_m$ be a free semicircular family. For each $N \geq 1$, let Let $H_1^N, \ldots, H_m^N$ be independent self-adjoint random matrices of dimension $d_N \geq N$ 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

$$\lim_{N\to\infty} \|\mathbf{E}[H_k^N]\| = \lim_{N\to\infty} \|\mathbf{E}[(H_k^N)^2] - \mathbf{1}\| = \lim_{N\to\infty} \mathbf{E}\left[\max_{1\leq i\leq M_N} \|Z_{ki}^N\|^2\right] = 0$$

and that

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

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

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

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

for every noncommutative polynomial p.

Theorem 2.16 applies to a large family of random matrices with non-Gaussian, nonhomogeneous, and dependent entries. To illustrate its assumptions, let us consider the special case of sparse Wigner matrices. Slightly improved rates are possible if we consider convergence in probability rather than a.s., cf. Corollary 7.12.

Definition 2.17. Let $(\eta_{ij})_{1\leq i\leq j<\infty}$ 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 $1 \leq i \leq j \leq d$.

Corollary 2.18 (Sparse Wigner matrices). *Let $(\eta_{rij})_{1\leq r\leq m, 1\leq i\leq j<\infty}$ be i.i.d. real-valued random variables with zero mean and unit variance, and fix any sequence G_N of k_N-regular graphs with $d_N \geq N$ vertices. Let H_r^N be the (G_N, η_r)-sparse Wigner matrix for each N, r. Then the conclusion of Theorem 2.16 holds when:*

a. $k_N \geq c d_N^{\frac{4}{p}} (\log d_N)^4$ for some $c > 0$ if $\mathbf{E}[|\eta_{kij}|^p] < \infty$ for some $p > 4$; or
b. $k_N \gg (\log d_N)^{4+2\beta}$ if $\mathbf{E}[|\eta_{kij}|^p]^{\frac{1}{p}} \leq C p^\beta$ for some $C, \beta \geq 0$ and all $p \geq 1$.

For dense Wigner matrices $k_N = d_N$, Corollary 2.18 yields strong asymptotic freeness as soon as η_{rij} have $4+\varepsilon$ moments for some $\varepsilon > 0$, as was previously shown by Anderson [1]. On the other hand, if η_{rij} have compact support, Corollary 2.18 yields strong asymptotic freeness for very sparse matrices that have only $k_N \gg (\log d_N)^4$ nonzero entries per row or column. For example, by choosing η_{rij} to be symmetric Bernoulli variables, this shows that only $O(d \log^{4+\varepsilon} d)$ random bits in dimension d suffice to achieve strong asymptotic freeness.

More generally, Corollary 2.18 suggests that strong asymptotic freeness requires a tradeoff between sparsity and integrability of the entries. It is not difficult to show that such a tradeoff is in fact necessary, see Remark 7.13.

While we do not develop this direction systematically in this paper, our methods can also be used to obtain nonasymptotic bounds for the norms of polynomials of random matrices. The case of quadratic polynomials (that is, generalized sample covariance matrices) will be discussed further in section 8.

3. The cumulant method

The aim of this section is to introduce the basic method 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 [11, §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 2.7, classical methods yield a bound where the parameter $\sigma_q(X)^2 \le \sigma(X)^2 := \|\sum_{i=1}^n \mathbf{E} Z_i^2\|$ is replaced by $\sum_{i=1}^n \| \|Z_i\|^2 \|_\infty$, which is typically much larger.

The main 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. The advantage of this exact formula is that it will enable us to keep the summands Z_i together inside the formula, 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 [5], and has been routinely applied to the study of random matrices with independent entries since the work of Lytova and Pastur [17]. In the remainder of this section, we recall the relevant arguments of [5, 17] and spell out their (completely straightforward) extension to the multivariate case $N > 1$.

3.1. **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 $k \ge 3$ 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 [22, Proposition 3.2.1] expresses the relation between joint cumulants and moments.

Lemma 3.1 (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 [5, Lemma 1] and [17, Proposition 3.1]; the multivariate case follows precisely in the same manner.

Lemma 3.2. *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

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

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 3.1 and 3.2 reduce to

$$\mathbf{E}[W_1\cdots W_m] = \sum_{\pi\in\mathrm{P}_2([m])}\prod_{\{i,j\}\in\pi}\mathrm{Cov}(W_i,W_j) \tag{3.1}$$

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

$$\mathbf{E}[W_i f(W_1,\ldots,W_m)] = \sum_{j=1}^{m}\mathrm{Cov}(W_i,W_j)\,\mathbf{E}\bigg[\frac{\partial f}{\partial x_j}(W_1,\ldots,W_m)\bigg]. \tag{3.2}$$

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

3.2. **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 [5, 17] to the multivariate case; see, e.g., [17, Corollary 3.1].

Theorem 3.3. *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 3.2 conditionally on $\{U, (Y_k)_{k\neq i}\}$ to compute the first term in the sum, and applying (3.2) conditionally on $\{Y, (U_k)_{k\neq i}\}$ 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 3.3 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 3.3. The following result is a straightforward combination of Theorem 3.3 with Taylor expansion to order $p-1$.

Corollary 3.4. *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

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

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 3.2 conditionally on $\{U,(Y_k)_{k\ne i}\}$ 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

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

Thus the identity in the statement follows precisely as in the proof of Theorem 3.3 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. □

4. 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 4.1, 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 4.2, we develop concentration inequalities for the norm of the resolvent that will be needed to control the spectrum.

4.1. 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

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

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

Proposition 4.1. *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*

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

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

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

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

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 [30], and was previously used in [4, Lemma 4.5].)

Lemma 4.2. *Let $F : (L_\infty(S^d_\infty))^k \to \mathbb{C}$ 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 [8, §10.1] and the fact that the spaces $L_p(S^d_p)$ form a complex interpolation scale $L_r(S^d_r) = (L_p(S^d_p), L_q(S^d_q))_\theta$ with $\frac{1}{r} = \frac{1-\theta}{p} + \frac{\theta}{q}$ [24, §2]. □

Next, we recall a Hölder inequality for Schatten classes. We include a proof in order to illustrate Lemma 4.2; the same method will be used again below.

Lemma 4.3. *Let $1 \le p_1,\ldots,p_k \le \infty$ satisfy $\sum_{i=1}^k \frac{1}{p_i} = 1$. Then*

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

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

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

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

where $\Delta := \{x \in [0,1] : \sum_{i=1}^k x_i = 1\}$. By Lemma 4.2, it suffices to prove the claim only for the extreme points of Δ, that is, when $p_i = 1$ and $p_j = \infty$, $j \ne i$ for some i. But the latter case is elementary, as $|XY|^2 = Y^*X^*XY \le \|X\|^2|Y|^2$ and thus $|F(Y_1,\ldots,Y_k)| \le \|Y_{i+1}\cdots Y_k Y_1 \cdots Y_{i-1}\|_\infty \|Y_i\|_1 \le 1$. □

Finally, we recall without proof the Lieb-Thirring inequality [9, Theorem 7.4].

Lemma 4.4. *Let Y, Z be $d\times d$ positive semidefinite random matrices. Then*

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

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

We can now proceed to the proof of Proposition 4.1.

Proof of Proposition 4.1. 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 $L_\infty(S^d_\infty)$, 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 4.2, 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 4.3

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

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 $r \ge 1$

$$\begin{aligned}
&|\mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kY_k]| = |\mathbf{E}[\mathrm{tr}\,|Y_k|^{\frac12}\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kV_k|Y_k|^{\frac12}]| \\
&\le \mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\cdots\mathbf{Z}_{\frac k2}Y_{\frac k2}Y_{\frac k2}^*\mathbf{Z}_{\frac k2}\cdots Y_1^*\mathbf{Z}_1|Y_k|]^{\frac12}\cdot \\
&\qquad \mathbf{E}[\mathrm{tr}\,\mathbf{Z}_kY_{k-1}^*\mathbf{Z}_{k-1}\cdots Y_{\frac k2+1}^*\mathbf{Z}_{\frac k2+1}\mathbf{Z}_{\frac k2+1}Y_{\frac k2+1}\cdots\mathbf{Z}_{k-1}Y_{k-1}\mathbf{Z}_kV_k|Y_k|V_k^*]^{\frac12} \\
&= \mathbf{E}[\mathrm{tr}\,|\mathbf{Z}_1|^{1-\frac1r}\mathbf{U}_1^*Y_1\mathbf{Z}_2\cdots Y_{\frac k2-1}\mathbf{Z}_{\frac k2}Y_{\frac k2}\cdot \\
&\qquad\qquad Y_{\frac k2}^*\mathbf{Z}_{\frac k2}Y_{\frac k2-1}^*\cdots\mathbf{Z}_2Y_1^*\mathbf{U}_1|\mathbf{Z}_1|^{1-\frac1r}|\mathbf{Z}_1|^{\frac1r}|Y_k||\mathbf{Z}_1|^{\frac1r}]^{\frac12}\cdot \\
&\quad \mathbf{E}[\mathrm{tr}\,|\mathbf{Z}_k|^{1-\frac1r}\mathbf{U}_k^*Y_{k-1}^*\mathbf{Z}_{k-1}\cdots Y_{\frac k2+1}^*\mathbf{Z}_{\frac k2+1}\cdot \\
&\qquad\qquad \mathbf{Z}_{\frac k2+1}Y_{\frac k2+1}\cdots\mathbf{Z}_{k-1}Y_{k-1}\mathbf{U}_k|\mathbf{Z}_k|^{1-\frac1r}|\mathbf{Z}_k|^{\frac1r}V_k|Y_k|V_k^*|\mathbf{Z}_k|^{\frac1r}]^{\frac12}
\end{aligned}$$

by Cauchy-Schwarz. Now let

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

Then we have $2\frac{1-\frac1r}{q} + (k-2)\frac1q + \frac1r = 1$. We can therefore estimate

$$\begin{aligned}
&|\mathbf{E}[\mathrm{tr}\,\mathbf{Z}_1Y_1\mathbf{Z}_2Y_2\cdots\mathbf{Z}_kY_k]| \\
&\le \|\mathbf{Z}_1\|_q^{1-\frac1r}\|\mathbf{Z}_2\|_q\cdots\|\mathbf{Z}_{\frac k2}\|_q\,\||\mathbf{Z}_1|^{\frac1r}|Y_k||\mathbf{Z}_1|^{\frac1r}\|_r^{\frac12}\cdot \\
&\qquad \|\mathbf{Z}_k\|_q^{1-\frac1r}\|\mathbf{Z}_{\frac k2+1}\|_q\cdots\|\mathbf{Z}_{k-1}\|_q\,\||\mathbf{Z}_k|^{\frac1r}V_k|Y_k|V_k^*|\mathbf{Z}_k|^{\frac1r}\|_r^{\frac12} \\
&= n^{-\frac{k-2}{q-2}}R_q(X)^{\frac{(k-2)q}{q-2}}\,\||\mathbf{Z}_1|^{\frac1r}|Y_k||\mathbf{Z}_1|^{\frac1r}\|_r^{\frac12}\,\||\mathbf{Z}_k|^{\frac1r}V_k|Y_k|V_k^*|\mathbf{Z}_k|^{\frac1r}\|_r^{\frac12}
\end{aligned}$$

by Lemma 4.3, where we used that $\|\mathbf{Z}_j\|_q = n^{-\frac1q}R_q(X)$ and that $k - \frac2r = \frac{(k-2)q}{q-2}$. On the other hand, using Lemma 4.4 we obtain

$$\||\mathbf{Z}_1|^{\frac1r}|Y_k||\mathbf{Z}_1|^{\frac1r}\|_r^r \le \mathbf{E}[\mathrm{tr}\,|Y_k|^r\mathbf{Z}_1^2] = \mathbf{E}[\mathrm{tr}\,|Y_k|^r\mathbf{E}[\mathbf{Z}_1^2]] \le \|\mathbf{E}[\mathbf{Z}_1^2]\|_{\frac q2},$$

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

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

where we used that $\|\mathbf{E}[\mathbf{Z}_j^2]\|_{\frac q2} = n^{-1}\sigma_q(X)^2$ 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:

$$\begin{aligned}
&|\mathbf{E}[\operatorname{tr} \mathbf{Z}_1 Y_1 \mathbf{Z}_2 Y_2 \cdots \mathbf{Z}_k Y_k]| \\
&= |\mathbf{E}[\operatorname{tr} |Y_k|^{\frac{1}{2}} \mathbf{Z}_1 Y_1 \cdots \mathbf{Z}_{\frac{k-1}{2}} Y_{\frac{k-1}{2}} \mathbf{U}_{\frac{k+1}{2}} |\mathbf{Z}_{\frac{k+1}{2}}|^{\frac{1}{2}} \cdot \\
&\qquad\qquad |\mathbf{Z}_{\frac{k+1}{2}}|^{\frac{1}{2}} Y_{\frac{k+1}{2}} \mathbf{Z}_{\frac{k+3}{2}} \cdots Y_{k-1} \mathbf{Z}_k V_k |Y_k|^{\frac{1}{2}}]| \\
&\le \mathbf{E}[\operatorname{tr} \mathbf{Z}_1 Y_1 \cdots \mathbf{Z}_{\frac{k-1}{2}} Y_{\frac{k-1}{2}} \mathbf{U}_{\frac{k+1}{2}} |\mathbf{Z}_{\frac{k+1}{2}}| \mathbf{U}^*_{\frac{k+1}{2}} Y^*_{\frac{k-1}{2}} \mathbf{Z}_{\frac{k-1}{2}} \cdots Y_1^* \mathbf{Z}_1 |Y_k|]^{\frac{1}{2}} \cdot \\
&\qquad \mathbf{E}[\operatorname{tr} \mathbf{Z}_k Y^*_{k-1} \cdots \mathbf{Z}_{\frac{k+3}{2}} Y^*_{\frac{k+1}{2}} |\mathbf{Z}_{\frac{k+1}{2}}| Y_{\frac{k+1}{2}} \mathbf{Z}_{\frac{k+3}{2}} \cdots Y_{k-1} \mathbf{Z}_k V_k |Y_k| V_k^*]^{\frac{1}{2}}.
\end{aligned}$$

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

4.2. **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. The requisite concentration inequalities are developed in this section.

4.2.1. *The Gaussian case.* The Gaussian random matrix G is amenable to a routine application of Gaussian concentration [7, Theorem 5.6] as in [4, Lemma 6.5]. For completeness, we spell out the argument.

Lemma 4.5. *Fix $z \in \mathbb{C}$ with $\operatorname{Im} z > 0$. Then we have for any $x \ge 0$*

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

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 $\|(z\mathbf{1} - Y)^{-1}\| \le (\operatorname{Im} z)^{-1}$ for any self-adjoint matrix Y, we obtain

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

where we used that

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

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 [7, Theorem 5.6], which states than an L-Lipschitz function of a standard Gaussian vector is L^2-subgaussian. □

4.2.2. *The non-Gaussian case.* We now aim to prove an analogue of Lemma 4.5 for the non-Gaussian random matrix X. Somewhat surprisingly, this setting does not appear to be amenable to off-the-shelf concentration inequalities: while convex Lipschitz functions 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 $(Z_1,\ldots,Z_n)\mapsto\|(z\mathbf{1}-X)^{-1}\|$.

Instead, we will prove a specialized concentration inequality using the entropy method [7, Chapter 6]. The result takes a more complicated form than Lemma 4.5, but will nonetheless prove to suffice for the purposes of this paper.

Proposition 4.6. *Fix $z\in\mathbb{C}$ with $\mathrm{Im}\,z>0$. Then we have*

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

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 4.7. *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

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

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 4.8. *Let W be defined as in Lemma 4.7. 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

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

To estimate the first term, we apply [7, Theorem 11.8] to obtain

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

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

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

To estimate the second term, note that

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

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 4.9. *Let W be defined as in Lemma 4.7, and define*

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

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

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

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

We can now complete the proof of Proposition 4.6.

Proof of Proposition 4.6. We begin by noting that the self-bounding property established in Lemma 4.9 implies, by [7, Theorem 6.19], that
$$\log \mathbf{E}[e^{W/a}] \le \frac{2}{a}\,\mathbf{E}[W], \qquad a = \frac{16R(X)^2}{(\operatorname{Im} z)^4} + \frac{64R(X)^4}{(\operatorname{Im} z)^6}. \tag{4.1}$$
On the other hand, the estimate of Lemma 4.7 implies, by the exponential Poincaré inequality [7, Theorem 6.16], that
$$\log \mathbf{E}[e^{\lambda\{\|(z\mathbf{1}-X)^{-1}\| - \mathbf{E}\|(z\mathbf{1}-X)^{-1}\|\}}] \le \frac{\lambda^2 a}{1-\lambda^2 a} \log \mathbf{E}[e^{W/a}]$$
for $0 \le \lambda < a^{-\frac{1}{2}}$. Combining these estimates with a Chernoff bound [7, 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 [7, Theorem 6.16], whose proof we spell out for completeness. The last inequality of [7, Theorem 6.15] and Lemma 4.7 imply
$$\operatorname{Ent}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big] \le \lambda^2 \vartheta(\lambda b)\,\mathbf{E}\big[We^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big], \qquad b = \frac{2R(X)}{(\operatorname{Im} z)^2}$$
for $\lambda \ge 0$, where $\operatorname{Ent}(Z) := \mathbf{E}[Z\log Z] - \mathbf{E}[Z]\log \mathbf{E}[Z]$ and we used that $\vartheta(x) := \frac{e^x-1}{x}$ is a positive increasing function. In particular, as $b^2 \le a$ and $\vartheta(1) \le 2$,
$$\operatorname{Ent}\big[e^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big] \le 2\lambda^2\,\mathbf{E}\big[We^{-\lambda\|(z\mathbf{1}-X)^{-1}\|}\big]$$

for $0 \le \lambda \le a^{-\frac{1}{2}}$. Applying the duality formula of entropy as in [7, 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 $0 \le \lambda \le (2a)^{-\frac{1}{2}}$. Therefore

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

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

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

where we used $\mathrm{arctanh}(x) \le \frac{x}{1-x}$ in the second inequality and (4.1) in the last inequality. We can now apply a Chernoff bound [7, 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 4.8 to estimate $\mathbf{E}[W]$. □

5. Universality of spectral statistics

The aim of this section is to prove the universality principles for spectral statistics that were formulated in section 2.2.2. 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. This program will be implemented for the moments, resolvent, and Stieltjes transform in sections 5.2, 5.3, and 5.4, respectively. Before we do so, however, we first introduce some basic constructions that are common to all the proofs.

5.1. **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 entails no loss of generality as the results of section 2.2.2 are independent of the joint distribution of X, G), and 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 $X(0) = G$. 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 3.3 and Corollary 3.4 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 3.1. Before we can do so, we must introduce a simple construction that will facilitate working with the second identity of Lemma 3.1. 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 3.3 in the present setting.

Corollary 5.1. *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 $\iota(M):=(\operatorname{Re}M_{uv},\operatorname{Im}M_{uv})_{u,v\in[d]}$. Let $Y_i=\iota(Z_i)$, and define $Y(t)$ as in Theorem 3.3. 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 3.3 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 $M^{\mathrm{R}}_{uv}:=\operatorname{Re}M_{uv}$ and $M^{\mathrm{I}}_{uv}:=\operatorname{Im}M_{uv}$. The conclusion follows by applying the second identity of Lemma 3.1 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 3.4.

Corollary 5.2. *For any $p\geq 3$ and smooth function $f:\mathrm{M}_d(\mathbb{C})\to\mathbb{C}$ we have*

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

where the remainder term satisfies

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

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

Proof. The conclusion follows from Corollary 3.4 in exactly the same manner as we derived Corollary 5.1 from Theorem 3.3. □

5.2. **Proof of Theorem 2.7.** 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 5.3. *Let* $p \in \mathbb{N}$ *and* $B_1, \ldots, B_k \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$. *Then*

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

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

We will also need the following estimate.

Lemma 5.4. *For any* $k \in \mathbb{N}$, *we have*

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

Proof. We first crudely estimate

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

Now note that any partition of $[k]$ into m parts can be generated by first choosing $r_1, \ldots, r_m \geq 1$ 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

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

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

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

Proof. Combining Corollary 5.1 and Lemma 5.3 yields

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

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 5.3 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 4.1 to estimate

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

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

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

using Lemma 5.4, $t\leq 1$, and that the number of $(k+1)$-tuples of nonnegative integers that sum to $2p-k$ is $\binom{2p}{k} \leq \frac{(2p)^k}{k!}$. To simplify the expression, we estimate

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

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 $R^{\frac{q}{q-2}-1}\sigma^{\frac{2(q-3)}{q-2}} \leq (\max\{R,\sigma\})^2$ and $R^{\frac{(2p-2)q}{q-2}-(2p-3)-1}\sigma^{\frac{2(q-2p)}{q-2}} \leq (\max\{R,\sigma\})^2$ (as $q\geq 2p\geq 4$ 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 5.5. To this end we will use the following simple lemma.

Lemma 5.6. *Let $f:[0,1]\to\mathbb{R}_+$, $C,K\geq 0$, and $\alpha\in[0,1]$. Suppose that*

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

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

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

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 2.7.

Proof of Theorem 2.7. 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 5.6 with $\alpha = \frac{3}{2p} \in [0,1]$ and Proposition 5.5 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 $x, y \ge 0$. □

Remark 5.7 (A variant without $\sigma_q(X)$). By a slight variation of the proof, we can prove an analogue of Theorem 2.7 that does not involve the parameter $\sigma_q(X)$. The resulting bound has a worse dependence on p but a better dependence on $R_q(X)$, which yields sharper estimates when $R_q(X)\,p^2 \ll \sigma_q(X)$.

Theorem 5.8. *For any $p \in \mathbb{N}$ and $2p \le q \le \infty$, we have*

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

Proof. We begin by noting 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 $A \ge 0$ [9, 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 5.5 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

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

The conclusion follows from Lemma 5.6. □

Remark 5.9 (Subexponential inequalities). That there is considerable room in the proof of Theorem 2.7 is evident from the crude inequality in the first equation display of the proof of Lemma 5.4. This extra room can be used in some cases to obtain improved inequalities in situations where the summands Z_i may be unbounded. To illustrate this idea, let us formulate a variant of Theorem 2.7 for a matrix series model with subexponential coefficients.

Theorem 5.10. *Let $\eta_1, \ldots, \eta_n$ be independent random variables with zero mean $\mathbf{E}\eta_i = 0$, unit variance $\mathbf{E}\eta_i^2 = 1$, and κ-subexponential moments $\mathbf{E}|\eta_i|^p \le \kappa^p p!$ for $p \in \mathbb{N}$. Let $A_0, \ldots, A_n \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$, and let*

$$X = A_0 + \sum_{i=1}^n \eta_i A_i.$$

Then we have for any $p \in \mathbb{N}$ and $2p \le q \le \infty$

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

where $R_q'(X) := \big(\sum_{i=1}^n \|A_i\|_q^q\big)^{\frac{1}{q}}$ for $q < \infty$ and $R_\infty'(X) := \max_{1 \le i \le n} \|A_i\|$.

Sketch of proof. In the present setting (with $Z_i = \eta_i A_i$), the first equation display of the proof of Proposition 5.5 simplifies to

$$\frac{d}{dt}\mathbf{E}[\mathrm{tr}\, X(t)^{2p}] = \frac{1}{2}\sum_{k=3}^{2p} k t^{\frac{k}{2}-1} \sum_{\pi \in \mathrm{P}([k])} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{J \in \pi} \mathbf{E}[\eta_i^{|J|}] \times$$
$$\sum_{\substack{r_1,\ldots,r_{k+1} \ge 0 \\ r_1+\cdots+r_{k+1} = 2p-k}} \sum_{i=1}^n \mathbf{E}[\mathrm{tr}\, X(t)^{r_1} A_i X(t)^{r_2} A_i \cdots X(t)^{r_k} A_i X(t)^{r_{k+1}}].$$

The subexponential moment assumption implies

$$\sum_{\pi \in \mathrm{P}([k])} (|\pi|-1)! \prod_{J \in \pi} \mathbf{E}[\eta_i^{|J|}] \le (2\kappa)^k (k-1)!$$

by the identity in the proof of Lemma 5.4. On the other hand, note that the proof of Proposition 4.1 does not use the fact that $\mathbf{E}Z_{ij} = 0$, so that its conclusion may be applied with the choice $Z_{ij} = A_i$. The latter yields

$$\left|\sum_{i=1}^n \mathbf{E}[\mathrm{tr}\, X(t)^{r_1} A_i \cdots X(t)^{r_k} A_i X(t)^{r_{k+1}}]\right|$$
$$\le R_q'(X)^{\frac{(k-2)q}{q-2}} \sigma_q(X)^{\frac{2(q-k)}{q-2}}\, \mathbf{E}[\mathrm{tr}\, X(t)^{2p}]^{1-\frac{k}{2p}}.$$

The remainder of the proof is identical to the proof of Theorem 2.7. □

If we were to apply Theorem 2.7 to the model of Theorem 5.10, we would obtain

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

As $qp \ge p^2$ and $q^{\frac{1}{3}} p^{\frac{2}{3}} \ge p$, we see that the bound of Theorem 5.10 has an improved dependence on p compared to the bound of Theorem 2.7. On the other hand, it is not clear how to meaningfully formulate the subexponential moment assumption for general summands Z_i beyond the matrix series model.

5.3. **Proof of Theorem 2.9.** The proof of resolvent universality is similar in spirit to the the proof of moment universality. The main difference is that 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 5.2.

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

Lemma 5.11. *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*

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

In particular,

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

Proof. The identity follows by applying the product rule k times to $\mathrm{tr}\,|z\mathbf{1}-M|^{-2p} = \mathrm{tr}[R_M(z)^p R_M(\bar z)^p]$ and using that $\partial_B R_M(z) = R_M(z)BR_M(z)$. 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 $\|R_M(z)\| \le |\mathrm{Im}\,z|^{-1}$, 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 5.12. *For any $z \in \mathbb{C}$ with $\mathrm{Im}\,z>0$, $p \in \mathbb{N}$, and $t\in[0,1]$, we have*

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

Proof. Combining Corollary 5.2 with Lemma 5.11 yields

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

with

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

Here we eliminated the sum over permutations σ in the identity as in the proof of Proposition 5.5, and we used Lemma 5.4 and Hölder's inequality in the estimate of the remainder. To proceed, we apply Proposition 4.1 with $p_j = \frac{2p+k}{r_{j+1}}$ for $1 \le j < l$, $p_l = \frac{2p+k}{r_{l+1}+s_1}$, $p_j = \frac{2p+k}{s_{j-l+1}}$ for $l < j < k$, $p_k = \frac{2p+k}{s_{m+1}+r_1}$, and $q = \infty$ to estimate

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

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 5.4 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 5.11). 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 5.5. As the terms inside the sum are convex as a function of k, we can estimate

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

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 2.9 is now immediate.

Proof of Theorem 2.9. Combine Proposition 5.12 and Lemma 5.6. □

5.4. **Proof of Theorem 2.10.** The proof of Stieltjes transform universality is much simpler than that of moment or resolvent universality, as it follows readily from standard universality methods [11] without employing any cumulant expansion or nontrivial trace inequalities. For example, we can apply Corollary 5.2 with $p = 3$:

note that in this special case, the proof of Corollary 5.2 uses only Taylor expansion and no cumulants appear. We include the argument for completeness.

Proof of Theorem 2.10. As

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

we obtain

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

for any random matrices Y_1, Y_2, Y_3 using Hölder's inequality and $\|(z\mathbf{1}-M)^{-1}\| \le (\operatorname{Im} z)^{-1}$. Applying Corollary 5.2 with $p = 3$ readily yields

$$\left|\frac{d}{dt}\mathbf{E}[\operatorname{tr}(z\mathbf{1}-X(t))^{-1}]\right| \lesssim \frac{1}{(\operatorname{Im} z)^4}\sum_{i=1}^n \|Z_i\|_3^3.$$

The second part of the theorem now follows by [4, Lemma 5.11]. □

6. Universality of the spectrum

The aim of this section is to prove Theorem 2.4 and Corollary 2.5. 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 2.9 using a technique that was developed for Gaussian random matrices in [4, §6.2]. The main 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 6.1. In sections 6.2 and 6.3, 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 2.4. Finally, Corollary 2.5 will be proved in section 6.4.

6.1. **Preliminaries.** The basic principle behind te proof is the following deterministic lemma, which is a trivial modification of [4, Lemma 6.4].

Lemma 6.1. *Let $C, K_1, K_2, K_3 \ge 0$, and let $A, B \in \mathrm{M}_d(\mathbb{C})_{\mathrm{sa}}$ satisfy*

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

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 2.4 is that we will engineer the assumption of Lemma 6.1 by using Theorem 2.9 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 6.2. *We have*

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

and

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

for all $t \geq 0$, 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, $\mathrm{dist}(x, N) \leq \frac{1}{4}$ for all $x \in \mathbb{S}^{d-1}$. A routine estimate [27, p. 110] yields

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

By viewing $\mathbb{C}^d$ as a $2d$-dimensional real vector space, we may use a standard volume argument [27, Lemma 2.3.4] to choose the net N so that $|N| \leq C^d$ for a universal constant C. On the other hand, by Bernstein's inequality [7, Theorem 2.10]

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

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

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

for all $t \geq 0$, 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 $\max_i |\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|$ for self-adjoint matrices A, B [6, 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

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

(this follows from the Gaussian tail bound [7, p. 22] as the real and imaginary parts of $\langle v, (G - \mathbf{E}G)w\rangle$ are Gaussian variables with variance bounded by $\sigma_*(X)^2$). □

6.2. **Proof of Theorem 2.4: 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 6.1 holds with with high probability.

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

Lemma 6.3. *Let $z \in \mathbb{C}$ with $\mathrm{Im}\, z > 0$. Then*

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

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

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

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

By Theorem 2.9, the expectation inside the probability satisfies

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

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

To proceed, note first that Lemma 4.5 implies

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

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. [7, Theorem 2.1]). Another application of Lemma 4.5 therefore yields

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

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

Combining the above bounds yields

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

for $p \in \mathbb{N}$. The conclusion follows readily using $d^{\frac{1}{2p}} \leq e^{\frac{1}{2}}$ for $p \geq \log d$. □

We are now ready to prove one direction of Theorem 2.4.

Proposition 6.4. *For any $t \geq 0$, we have*

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

where C is a universal constant and $\varepsilon(t)$ is as defined in Theorem 2.4.

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 6.2. 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

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

for $x \ge \log d$, where we used Lemma 6.2 in the first inequality and a union bound and Lemma 6.3 in the third inequality (here $C > 1$ is the constant of Lemma 6.3).

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

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

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 6.1. □

6.3. **Proof of Theorem 2.4: 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 4.6. As before, we begin by establishing a resolvent norm inequality for a single choice of z.

Lemma 6.5. *Let $z \in \mathbb{C}$ with $\mathrm{Im}\, z > 0$. Then*

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

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

Proof. As in the proof of Lemma 6.3, 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 2.9.

To proceed, we use that Proposition 4.6 implies

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

for $p \in \mathbb{N}$ by [7, Theorem 2.3]. Another application of Proposition 4.6 yields

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

for $p \in \mathbb{N}$, provided the universal constant C is chosen sufficiently large. Now recall that the matrix Bernstein inequality [28, eq. (6.1.4)] implies

$$(\mathbf{E}\|X-\mathbf{E}X\|^2)^{\frac{1}{2}} \lesssim \sigma(X)\sqrt{\log d} + R(X)\log d \le \sigma(X)\sqrt{p} + R(X)p$$

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

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

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

Remark 6.6. 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 6.5 to estimate the matrix norms that appear in Proposition 4.6. 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 6.4 (or from Theorem 2.7 by choosing $p \asymp \log d$ and $q = \infty$) and the noncommutative Khintchine inequality.

We are now ready to prove the converse direction of Theorem 2.4.

Proposition 6.7. *For any $t \geq 0$, we have*

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

where C is a universal constant and $\varepsilon(t)$ is as defined in Theorem 2.4.

Proof. By following exactly the same steps as in the proof of Proposition 6.4, we can deduce using Lemmas 6.2 and Lemma 6.5 the inequality

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

for all $t, \varepsilon \geq 0$, where C is a universal constant. Thus Lemma 6.1 implies

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

for all $t \geq 0$ 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

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

by Young's inequality, concluding the proof. □

We now conclude the proof of Theorem 2.4.

Proof of Theorem 2.4. Combining Propositions 6.4 and 6.7 yields

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

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

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

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

6.4. Proof of Corollary 2.5.

Now that Theorem 2.4 has been established, the proof of Corollary 2.5 now follows by routine manipulations.

Proof of Corollary 2.5. We first note that

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

certainly implies

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

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

$$|\lambda_{\max}(A) - \lambda_{\max}(B)| \leq d_{\mathrm{H}}(\mathrm{sp}(A), \mathrm{sp}(B)),$$

and the first bound of Corollary 2.5 follows immediately from Theorem 2.4.

To prove the second bound, we note that a routine application of Gaussian concentration (see, e.g., [4, 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 2.5, 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 2.5 follows for a suitable choice of the universal constant (as in the last step of the proof of Theorem 2.4).

To bound the expectation, we note that

$$\begin{aligned}&|\mathbf{E}\lambda_{\max}(X) - \mathbf{E}\lambda_{\max}(G)| \leq \mathbf{E}|\lambda_{\max}(X) - \lambda_{\max}(G)| \\ &\leq C\varepsilon(\log d) + \int_{C\varepsilon(\log d)}^{\infty} \mathbf{P}\big[|\lambda_{\max}(X) - \lambda_{\max}(G)| \geq x\big]\,dx \\ &= C\varepsilon(\log d) + C\int_{\log d}^{\infty} \mathbf{P}\big[|\lambda_{\max}(X) - \lambda_{\max}(G)| \geq C\varepsilon(t)\big]\,\frac{d\varepsilon(t)}{dt}\,dt \\ &\leq C\varepsilon(\log d) + Cd\int_{\log d}^{\infty} e^{-t}\,\frac{d\varepsilon(t)}{dt}\,dt \lesssim \varepsilon(\log d),\end{aligned}$$

completing the proof. The analogous bounds where $\lambda_{\max}(X), \lambda_{\max}(G)$ are replaced by $\|X\|, \|G\|$ are proved in an identical manner. □

6.5. **Sharp matrix concentration inequalities.** The sharp matrix concentration inequalities of section 2.3 follow by a direct combination of the universality principles of this paper with the Gaussian results of [4].

Proof of Theorem 2.13. Applying [4, Theorem 2.1] to G yields

$$\mathbf{P}\big[\mathrm{sp}(G) \subseteq \mathrm{sp}(X_{\mathrm{free}}) + C\{v(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}(\log d)^{\frac{3}{4}} + \sigma_*(X)\sqrt{t}\}[-1,1]\big] \geq 1 - e^{-t}$$

for $t \geq 0$, where C is a universal constant. Thus Theorem 2.4 yields

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

and we can replace $d+1$ by d on the right-hand side by increasing the universal constant C as in the last step of the proof of Theorem 2.4. The tail bounds for $\lambda_{\max}(X)$ or $\|X\|$ follow immediately, and the bounds on $\mathbf{E}\lambda_{\max}(X)$ or $\mathbf{E}\|X\|$ follow by integrating the tail bounds as in the proof of Corollary 2.5. □

Proof of Theorem 2.15. That

$$\big|\mathbf{E}[\mathrm{tr}\, G^{2p}]^{\frac{1}{2p}} - (\mathrm{tr}\otimes\tau)(X_{\mathrm{free}}^{2p})^{\frac{1}{2p}}\big| \leq 2v(X)^{\frac{1}{2}}\sigma(X)^{\frac{1}{2}}p^{\frac{3}{4}}$$

follows from [4, Theorem 2.7], that

$$\big|\mathbf{E}[\mathrm{tr}\,|z\mathbf{1} - G|^{-2p}]^{\frac{1}{2p}} - (\mathrm{tr}\otimes\tau)(|z\mathbf{1} - X_{\mathrm{free}}|^{-2p})^{\frac{1}{2p}}\big| \lesssim \frac{v(X)^2\sigma(X)^2p^3}{(\mathrm{Im}\,z)^5}$$

follows from [4, Theorem 6.1], and that

$$|\mathbf{E}[\mathrm{tr}\, f(G)] - (\mathrm{tr}\otimes\tau)[f(X_{\mathrm{free}})]| \lesssim v(X)^2\sigma(X)^2\|f\|_{W^{6,1}(\mathbb{R})}$$

follows from [4, Corollary 2.9]. Thus the bounds of Theorem 2.15 follow by applying Theorem 2.7, Theorem 2.9, and Theorem 2.10, respectively. □

7. Strong asymptotic freeness

The aim of this section is to prove Theorem 2.16 and Corollary 2.18, which establish strong asymptotic freeness of very general random matrix models.

The main features of Theorem 2.16 already arise when the coefficient matrices Z_{ki}^N are uniformly bounded. We will therefore first prove the following weaker form of Theorem 2.16 that imposes assumptions on the parameter $R(X)$.

Theorem 7.1. *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 2.16. Suppose*

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

and

$$\lim_{N\to\infty} (\log d_N)^{\frac{3}{2}} v(H_k^N) = \lim_{N\to\infty} (\log d_N)^2 R(H_k^N) = 0$$

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

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

for every noncommutative polynomial p.

This result will be proved in two parts. First, we prove weak asymptotic freeness (convergence of the traces of polynomials) in section 7.1 using a simple form of our universality method. We subsequently prove strong asymptotic freeness (convergence of the norms of polynomials) in section 7.2.

Once Theorem 7.1 has been proved, we will extend its conclusion unbounded models in section 7.3 by means of a truncation argument, completing the proof of Theorem 2.16. Finally, Corollary 2.18 will be proved in section 7.4.

7.1. **Proof of Theorem 7.1: weak asymptotic freeness.** In this section, we assume that $H_1^N, \ldots, H_m^N$ satisfy the assumptions of Theorem 7.1, and denote by $G_1^N, \ldots, G_m^N$ independent self-adjoint random matrices whose entries are jointly Gaussian with the same mean and covariance as $H_1^N, \ldots, H_m^N$, respectively. We further assume that $H_1^N, \ldots, H_m^N$ and $G_1^N, \ldots, G_m^N$ are independent.

The aim of this section is to prove that

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

By linearity of the trace, it evidently suffices to assume that

$$p(H_1, \ldots, H_m) = H_{k_1} \cdots H_{k_q}$$

is a monomial of degree q.

Lemma 7.2. *For any $q \in \mathbb{N}$ and $1 \le k_1, \ldots, k_q \le m$, we have*

$$\lim_{N\to\infty} \big|\mathbf{E}[\operatorname{tr} H_{k_1}^N \cdots H_{k_q}^N] - \mathbf{E}[\operatorname{tr} G_{k_1}^N \cdots G_{k_q}^N]\big| = 0.$$

Proof. As the degree q is fixed as $N \to \infty$, it suffices by Theorem 3.3 to show that

$$\Delta_{k,r}^N(t) := \sum_{i=1}^{M_N} \sum_{a_1,b_1,\ldots,a_r,b_r=1}^{d_N} \kappa\big((Z_{ki}^N)_{a_1b_1}, \ldots, (Z_{ki}^N)_{a_rb_r}\big) \times$$
$$\mathbf{E}\left[\frac{\partial^r}{\partial (Z_{ki}^N)_{a_1b_1} \cdots (Z_{ki}^N)_{a_rb_r}} \operatorname{tr} H_{k_1}^N(t) \cdots H_{k_q}^N(t)\right] \xrightarrow{N\to\infty} 0$$

for all $3 \le r \le q$, $1 \le k \le m$, and $t \in [0,1]$, where

$$H_k^N(t) := \mathbf{E}[H_k^N] + \sqrt{t}\,(H_k^N - \mathbf{E}H_k^N) + \sqrt{1-t}\,(G_k^N - \mathbf{E}G_k^N).$$

Applying the second identity of Lemma 3.1 as in section 5.1, we readily obtain

$$\Delta_{k,r}^N(t) \le C_q\, R(H_k^N)^{r-2}\sigma(H_k^N)^2 \max_j \mathbf{E}[\mathrm{tr}\,|H_{k_j}^N(t)|^{q-r}]$$

using the trace inequality of Proposition 4.1, where C_q is a universal constant that depends on q only. On the other hand, note that

$$\begin{aligned}\|H_k^N(t)\|_{q-r} &\le \|\mathbf{E}[H_k^N]\|_{q-r} + \sqrt{t}\,\|H_k^N - \mathbf{E}H_k^N\|_{q-r} + \sqrt{1-t}\,\|G_k^N - \mathbf{E}G_k^N\|_{q-r} \\ &\lesssim \|\mathbf{E}[H_k^N]\| + \sigma(H_k^N)\sqrt{q} + R(H_k^N)q\end{aligned}$$

by the noncommutative Khintchine inequality and Theorem 2.7, cf. Example 2.8. The conclusion now follows directly, as the assumptions of Theorem 7.1 imply that $\|\mathbf{E}[H_k^N]\| = o(1)$, $\sigma(H_k^N) = O(1)$ and $R(H_k^N) = o(1)$ as $N \to \infty$. □

Combining Lemma 7.2 with asymptotic freeness for Gaussian random matrices as established in [4, Theorem 2.10], we immediately conclude:

Corollary 7.3. *For any $q \in \mathbb{N}$ and $1 \le k_1, \ldots, k_q \le m$, we have*

$$\lim_{N\to\infty} \mathbf{E}[\mathrm{tr}\, H_{k_1}^N \cdots H_{k_q}^N] = \tau(s_{k_1} \cdots s_{k_q}).$$

It remains to strengthen convergence in expectation to almost sure convergence. To this end, we must prove an appropriate concentration inequality. The proof of the following result is similar in spirit to that of Proposition 4.6.

Proposition 7.4. *For any $q \in \mathbb{N}$ and $1 \le k_1, \ldots, k_q \le m$, we have*

$$\mathbf{P}\Bigg[\big|\,\mathrm{tr}\, H_{k_1}^N \cdots H_{k_q}^N - \mathbf{E}[\mathrm{tr}\, H_{k_1}^N \cdots H_{k_q}^N]\,\big| \ge C_q\big(E + \sigma_* p^{\frac12} + Rp\big)^{q-1}\big\{\sigma_* p^{\frac12} + R^{\frac12}E^{\frac12}p^{\frac12} + Rp\big\}\Bigg] \le e^{-p}$$

for every $p \ge 2$, where C_q is a universal constant that depends only on q and $\sigma_* := \max_k \sigma_*(H_k^N)$, $R := \max_k R(H_k^N)$, and $E := \max_k \mathbf{E}\|H_k^N\| + \max_k \sigma(H_k^N)$.

Proof. Let $(\tilde Z_{ki}^N)$ be an independent copy of (Z_{ki}^N), and define

$$H_k^{N\sim ki} := H_k^N - Z_{ki}^N + \tilde Z_{ki}^N, \qquad H_l^{N\sim ki} = H_l^N \text{ for } l \ne k.$$

Then we have the telescoping identity

$$\begin{aligned}&H_{k_1}^N \cdots H_{k_q}^N - H_{k_1}^{N\sim ki} \cdots H_{k_q}^{N\sim ki} \\ &= \sum_{r=1}^q 1_{k_r=k}\, H_{k_1}^{N\sim ki} \cdots H_{k_{r-1}}^{N\sim ki}\{Z_{ki}^N - \tilde Z_{ki}^N\} H_{k_{r+1}}^N \cdots H_{k_q}^N \\ &= \sum_{r=1}^q 1_{k_r=k}\, H_{k_1}^N \cdots H_{k_{r-1}}^N\{Z_{ki}^N - \tilde Z_{ki}^N\} H_{k_{r+1}}^N \cdots H_{k_q}^N \\ &\quad - \sum_{r=1}^q \sum_{s=1}^{r-1} 1_{k_r=k_s=k}\, H_{k_1}^{N\sim ki} \cdots H_{k_{s-1}}^{N\sim ki}\{Z_{ki}^N - \tilde Z_{ki}^N\}\cdot \\ &\qquad\qquad H_{k_{s+1}}^N \cdots H_{k_{r-1}}^N\{Z_{ki}^N - \tilde Z_{ki}^N\} H_{k_{r+1}}^N \cdots H_{k_q}^N.\end{aligned}$$

Now note that

$$\sum_{i=1}^{M_N} \big| \operatorname{tr} H^N_{k_1} \cdots H^N_{k_{r-1}} \{ Z^N_{ki} - \tilde Z^N_{ki} \} H^N_{k_{r+1}} \cdots H^N_{k_q} \big|^2$$
$$\leq \big(\operatorname{tr} \big| H^N_{k_{r+1}} \cdots H^N_{k_q} H^N_{k_1} \cdots H^N_{k_{r-1}} \big| \big)^2 \sup_{\|v\|=\|w\|=1} \sum_{i=1}^{M_N} \big| \langle v, (Z^N_{ki} - \tilde Z^N_{ki}) w \rangle \big|^2,$$

as the extreme points of the set $\{M \in \mathrm{M}_d(\mathbb{C}) : \operatorname{Tr}|M| \leq 1\}$ are the matrices wv^* with $\|v\| = \|w\| = 1$ (by singular value decomposition). On the other hand,

$$\sum_{i=1}^{M_N} \big| \operatorname{tr} H^{N\sim ki}_{k_1} \cdots H^{N\sim ki}_{k_{s-1}} \{ Z^N_{ki} - \tilde Z^N_{ki} \} H^N_{k_{s+1}} \cdots H^N_{k_{r-1}} \{ Z^N_{ki} - \tilde Z^N_{ki} \} H^N_{k_{r+1}} \cdots H^N_{k_q} \big|^2$$
$$\leq 4R(H^N_k)^2 \max_{i,j} \|H^{N\sim ki}_{k_j}\|^{2(s-1)} \max_j \|H^N_{k_j}\|^{2(q-s-1)} \operatorname{tr} \left[\sum_{i=1}^{M_N} (Z^N_{ki} - \tilde Z^N_{ki})^2 \right]$$

by Cauchy-Schwarz. As $\|H^{N\sim ki}_{k_j}\| \leq \|H^N_{k_j}\| + 2R(H^N_k)$, we can therefore estimate

$$\sum_{i=1}^{M_N} \big| \operatorname{tr} H^N_{k_1} \cdots H^N_{k_q} - \operatorname{tr} H^{N\sim ki}_{k_1} \cdots H^{N\sim ki}_{k_q} \big|^2 \leq C_q V_k,$$

where C_q is a universal constant that depends only on q and

$$V_k := W_k^{2q-2} \sup_{\|v\|=\|w\|=1} \sum_{i=1}^{M_N} \big| \langle v, (Z^N_{ki} - \tilde Z^N_{ki}) w \rangle \big|^2 + W_k^{2q-4} R(H^N_k)^2 \operatorname{tr} \left[\sum_{i=1}^{M_N} (Z^N_{ki} - \tilde Z^N_{ki})^2 \right]$$

with $W_k := \max_j \|H^N_{k_j}\| + R(H^N_k)$. We may now apply [7, Theorem 15.5] to the real and imaginary parts of $\operatorname{tr} H^N_{k_1} \cdots H^N_{k_q}$ to estimate

$$\big\| \operatorname{tr} H^N_{k_1} \cdots H^N_{k_q} - \mathbf{E}[\operatorname{tr} H^N_{k_1} \cdots H^N_{k_q}] \big\|_p \leq C_q \sqrt{p} \left\| \sum_{j=1}^q V_{k_j} \right\|_{\frac{p}{2}}^{\frac12} \leq C_q' \sqrt{p} \max_k \|V_k\|_{\frac{p}{2}}^{\frac12}$$

for all $p \geq 2$, where C_q, C_q' are universal constants that depend only on q.

We must now estimate the right-hand side. To this end, we begin by noting that a routine application of Talagrand's concentration inequality for empirical processes (see, e.g., [4, p. 48]) yields the tail bound

$$\mathbf{P}\big[\|H^N_k - \mathbf{E}H^N_k\| \geq 2\,\mathbf{E}\|H^N_k - \mathbf{E}H^N_k\| + \sigma_*(H^N_k)\sqrt{t} + R(H^N_k)t \big] \leq e^{-Ct}$$

for all $t \geq 0$, where C is a universal constant. Thus (cf. [7, p. 30])

$$\begin{aligned}
\max_k \|W_k\|_{pq} &\leq q \max_k \|\|H^N_k\|\|_{pq} + \max_k R(H^N_k) \\
&\leq q \max_k \big\{ \|\mathbf{E}H^N_k\| + 2\,\mathbf{E}\|H^N_k - \mathbf{E}H^N_k\| + \\
&\qquad\qquad \|(\|H^N_k - \mathbf{E}H^N_k\| - 2\,\mathbf{E}\|H^N_k - \mathbf{E}H^N_k\|)_+\|_{pq} \big\} + \max_k R(H^N_k) \\
&\leq q \max_k \big\{ \|\mathbf{E}H^N_k\| + 2\,\mathbf{E}\|H^N_k - \mathbf{E}H^N_k\| + \\
&\qquad\qquad C\sigma_*(H^N_k)\sqrt{pq} + CR(H^N_k)pq \big\} + \max_k R(H^N_k) \\
&\leq C_q \max_k \big\{ \mathbf{E}\|H^N_k\| + \sigma_*(H^N_k)\sqrt{p} + R(H^N_k)p \big\}
\end{aligned}$$

for all $p \ge 2$, where C_q is a universal constant that depends only on q.

Next, we note that by the self-bounding property and expectation bound established in the proofs of Lemmas 4.9 and 4.8, respectively, we obtain

$$\begin{aligned}\Bigg\| \sup_{\|v\|=\|w\|=1} \sum_{i=1}^{M_N} \big|\langle v, (Z_{ki}^N - \tilde{Z}_{ki}^N) w\rangle\big|^2 \Bigg\|_{\frac{pq}{2}} &\lesssim 2\,\mathbf{E}\Bigg[\sup_{\|v\|=\|w\|=1} \sum_{i=1}^{M_N} \big|\langle v, (Z_{ki}^N - \tilde{Z}_{ki}^N) w\rangle\big|^2 \Bigg] + pqR(H_k^N)^2 \\ &\lesssim \sigma_*(H_k^N)^2 + R(H_k^N)\mathbf{E}\|H_k^N - \mathbf{E}H_k^N\| + pqR(H_k^N)^2\end{aligned}$$

for all $p \ge 2$, where we used [7, Corollary 15.8] in the first inequality.

Finally, as $\sum_{i=1}^{M_N} \mathrm{tr}[(Z_{ki}^N - \tilde{Z}_{ki}^N)^2]$ is a sum of independent bounded random variables, we obtain using Bernstein's inequality [7, Theorem 2.10 and Theorem 2.3]

$$\begin{aligned}\Bigg\| \mathrm{tr}\Bigg[\sum_{i=1}^{M_N} (Z_{ki}^N - \tilde{Z}_{ki}^N)^2 \Bigg] \Bigg\|_{\frac{pq}{2}} &\lesssim \sum_{i=1}^{M_N} \mathbf{E}[\mathrm{tr}\,(Z_{ki}^N - \tilde{Z}_{ki}^N)^2] + \sqrt{pq}\Bigg[\sum_{i=1}^{M_N} \mathbf{E}[(\mathrm{tr}\,(Z_{ki}^N - \tilde{Z}_{ki}^N)^2)^2] \Bigg]^{\frac{1}{2}} + pqR(H_k^N)^2 \\ &\lesssim \sigma(H_k^N)^2 + \sqrt{pq}\,R(H_k^N)\sigma(H_k^N) + pqR(H_k^N)^2 \\ &\lesssim \sigma(H_k^N)^2 + pqR(H_k^N)^2,\end{aligned}$$

for all $p \ge 2$, where we used Young's inequality in the last line.

Combining all the above estimates and using Hölder's inequality yields

$$\begin{aligned}\max_k \|V_k\|_{\frac{p}{2}}^{\frac{1}{2}} \le \max_k \|W_k\|_{pq}^{q-1} \max_k \Bigg\| \sup_{\|v\|=\|w\|=1} \sum_{i=1}^{M_N} \big|\langle v, (Z_{ki}^N - \tilde{Z}_{ki}^N) w\rangle\big|^2 \Bigg\|_{\frac{pq}{2}}^{\frac{1}{2}} \\ + \max_k \|W_k\|_{pq}^{q-2} \max_k R(H_k^N) \max_k \Bigg\| \mathrm{tr}\Bigg[\sum_{i=1}^{M_N} (Z_{ki}^N - \tilde{Z}_{ki}^N)^2 \Bigg] \Bigg\|_{\frac{pq}{2}}^{\frac{1}{2}} \\ \le C_q\big(E + \sigma_*\sqrt{p} + Rp\big)^{q-1}\big\{\sigma_* + R^{\frac{1}{2}}E^{\frac{1}{2}} + \sqrt{p}R\big\},\end{aligned}$$

were C_q is a universal constant that depends only on q and σ_*, R, E are as defined in the statement. The conclusion follows as

$$\mathbf{P}\Big[\big|\,\mathrm{tr}\,H_{k_1}^N \cdots H_{k_q}^N - \mathbf{E}[\mathrm{tr}\,H_{k_1}^N \cdots H_{k_q}^N]\,\big| \ge e\,\big\|\,\mathrm{tr}\,H_{k_1}^N \cdots H_{k_q}^N - \mathbf{E}[\mathrm{tr}\,H_{k_1}^N \cdots H_{k_q}^N]\,\big\|_p\Big] \le e^{-p}$$

for all p by Markov's inequality. □

We can now complete the proof of weak asymptotic freeness.

Corollary 7.5. *For every $q \in \mathbb{N}$ and $1 \le k_1, \ldots, k_q \le m$, we have*

$$\lim_{N\to\infty} \mathrm{tr}\,H_{k_1}^N \cdots H_{k_q}^N = \tau(s_{k_1} \cdots s_{k_q}) \quad a.s.$$

Proof. We first note that the assumptions of Theorem 7.1 imply that

$$\|\mathbf{E}H_k^N\| = o(1), \qquad \sigma(H_k^N) = 1 + o(1),$$

and that

$$\sigma_*(H_k^N) \le v(H_k^N) = o((\log d_N)^{-\frac{3}{2}}), \qquad R(H_k^N) = o((\log d_N)^{-2}).$$

Thus Theorem 2.13 yields

$$\mathbf{E}\|H_k^N\| \le 2 + o(1),$$

where we used that $\|X_{\rm free}\| \le \|\mathbf{E}X\| + 2\sigma(X)$ by [4, Lemma 2.5]. Applying Proposition 7.4 with $p = 2\log d_N$ yields

$$\mathbf{P}\Big[\big|\operatorname{tr} H_{k_1}^N \cdots H_{k_q}^N - \mathbf{E}[\operatorname{tr} H_{k_1}^N \cdots H_{k_q}^N]\big| \ge o(1)\Big] \le \frac{1}{d_N^2} \le \frac{1}{N^2}$$

as $N \to \infty$, where we used $d_N \ge N$. Thus

$$\lim_{N\to\infty} \big|\operatorname{tr} H_{k_1}^N \cdots H_{k_q}^N - \mathbf{E}[\operatorname{tr} H_{k_1}^N \cdots H_{k_q}^N]\big| = 0 \quad \text{a.s.}$$

by the Borel-Cantelli lemma. It remains to apply Corollary 7.3. □

7.2. **Proof of Theorem 7.1: strong asymptotic freeness.** In this section, we assume again that $H_1^N, \ldots, H_m^N$ satisfy the assumptions of Theorem 7.1, and aim to prove that for every noncommutative polynomial p

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

To this end, we will use a linearization argument due to Haagerup and Thorbjørnsen [14, Lemma 1] to reduce the problem to the study of certain random matrices of the form (2.1), following the same steps as in [4, §7.2].

The basic observation is the following consequence of 2.13.

Lemma 7.6. *Let $d' \in \mathbb{N}$ and $A_0, \ldots, A_m \in \mathrm{M}_{d'}(\mathbb{C})_{\rm sa}$. Define the random matrix*

$$\Xi^N = A_0 \otimes \mathbf{1} + \sum_{k=1}^m A_k \otimes H_k^N,$$

and let $\Xi_{\rm free}^N$ be the associated noncommutative model. Then

$$\operatorname{sp}(\Xi^N) \subseteq \operatorname{sp}(\Xi_{\rm free}^N) + [-\varepsilon, \varepsilon]$$

eventually as $N \to \infty$ a.s. for every $\varepsilon > 0$.

Proof. It follows as in the proof of [4, 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 2.13 with $t = 3\log d_N$ yields

$$\mathbf{P}\big[\operatorname{sp}(\Xi^N) \subseteq \operatorname{sp}(\Xi_{\rm free}^N) + o(1)[-1,1]\big] \ge 1 - \frac{d'}{d_N^2} \ge 1 - \frac{d'}{N^2},$$

and the conclusion follows by the Borel-Cantelli lemma. □

We can now complete the proof of strong asymptotic freeness.

Corollary 7.7. *For every noncommutative polynomial p, we have*

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

Proof. Let Ξ^N and $\Xi^N_{\rm free}$ be as in Lemma 7.6. It is shown in [4, Lemma 7.10] that

$$\mathrm{sp}(\Xi^N_{\rm free}) \subseteq \mathrm{sp}(A_0 \otimes \mathbf{1} + \textstyle\sum_{k=1}^m A_k \otimes s_k) + [-\varepsilon, \varepsilon]$$

eventually as $N \to \infty$ a.s. for every $\varepsilon > 0$. Combining this fact with Lemma 7.6 and the Haagerup-Thorbjørnsen linearization argument [4, Theorem 7.7] yields

$$\limsup_{N\to\infty} \|p(H_1^N, \ldots, H_m^N)\| \le \|p(s_1, \ldots, s_m)\| \quad \text{a.s.}$$

for every noncommutative polynomial p. In the converse direction, note that

$$\begin{aligned}\liminf_{N\to\infty} \|p(H_1^N, \ldots, H_m^N)\| &\ge \liminf_{N\to\infty} (\mathrm{tr}\, |p(H_1^N, \ldots, H_m^N)|^{2r})^{\frac{1}{2r}} \\ &= \tau(|p(s_1, \ldots, s_m)|^{2r})^{\frac{1}{2r}} \quad \text{a.s.}\end{aligned}$$

for any $r \in \mathbb{N}$ by Corollary 7.5, where we used that $|p|^{2r}$ is also a polynomial. As

$$\lim_{r\to\infty} \tau(|p(s_1, \ldots, s_m)|^{2r})^{\frac{1}{2r}} = \|p(s_1, \ldots, s_m)\|$$

(here we use that τ is faithful), the conclusion follows. □

Combining Corollaries 7.5 and 7.7 completes the proof of Theorem 7.1.

7.3. **Proof of Theorem 2.16.** In contrast to Theorem 7.1, which assumes that $R(H_k^N) \to 0$, the setting of Theorem 2.16 allows the summands Z_{ki}^N to be unbounded. To handle this situation, we will employ a simple truncation argument that reduces the proof of Theorem 2.16 to the setting of Theorem 7.1.

Before we proceed to the proof of Theorem 2.16, we first state an elementary probabilistic lemma that will be needed below.

Lemma 7.8. *Let $(Y_n)_{n\ge1}$ 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 2.16.

Proof of Theorem 2.16. By Lemma 7.8, 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

$$\tilde H_k^N := Z_{k0}^N + \sum_{i=1}^{M_N} 1_{\|Z_{ki}^N\| \le (\log d_N)^{-2} a_N} Z_{ki}^N.$$

Then $\tilde H_k^N = H_k^N$ eventually as $N \to \infty$ a.s. for all k. It therefore suffices to show that the matrices $\tilde H_1^N, \ldots, \tilde H_m^N$ satisfy the assumptions of Theorem 7.1.

Let us write $\tilde{H}_k^N$ in the form (2.1)

$$\tilde{H}_k^N = \tilde{Z}_{k0}^N + \sum_{i=1}^{M_N} \tilde{Z}_{ki}^N$$

with

$$\tilde{Z}_{k0}^N := Z_{k0}^N + \sum_{i=1}^{M_N} \mathbf{E}\big[1_{\|Z_{ki}^N\| \le (\log d_N)^{-2} a_N} Z_{ki}^N\big],$$
$$\tilde{Z}_{ki}^N := 1_{\|Z_{ki}^N\| \le (\log d_N)^{-2} a_N} Z_{ki}^N - \mathbf{E}\big[1_{\|Z_{ki}^N\| \le (\log d_N)^{-2} a_N} Z_{ki}^N\big]$$

for $1 \le i \le m$. Note first that by construction,

$$(\log d_N)^2 R(\tilde{H}_k^N) \le 2a_N \xrightarrow{N\to\infty} 0.$$

On the other hand, as $\mathrm{Var}(1_A Y) \le \mathbf{E}[1_A Y^2] \le \mathbf{E}[Y^2] = \mathrm{Var}(Y)$ for any centered random variable Y and event A, we can estimate

$$\mathrm{Cov}(\tilde{H}_k^N) = \sum_{i=1}^{M_N} \mathrm{Cov}(\tilde{Z}_{ki}^N) \le \sum_{i=1}^{M_N} \mathrm{Cov}(Z_{ki}^N) = \mathrm{Cov}(H_k^N).$$

Thus we 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 $\mathbf{E}[(\tilde{H}_k^N)^2]$.

Note first that

$$b_{ki}^N\, \mathbf{E}\big[1_{\|Z_{ki}^N\| \le (\log d_N)^{-2} a_N} Z_{ki}^N\big] = \mathbf{E}\big[1_{\max_j \|Z_{kj}^N\| \le (\log d_N)^{-2} a_N} Z_{ki}^N\big]$$

with

$$b_{ki}^N := \mathbf{P}\big[\max_{j\ne i} \|Z_{kj}^N\| \le (\log d_N)^{-2} a_N\big].$$

We can therefore write

$$\mathbf{E}[\tilde{H}_k^N] = \mathbf{E}[H_k^N] + \mathbf{E}\left[1_{\max_j \|Z_{kj}^N\| \le (\log d_N)^{-2} a_N} \sum_{i=1}^{M_N} \frac{Z_{ki}^N}{b_{ki}^N}\right].$$

Now note that

$$\begin{aligned}
&\left\|\mathbf{E}\left[1_{\max_j \|Z_{kj}^N\| \le (\log d_N)^{-2} a_N} \sum_{i=1}^{M_N} \frac{Z_{ki}^N}{b_{ki}^N}\right]\right\| \\
&\quad= \sup_{\|v\|=\|w\|=1} \left|\mathbf{E}\left[1_{\max_j \|Z_{kj}^N\| \le (\log d_N)^{-2} a_N} \sum_{i=1}^{M_N} \frac{\langle v, Z_{ki}^N w\rangle}{b_{ki}^N}\right]\right| \\
&\quad\le \sup_{\|v\|=\|w\|=1} \mathbf{E}\left[\sum_{i=1}^{M_N} \frac{|\langle v, Z_{ki}^N w\rangle|^2}{(b_{ki}^N)^2}\right]^{\frac{1}{2}} \le \frac{\sigma_*(H_k^N)}{\min_i b_{ki}^N}
\end{aligned}$$

using Cauchy-Schwarz. But as $\max_j \|Z_{jk}^N\| \le (\log d_N)^{-2} a_N$ eventually a.s., we have $\min_i b_{ki}^N \ge \mathbf{P}[\max_j \|Z_{jk}^N\| \le (\log d_N)^{-2} a_N] \to 1$ as $N \to \infty$. It follows that

$$\|\mathbf{E}[\tilde{H}_k^N]\| \le \|\mathbf{E}[H_k^N]\| + (1 + o(1))\sigma_*(H_k^N) \xrightarrow{N\to\infty} 0,$$

as $\|\mathbf{E}[H_k^N]\| = o(1)$ and $\sigma_*(H_k^N) \le v(H_k^N) = o(1)$ by assumption.

Next, we note that

$$\begin{aligned}\mathbf{E}[(\tilde H_k^N)^2]-\mathbf{E}[(H_k^N)^2] &= \mathbf{E}[\tilde H_k^N]^2-\mathbf{E}[H_k^N]^2\\ &\quad -\sum_{i=1}^{M_N}\mathbf{E}\big[1_{\|Z_{ki}^N\|>(\log d_N)^{-2}a_N}Z_{ki}^N\big]^2-\sum_{i=1}^{M_N}\mathbf{E}[1_{\|Z_{ki}^N\|>(\log d_N)^{-2}a_N}(Z_{ki}^N)^2],\end{aligned}$$

where we used that $\mathbf{E}\big[1_{\|Z_{ki}^N\|\le(\log d_N)^{-2}a_N}Z_{ki}^N\big]=-\mathbf{E}\big[1_{\|Z_{ki}^N\|>(\log d_N)^{-2}a_N}Z_{ki}^N\big]$ when $\mathbf{E}[Z_{ki}^N]=0$. As $\mathbf{E}[Z]^2\le\mathbf{E}[Z^2]$ for any self-adjoint random matrix Z, we obtain

$$\begin{aligned}\|\mathbf{E}[(\tilde H_k^N)^2]-\mathbf{1}\| &\le \|\mathbf{E}[(H_k^N)^2]-\mathbf{1}\|+\|\mathbf{E}[\tilde H_k^N]\|^2+\|\mathbf{E}[H_k^N]\|^2\\ &\qquad +2\Bigg\|\sum_{i=1}^{M_N}\mathbf{E}[1_{\|Z_{ki}^N\|>(\log d_N)^{-2}a_N}(Z_{ki}^N)^2]\Bigg\|\\ &\le o(1)+2\Bigg\|\mathbf{E}\Bigg[1_{\max_j\|Z_{kj}^N\|>(\log d_N)^{-2}a_N}\sum_{i=1}^{M_N}(Z_{ki}^N)^2\Bigg]\Bigg\|.\end{aligned}$$

Let $(Z_{ki}^{N\prime})$ be an independent copy of (Z_{ki}^N). Then we can estaimate

$$\begin{aligned}&\Bigg\|\mathbf{E}\Bigg[1_{\max_j\|Z_{kj}^N\|>(\log d_N)^{-2}a_N}\sum_{i=1}^{M_N}(Z_{ki}^N)^2\Bigg]\Bigg\|\\ &\quad\le\Bigg\|\mathbf{E}\Bigg[1_{\max_j\|Z_{kj}^N\|>(\log d_N)^{-2}a_N}\sum_{i=1}^{M_N}\{(Z_{ki}^N)^2-(Z_{ki}^{N\prime})^2\}\Bigg]\Bigg\|\\ &\qquad+\mathbf{P}\big[\max_j\|Z_{kj}^N\|>(\log d_N)^{-2}a_N\big]\Bigg\|\mathbf{E}\Bigg[\sum_{i=1}^{M_N}(Z_{ki}^N)^2\Bigg]\Bigg\|.\end{aligned}$$

As $\max_j\|Z_{kj}^N\|\le(\log d_N)^{-2}a_N$ eventually a.s. and $\|\mathbf{E}\sum_{i=1}^{M_N}(Z_{ki}^N)^2\|=\|\mathbf{E}[(H_k^N)^2]-\mathbf{E}[H_k^N]^2\|=O(1)$, the last term $o(1)$. To estimate the remaining term, note that

$$\begin{aligned}&\Bigg\|\mathbf{E}\Bigg[1_{\max_j\|Z_{kj}^N\|>(\log d_N)^{-2}a_N}\sum_{i=1}^{M_N}\{(Z_{ki}^N)^2-(Z_{ki}^{N\prime})^2\}\Bigg]\Bigg\|\\ &=\sup_{\|v\|=\|w\|=1}\Bigg|\mathbf{E}\Bigg[1_{\max_j\|Z_{kj}^N\|>(\log d_N)^{-2}a_N}\sum_{i=1}^{M_N}\langle v,\{(Z_{ki}^N)^2-(Z_{ki}^{N\prime})^2\}w\rangle\Bigg]\Bigg|\\ &\le\sup_{\|v\|=\|w\|=1}\mathbf{E}\Bigg[1_{\max_j\{\|Z_{kj}^N\|\vee\|Z_{kj}^{N\prime}\|\}>(\log d_N)^{-2}a_N}\Bigg|\sum_{i=1}^{M_N}\langle v,\{(Z_{ki}^N)^2-(Z_{ki}^{N\prime})^2\}w\rangle\Bigg|\Bigg]\\ &=\sup_{\|v\|=\|w\|=1}\mathbf{E}\Bigg[1_{\max_j\{\|Z_{kj}^N\|\vee\|Z_{kj}^{N\prime}\|\}>(\log d_N)^{-2}a_N}\Bigg|\sum_{i=1}^{M_N}\varepsilon_i\langle v,\{(Z_{ki}^N)^2-(Z_{ki}^{N\prime})^2\}w\rangle\Bigg|\Bigg],\end{aligned}$$

where (ε_i) are i.i.d. random signs independent of $(Z_{ki}^N,Z_{ki}^{N\prime})$, and we used in the last line that the distribution of $(Z_{ki}^N,Z_{ki}^{N\prime})$ is invariant under exchanging Z_{kj}^N and

$Z_{ki}^{N\prime}$ for any j. But note that by Jensen and the triangle inequality

$$\mathbf{E}_\varepsilon\left[\left|\sum_{i=1}^{M_N}\varepsilon_i\langle v,\{(Z_{ki}^N)^2-(Z_{ki}^{N\prime})^2\}w\rangle\right|\right]\le\left(\sum_{i=1}^{M_N}|\langle v,\{(Z_{ki}^N)^2-(Z_{ki}^{N\prime})^2\}w\rangle|^2\right)^{\frac12}$$
$$\le\left(\sum_{i=1}^{M_N}\|Z_{ki}^N w\|^2\right)^{\frac12}\max_j\|Z_{kj}^N\|+\left(\sum_{i=1}^{M_N}\|Z_{ki}^{N\prime}w\|^2\right)^{\frac12}\max_j\|Z_{kj}^{N\prime}\|,$$

where $\mathbf{E}_\varepsilon$ denotes the expectation with respect to the variables (ε_i) only. Thus

$$\left\|\mathbf{E}\left[1_{\max_j\|Z_{kj}^N\|>(\log d_N)^{-2}a_N}\sum_{i=1}^{M_N}\{(Z_{ki}^N)^2-(Z_{ki}^{N\prime})^2\}\right]\right\|$$
$$\le 2\,\mathbf{E}\left[\max_{1\le j\le M_N}\|Z_{kj}^N\|^2\right]^{\frac12}\|\mathbf{E}[(H_k^N)^2]-\mathbf{E}[H_k^N]^2\|=o(1)$$

by Cauchy-Schwarz and the assumptions of the theorem. It follows that

$$\|\mathbf{E}[(\tilde H_k^N)^2]-\mathbf{1}\|\xrightarrow{N\to\infty}0.$$

Combining all the above estimates, we have now verified that $\tilde H_1^N,\ldots,\tilde H_m^N$ satisfy all the assumptions of Theorem 7.1, completing the proof. □

Theorem 2.16 remains valid if almost sure convergence is replaced by convergence in probability. While this yields a slightly weaker conclusion, the assumption is often easier to check and holds in more general situations.

Corollary 7.9. *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 2.16. Suppose*

$$\lim_{N\to\infty}\|\mathbf{E}[H_k^N]\|=\lim_{N\to\infty}\|\mathbf{E}[(H_k^N)^2]-\mathbf{1}\|=\lim_{N\to\infty}\mathbf{E}\left[\max_{1\le i\le M_N}\|Z_{ki}^N\|^2\right]=0$$

and that

$$\lim_{N\to\infty}(\log d_N)^{\frac32}v(H_k^N)=0,\qquad\lim_{N\to\infty}(\log d_N)^2\max_{1\le i\le M_N}\|Z_{ki}^N\|=0\ \textit{in probability}$$

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

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

for every noncommutative polynomial p.

Proof. This follows immedately from Theorem 2.16 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. □

7.4. **Proof of Corollary 2.18.** To prove Corollary 2.18, we need only verify that the assumptions of Theorem 2.16 are satisfied in this setting. Before we proceed to the proof, we first state another elementary probabilistic lemma.

Lemma 7.10. *Let $(Y_n)_{n\ge1}$ be a sequence of i.i.d. random variables with $\mathbf{E}[|Y_n|^p]<\infty$ for some $p>0$. Then $\lim_{n\to\infty}n^{-\frac1p}\max_{m\le n}|Y_m|=0$ a.s.*

Proof. By the union bound and as $\sum_{k\geq 0} 2^k 1_{2^k\leq x} \leq 2x$, we can estimate

$$\sum_{k\geq 0}\mathbf{P}\left[2^{-\frac{k}{p}}\max_{m\leq 2^k}|Y_m|\geq\varepsilon\right]\leq\sum_{k\geq 0}2^k\,\mathbf{P}\left[|Y_1|^p\geq 2^k\varepsilon^p\right]\leq\frac{2\,\mathbf{E}[|Y_1|^p]}{\varepsilon^p}<\infty$$

for any $\varepsilon>0$. Thus

$$\lim_{k\to\infty}\max_{2^{k-1}\leq n<2^k} n^{-\frac{1}{p}}\max_{m\leq n}|Y_m|\leq 2^{\frac{1}{p}}\lim_{k\to\infty}2^{-\frac{k}{p}}\max_{m\leq 2^k}|Y_m|=0\quad\text{a.s.}$$

by the Borel-Cantelli lemma. □

We can now complete the proof of Corollary 2.18.

Proof of Corollary 2.18. Let $\mathrm{G}_N=([d_N],E_N)$ be k_N-regular, and write

$$H_k^N=\sum_{i<j:\{i,j\}\in E_N}\frac{\eta_{kij}}{\sqrt{k_N}}(e_ie_j^*+e_je_i^*).$$

Then $\mathbf{E}[H_k^N]=0$ and $\mathbf{E}[(H_k^N)^2]=\mathbf{1}$ by construction. Furthermore, we have

$$\lim_{N\to\infty}(\log d_N)^{\frac{3}{2}}v(H_k^N)=\sqrt{2}\lim_{N\to\infty}(\log d_N)^{\frac{3}{2}}k_N^{-\frac{1}{2}}=0$$

under the assumptions of either parts a or b.

It remains to verify the requisite assumptions on

$$\max_{1\leq i\leq M_N}\|Z_{ki}^N\|=k_N^{-\frac{1}{2}}\max_{i<j:\{i,j\}\in E_N}|\eta_{kij}|\leq k_N^{-\frac{1}{2}}\max_{1\leq i<j\leq d_N}|\eta_{kij}|.$$

We consider parts a and b separately.

Part a. Let η have the same distribution as η_{kij}. Then we can estimate

$$\mathbf{E}\left[\max_{1\leq i\leq M_N}\|Z_{ki}^N\|^2\right]\leq\frac{1}{k_N}\mathbf{E}\left[\max_{1\leq i<j\leq d_N}|\eta_{kij}|^p\right]^{\frac{2}{p}}\leq\frac{d_N^{\frac{4}{p}}}{k_N}\mathbf{E}[|\eta|^p]^{\frac{2}{p}}.$$

As the assumption of part a implies $d_N^{\frac{4}{p}}k_N^{-1}\lesssim(\log d_N)^{-4}$, we have shown

$$\lim_{N\to\infty}\mathbf{E}\left[\max_{1\leq i\leq M_N}\|Z_{ki}^N\|^2\right]=0.$$

On the other hand, Lemma 7.10 yields

$$\limsup_{N\to\infty}(\log d_N)^2\max_{1\leq i\leq M_N}\|Z_{ki}^N\|\lesssim\lim_{N\to\infty}d_N^{-\frac{2}{p}}\max_{1\leq i<j\leq d_N}|\eta_{kij}|=0\quad\text{a.s.},$$

where we used that $(\log d_N)^2k_N^{-\frac{1}{2}}\lesssim d_N^{-\frac{2}{p}}$ by the assumption of part a. Thus all the assumptions of Theorem 2.16 are satisfied, completing the proof.

Part b. As in the first part, we estimate

$$\mathbf{E}\left[\max_{1\leq i\leq M_N}\|Z_{ki}^N\|^2\right]\leq\frac{d_N^{\frac{4}{p}}}{k_N}\mathbf{E}[|\eta|^p]^{\frac{2}{p}}\leq\frac{C^2p^{2\beta}d_N^{\frac{4}{p}}}{k_N}$$

for all $p\geq 1$. If we choose $p=\log d_N$, we obtain

$$\lim_{N\to\infty}\mathbf{E}\left[\max_{1\leq i\leq M_N}\|Z_{ki}^N\|^2\right]=0$$

as the assumption of part b implies $(\log d_N)^{2\beta}k_N^{-1}\ll(\log d_N)^{-4}$.

To establish almost sure convergence, note that

$$\mathbf{P}\bigg[\max_{1\le i<j\le d_N}|\eta_{kij}|\ge Ce(4\log d_N)^\beta\bigg]\le d_N^2\,\mathbf{P}\big[|\eta|\ge e\|\eta\|_{4\log d_N}\big]\le\frac{1}{d_N^2}\le\frac{1}{N^2},$$

where we used the union bound in the first inequality and Markov's inequality in the second inequality. Thus the Borel-Cantelli lemma implies that

$$(\log d_N)^2\max_{1\le i\le M_N}\|Z_{ki}^N\|\le(\log d_N)^2k_N^{-\frac12}\max_{1\le i<j\le d_N}|\eta_{kij}|\lesssim\frac{(\log d_N)^{2+\beta}}{k_N^{\frac12}}$$

eventually as $N\to\infty$ a.s. But as $(\log d_N)^{2+\beta}\ll k_N^{\frac12}$ by the assumption of part b, all the assumptions of Theorem 2.16 are satisfied, completing the proof. □

Remark 7.11 (Sharper rate). An inefficiency in the proof of Corollary 2.18 is that we estimated the maximum of $O(k_Nd_N)$ variables $\max_{i<j:\{i,j\}\in E_N}|\eta_{kij}|$ by the maximum of $O(d_N^2)$ variables $\max_{i<j\le d_N}|\eta_{kij}|$. This is necessary in order to apply Lemma 7.10, as the set of variables $(\eta_{kij})_{i<j:\{i,j\}\in E_N}$ that are used by H_k^N need not be increasing in d_N. When this is the case, however, or if we weaken a.s. convergence to convergence in probability, we obtain a better rate in part a of Corollary 2.18. This rate is nearly optimal, as will be explained in Remark 7.13 below.

Corollary 7.12. *Let $H_1^N,\ldots,H_m^N$ be (G_N,η_r)-sparse Wigner matrices as in Corollary 2.18, and suppose that $\mathbf{E}[|\eta_{kij}|^p]<\infty$. If $k_N\gg d_N^{\frac{2}{p-2}}(\log d_N)^{\frac{4p}{p-2}}$, then*

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

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

Proof. For the first part of the statement, it suffices to verify the assumptions of Corollary 7.9. The assumptions on $\mathbf{E}[H_k^N]$, $\mathbf{E}[(H_k^N)^2]$, and $v(H_k^N)$ were already verified in the proof of Corollary 2.18. On the other hand, we can now estimate

$$\mathbf{E}\bigg[\max_{1\le i\le M_N}\|Z_{ki}^N\|^2\bigg]\le\frac{1}{k_N}\mathbf{E}\bigg[\max_{i<j:\{i,j\}\in E_N}|\eta_{kij}|^p\bigg]^{\frac2p}\le\frac{(k_Nd_N)^{\frac2p}}{k_N}\mathbf{E}[|\eta|^p]^{\frac2p}.$$

As the present assumption implies $(k_Nd_N)^{\frac2p}k_N^{-1}\ll(\log d_N)^{-4}$, we have shown

$$\lim_{N\to\infty}(\log d_N)^4\,\mathbf{E}\bigg[\max_{1\le i\le M_N}\|Z_{ki}^N\|^2\bigg]=0,$$

which simultaneously verifies both remaining assumptions of Corollary 7.9.

Now note that if E_N is increasing, we can use Lemma 7.10 to obtain

$$\limsup_{N\to\infty}(\log d_N)^2\max_{1\le i\le M_N}\|Z_{ki}^N\|\lesssim\lim_{N\to\infty}(k_Nd_N)^{-\frac1p}\max_{i<j:\{i,j\}\in E_N}|\eta_{kij}|=0\quad\text{a.s.},$$

where we used that $(\log d_N)^2k_N^{-\frac12}\lesssim(k_Nd_N)^{-\frac1p}$ under the present assumption. The second part of the statement therefore follows from Theorem 2.16. □

Remark 7.13 (Lower bounds). A key feature of Corollaries 2.18 and 7.12 is the tradeoff between sparsity and integrability of the entries. In general, such a tradeoff is not only sufficient, but also necessary, for the strong asymptotic freeness property to hold. Let us illustrate this in a concrete example.

As above, we denote by η a random variable with the same distribution as η_{kij}. Fix $p \geq 4$ and choose an entry distribution such that $\mathbf{P}[|\eta| > x] = (x \log x)^{-p}$ for $x \geq e$, so that $\mathbf{E}[|\eta|^p] < \infty$ but $\mathbf{E}[|\eta|^{p+\delta}] = \infty$ for any $\delta > 0$. Corollary 7.12 implies

$$k_N \gg d_N^{\frac{2}{p-2}} (\log d_N)^{\frac{4p}{p-2}} \implies \lim_{N\to\infty} \|p(H_1^N, \ldots, H_m^N)\| = \|p(s_1, \ldots, s_m)\| \text{ in prob.}$$

We will presently show that

$$\lim_{N\to\infty} \|p(H_1^N, \ldots, H_m^N)\| = \|p(s_1, \ldots, s_m)\| \text{ in prob.} \implies k_N \gg d_N^{\frac{2}{p-2}} (\log d_N)^{-\frac{4p}{p-2}}$$

as $N \to \infty$. In particular, this shows that the polynomial rate of Corollary 7.12 is the best possible up to the logarithmic factor.

To establish the claim, note that as $\|M\| \geq \max_{i,j} |M_{ij}|$, we can estimate

$$\mathbf{P}\big[\|H_1^N\| > k_N^{-\frac{1}{2}} x\big] \geq \mathbf{P}\Bigg[\max_{i<j:\{i,j\}\in E_N} |(H_1^N)_{ij}| > k_N^{-\frac{1}{2}} x\Bigg] = 1 - \big(1 - (x \log x)^{-p}\big)^{\frac{k_N d_N}{2}}$$

for $x \geq e$. Choosing $x = (k_N d_N)^{\frac{1}{p}} (\log d_N)^{-1}$ yields $(x \log x)^{-p} \geq 2(k_N d_N)^{-1}$ for all sufficiently large N, where we used $k_N \leq d_N$. We have therefore shown that

$$\mathbf{P}\big[\|H_1^N\| > k_N^{\frac{1}{p}-\frac{1}{2}} d_N^{\frac{1}{p}} (\log d_N)^{-1}\big] \geq 1 - e^{-1}$$

for all sufficiently large N. On the other hand, if the conclusion of Corollary 7.12 holds, then certainly $\|H_1^N\| \to \|s_1\| = 2$ in probability, and thus

$$\limsup_{N\to\infty} k_N^{\frac{1}{p}-\frac{1}{2}} d_N^{\frac{1}{p}} (\log d_N)^{-1} \leq 2.$$

The latter implies $k_N \gtrsim d_N^{\frac{2}{p-2}} (\log d_N)^{-\frac{2p}{p-2}} \gg d_N^{\frac{2}{p-2}} (\log d_N)^{-\frac{4p}{p-2}}$ for large N.

8. Sample covariance matrices

In order to illustrate some subtleties of the model (2.1) from the viewpoint of universality, we will presently discuss a classical model of random matrix theory that arises frequently in applied mathematics: the sample covariance matrix. The simplest model of this kind is defined as follows. Let $Y_1, \ldots, Y_n$ be i.i.d., centered random vectors in $\mathbb{R}^d$ with covariance matrix Σ, and consider the random matrix

$$S = \frac{1}{n} \sum_{i=1}^{n} Y_i Y_i^*. \tag{8.1}$$

A typical problem in this setting is to control the deviation of the sample covariance matrix from its mean $\|S - \mathbf{E}S\|$. As $X = S - \mathbf{E}S$ is a sum of independent random matrices, it is tempting to view X as a special instance of (2.1). As the following example illustrates, however, this viewpoint is not always meaningful.

Example 8.1. Suppose that the entries of the random vectors Y_i are i.i.d. random signs. Then the Gaussian model G associated to $X = S - \mathbf{E}S$ has i.i.d. entries $(G_{ij})_{i>j}$ with $\mathbf{E}[G_{ij}] = 0$ and $\mathrm{Var}(G_{ij}) = \frac{1}{n}$, and diagonal entries $G_{ii} = 0$. The limitation of this model becomes evident in the asymptotic regime where $d, n \to \infty$ in fixed proportion $\frac{d}{n} \to \gamma$. Recall that by a classical result of Bai and Yin [3]

$$\|S - \mathbf{E}S\| \xrightarrow[d/n\to\gamma]{d,n\to\infty} 2\sqrt{\gamma} + \gamma \quad \text{a.s.}$$

On the other hand, the Wigner matrix G satisfies [2, Theorem 2.1.22]

$$\|G\| \xrightarrow[d/n\to\gamma]{d,n\to\infty} 2\sqrt{\gamma} \quad \text{a.s.}$$

Thus it is clear that the universality phenomenon captured by the main results of this paper *cannot* arise in this manner when $\gamma > 0$.

We can nonetheless apply Corollary 2.5, for example, to estimate

$$\big|\mathbf{E}\|S-\mathbf{E}S\| - \mathbf{E}\|G\|\big| \lesssim \sqrt{\frac{\log d}{n}} + \left(\frac{d\log d}{n}\right)^{\frac{2}{3}} + \frac{d\log d}{n},$$

as we may readily compute $\sigma(X)^2 = \frac{d-1}{n}$, $\sigma_*(X)^2 \le \frac{2}{n}$, and $R(X) = \frac{d}{n}$. As $(\mathbf{E}\|G\|)^2 \asymp \frac{d}{n}$ for any d,n (see, e.g., [12]), the above inequality yields universality in the regime $d,n\to\infty$ with $n \gg d\log^4 d$. This is a nontrivial result, but does not provide any understanding of the proportional dimension regime $d \propto n$.

The fundamental issue that is highlighted by the above example is that, despite the fact that (8.1) is a sum of independent random matrices, the Gaussian model G is not the correct universal model associated to the random matrix $X - \mathbf{E}X$ in the proportional dimension regime. The natural universal model in this setting is not obtained by replacing the entries of X by Gaussian random variables, but rather by replacing the vectors Y_i by Gaussian vectors with the same mean and covariance. While the distinction between these two universal models is well understood in classical random matrix theory, it is largely invisible from the viewpoint of classical matrix concentration inequalities because these inequalities are suboptimal by logarithmic factors, cf. [28, §1.5–1.6]. On the other hand, it is clear that in order to obtain *sharp* matrix concentration inequalities, it is essential to apply the universality principle to the correct universal model.

To this end, let us view the sample covariance matrix as $S = YY^*$, where Y is the $d\times n$ random matrix whose columns are $\frac{1}{\sqrt{n}}Y_1,\ldots,\frac{1}{\sqrt{n}}Y_n$. The natural universal model associated to S is $S_{\rm univ} = HH^*$, where H is the $d\times n$ random matrix such that the real and imaginary parts of its entries are jointly Gaussian with the same mean and covariance and those of Y. To show that the spectral properties of S are close to those of $S_{\rm univ}$, we could define the interpolation

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

and prove universality principles by computing $\frac{d}{dt}\mathbf{E}[\mathrm{tr}\, f(Y(t)Y(t)^*)]$ for appropriate spectral statistics f, precisely as we did in the proofs of the main results of this paper. However, as S is merely a quadratic polynomial of Y, we can also directly apply our main results to the present setting using a simple linearization argument as in [4, §3.3], as we will now illustrate. Note that the setting of the following result is far more general than the model (8.1), in that the random vectors $Y_1,\ldots,Y_n$ are not assumed to be centered, independent, or identically distributed.

Theorem 8.2. *Let Z_0 be a $d\times n$ deterministic matrix, let $Z_1,\ldots,Z_m$ be $d\times n$ random matrices with zero mean $\mathbf{E}[Z_i]=0$, and define*

$$Y = Z_0 + \sum_{i=1}^m Z_i.$$

Let H be the $d\times n$ random matrix so that the real and imaginary parts of its entries are jointly Gaussian with the same mean and covariance and those 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

$$\delta=\sigma_*(Y)(\log(d+n))^{\frac{1}{2}}+R(Y)^{\frac{1}{3}}\sigma(Y)^{\frac{2}{3}}(\log(d+n))^{\frac{2}{3}}+R(Y)\log(d+n).$$

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

Sketch of proof. Let $A_\varepsilon=(\|\mathbf{E}YY^*\|+4\varepsilon^2)\mathbf{1}-\mathbf{E}YY^*$, and define

$$\breve{Y}_\varepsilon=\begin{bmatrix}0&0&Y^*\\0&0&A_\varepsilon^{\frac{1}{2}}\\Y&A_\varepsilon^{\frac{1}{2}}&0\end{bmatrix},\qquad\qquad \breve{H}_\varepsilon=\begin{bmatrix}0&0&H^*\\0&0&A_\varepsilon^{\frac{1}{2}}\\H&A_\varepsilon^{\frac{1}{2}}&0\end{bmatrix}.$$

By [4, Remark 2.6], we have $\sigma_*(\breve{Y}_\varepsilon)=\sigma_*(Y)$, $\sigma(\breve{Y}_\varepsilon)=\sigma(Y)$, and $R(\breve{Y}_\varepsilon)=R(Y)$. Moreover, arguing as in the proofs of [4, Theorem 3.12 and Lemma 3.14] yields

$$d_{\mathrm{H}}\big(\mathrm{sp}(\breve{Y}_\varepsilon),\mathrm{sp}(\breve{H}_\varepsilon)\big)\le\varepsilon\qquad\Longrightarrow$$
$$\big|\|YY^*-\mathbf{E}YY^*\|-\|HH^*-\mathbf{E}HH^*\|\big|\le 2\varepsilon\big\{\max\{\|H\|,\|Y\|\}+\|\mathbf{E}YY^*\|^{\frac{1}{2}}\big\}+5\varepsilon^2$$

for any $\varepsilon>0$. Thus Theorem 2.4 yields

$$\mathbf{P}\big[\big|\|YY^*-\mathbf{E}YY^*\|-\|HH^*-\mathbf{E}HH^*\|\big|>\\ C\varepsilon(t)\big\{\max\{\|H\|,\|Y\|\}+\|\mathbf{E}YY^*\|^{\frac{1}{2}}\big\}+C\varepsilon(t)^2\big]\le(2d+n)e^{-t}$$

for all $t>0$, where C is a universal constant and

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

On the other hand, we have

$$\mathbf{P}[\|H\|>\mathbf{E}\|H\|+C\varepsilon(t)]\le e^{-t},$$
$$\mathbf{P}[\|Y\|>2\mathbf{E}\|Y\|+C\varepsilon(t)]\le e^{-t}$$

for all $t>0$ by Gaussian concentration and Talagrand's concentration inequality, respectively, cf. [4, Corollary 4.14 and p. 48]. It follows that

$$\mathbf{P}\big[\big|\|YY^*-\mathbf{E}YY^*\|-\|HH^*-\mathbf{E}HH^*\|\big|>\\ C\varepsilon(t)\big\{\max\{\mathbf{E}\|H\|,\mathbf{E}\|Y\|\}+\|\mathbf{E}YY^*\|^{\frac{1}{2}}\big\}+C\varepsilon(t)^2\big]\le C(d+n)e^{-t}.$$

In particular, integrating the above bound as in the proof of Corollary 2.5 yields

$$\big|\mathbf{E}\|YY^*-\mathbf{E}YY^*\|-\mathbf{E}\|HH^*-\mathbf{E}HH^*\|\big|\lesssim\delta\big\{\max\{\mathbf{E}\|H\|,\mathbf{E}\|Y\|\}+\|\mathbf{E}YY^*\|^{\frac{1}{2}}\big\}+\delta^2.$$

To conclude the proof, it suffices to note that

$$\mathbf{E}\|Y\|\le\mathbf{E}\|H\|+\delta$$

by Corollary 2.5 and Remark 2.1, while

$$\|\mathbf{E}YY^*\|^{\frac{1}{2}}=\|\mathbf{E}HH^*\|^{\frac{1}{2}}\le(\mathbf{E}\|H\|^2)^{\frac{1}{2}}\lesssim\mathbf{E}\|H\|+\delta$$

follows from the Gaussian concentration inequality that was used above. □

To illustrate Theorem 8.2, let us revisit Example 8.1.

Example 8.3. Let Y be the $d \times n$ random matrix defined by

$$Y = \sum_{i=1}^{d} \sum_{j=1}^{n} \frac{\varepsilon_{ij}}{\sqrt{n}} e_i e_j^*,$$

where $(\varepsilon_{ij})_{i \le d, j \le n}$ are i.i.d. random signs. Then the sample covariance matrix of Example 8.1 may be expressed as $S = YY^*$.

We can now apply Theorem 8.2 as follows. The Gaussian model H associated to Y has i.i.d. centered Gaussian entries with variance $\frac{1}{n}$. Moreover, we can readily compute $\sigma(Y)^2 = \max\{\frac{d}{n}, 1\}$ and $\sigma_*(Y)^2 = R(Y)^2 = \frac{1}{n}$, while in this case $\mathbf{E}\|H\| \asymp \sigma(Y)$ for any d, n (see, e.g., [12]). Thus Theorem 8.2 yields

$$\big|\mathbf{E}\|S - \mathbf{E}S\| - \mathbf{E}\|HH^* - \mathbf{E}HH^*\|\big| \lesssim \left(1 + \sqrt{\frac{d}{n}}\right)\delta + \delta^2$$

with

$$\delta \lesssim \left(\frac{\log^2(d+n)}{\sqrt{n}}\right)^{\frac{1}{3}} + \left(\frac{d \log^2(d+n)}{n\sqrt{n}}\right)^{\frac{1}{3}} + \frac{\log(d+n)}{\sqrt{n}}.$$

In the present setting, it is known that

$$\mathbf{E}\|HH^* - \mathbf{E}HH^*\| \asymp \sqrt{\frac{d}{n}} + \frac{d}{n}$$

for all d, n [15]. Thus the above inequality yields universality in the regime $d, n \to \infty$ with $d \gg n^{\frac{2}{3}} \log^{\frac{4}{3}} n$, which includes the case of proportional dimension.

When combined, Examples 8.1 and 8.3 capture the entire random matrix regime $d, n \to \infty$ for the particular model under consideration. However, one must apply the universality principles of this paper in two different ways depending on what universal model captures the behavior of the sample covariance matrix in the given regime (note that the two universality principles overlap in a regime in which the two universal models G and $HH^* - \mathbf{E}HH^*$ themselves behave alike). While this example is a particularly simple one, it serves to illustrate both the versatility of the main results of this paper, and the importance of applying them in the right manner in order to obtain meaningful results.

As was noted above, the setting of Theorem 8.2 is far more general than that of sample covariance matrices. It may be combined with [4, Theorem 3.12] to obtain a very general matrix concentration inequality for $\|YY^* - \mathbf{E}YY^*\|$ in terms of the associated noncommutative model $\|Y_{\text{free}}Y_{\text{free}}^* - \mathbf{E}YY^* \otimes \mathbf{1}\|$. This yields concrete nonasymptotic bounds for generalized sample covariance matrices that may have nonhomogeneous and dependent samples. On the other hand, even in the setting of classical sample covariance matrices (8.1), the kind of dependence that is captured by Theorem 8.2 can be somewhat restrictive. To illustrate this point, let us revisit the model of Examples 8.1 and 8.3 under weaker assumptions.

Example 8.4. Consider the classical sample covariance matrix (8.1) where $Y_1, \ldots, Y_n$ are i.i.d. isotropic random vectors in $\mathbb{R}^d$, that is, $\mathbf{E}[Y_i] = 0$ and $\mathrm{Cov}(Y_i) = \mathbf{1}$.

Let us write as above $S = YY^*$, where Y is the $d \times n$ matrix with columns $\frac{1}{\sqrt{n}}Y_1, \ldots, \frac{1}{\sqrt{n}}Y_n$. Then the Gaussian model H associated to Y is precisely the same as in Example 8.3, and we recall in particular that

$$\mathbf{E}\|H\| \asymp 1 + \sqrt{\frac{d}{n}}, \qquad \mathbf{E}\|HH^* - \mathbf{E}HH^*\| \asymp \sqrt{\frac{d}{n}} + \frac{d}{n}.$$

On the other hand, as we have presently not made any assumptions on the distribution of Y_i beyond its mean and covariance, we can at best represent Y as a sum of independent random matrices in the form

$$Y = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i e_i^*,$$

for which $R(Y)^2 \geq \frac{1}{n} \max_i \mathbf{E}\|Y_i\|^2 = \frac{d}{n}$. We therefore obtain

$$\delta\, \mathbf{E}\|H\| + \delta^2 \gtrsim \mathbf{E}\|HH^* - \mathbf{E}HH^*\| \log(d+n)$$

in Theorem 8.2. In particular, this shows that Theorem 8.2 cannot yield any meaningful universality statement under the present assumptions. (However, note that the computation of Example 8.1 remains valid in the present setting, so we still obtain universal behavior when $n \gg d \log^4 d$ provided $\|Y_i\|^2 \leq Cd$ a.s.)

The problem that arises here is not an inefficiency in Theorem 8.2, but is rather due to the fact that the assumption that $Y_1, \ldots, Y_n$ are i.i.d. and isotropic does not in itself suffice to yield the kind of universality that is captured by Theorem 8.2. The following counterexample illustrates this point. 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. random variables that are uniformly distributed on $[d]$. Then $Y_1, \ldots, Y_n$ are i.i.d. and isotropic. However, as YY^* is diagonal with multinomially distributed diagonal entries, we have [25]

$$\mathbf{E}\|YY^* - \mathbf{E}YY^*\| = (1 + o(1)) \frac{\gamma \log d}{\log(\gamma \log d)} \quad \text{as } d, n \to \infty,\ \tfrac{d}{n} \to \gamma.$$

This example therefore fails to exhibit universality in the regime $d \propto n$.

The above example illustrates that even in the classical setting of isotropic sample covariance matrices, some additional assumption on the distribution of the vectors Y_i is needed in order to obtain universality in the proportional dimension regime. It suffices to assume that the entries of the vectors Y_i are themselves i.i.d., as was illustrated in Example 8.3. More generally, for isotropic covariance matrices, Theorem 8.2 yields universality when $d \propto n$ in the setting where each Y_i is itself a sum of independent random vectors of sufficiently small norm, that is, if

$$Y_i = \sum_{r=1}^{m} Y_i^{(r)} \quad \text{with } Y_i^{(1)}, \ldots, Y_i^{(m)} \text{ independent}, \quad \max_{r \leq m} \|Y_i^{(r)}\|^2 \ll \frac{d}{(\log d)^4} \text{ a.s.}$$

This is the case, for example, if the entries of Y_i form sufficiently small independent blocks, or if each Y_i is itself an average of i.i.d. data.

However, while the model (2.1) of sums of independent random matrices is motivated by a wide variety of applications [28], sums of independent random vectors do not appear to arise as often in practice (except in the special case of independent entries). For isotropic sample covariance matrices, much less restrictive assumptions on the distribution of the vectors Y_i are known to suffice to obtain universal behavior. In particular, it was shown in [10] that to obtain norm universality of isotropic sample covariance matrices, it suffices to assume that the Euclidean norm of every projection of Y_i is sufficiently concentrated. While the model of isotropic sample covariance matrices is very special, the above assumption on the vectors Y_i captures interesting examples that are outside the reach of the methods of this paper (e.g., when Y_i are isotropic log-concave vectors).

The above considerations motivate the potential utility of universality principles for random matrices that admit more general dependence structures than can be captured by the model (2.1). It would be of interest, for example, to obtain general universality principles in the spirit of this paper that are able to explain the behavior of isotropic sample covariance matrices in [10]. It is far from clear, however, what could be a natural formulation of such a principle. A different dependence structure of potential interest in certain applications is to impose a form of mixing condition on the entries of the random matrix; some results in this spirit for the classical random matrix regime may be found in [13].

Acknowledgments. This work was supported in part by the NSF grants DMS-1856221 and DMS-2054565. The authors thanks Afonso Bandeira, March Boedihardjo, and Joel Tropp for many discussions on the topic of this paper.

References

[1] G. W. Anderson. Convergence of the largest singular value of a polynomial in independent Wigner matrices. *Ann. Probab.*, 41(3B):2103–2181, 2013.
[2] 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.
[3] 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.
[4] A. S. Bandeira, M. T. Boedihardjo, and R. van Handel. Matrix concentration inequalities and free probability, 2021. Preprint arxiv:2108.06312v1.
[5] A. D. Barbour. Asymptotic expansions based on smooth functions in the central limit theorem. *Probab. Theory Relat. Fields*, 72(2):289–303, 1986.
[6] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.
[7] 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.
[8] A.-P. Calderón. Intermediate spaces and interpolation, the complex method. *Studia Math.*, 24:113–190, 1964.
[9] 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.
[10] 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.
[11] 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.
[12] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001.
[13] L. Erdős, T. Krüger, and D. Schröder. Random matrices with slow correlation decay. *Forum Math. Sigma*, 7:Paper No. e8, 89, 2019.
[14] 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.
[15] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.
[16] F. Lehner. Computing norms of free operators with matrix coefficients. *Amer. J. Math.*, 121(3):453–486, 1999.
[17] 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.
[18] 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.

[19] 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.
[20] 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.
[21] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges, 2009. Preprint arxiv:0911.0600.
[22] 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.
[23] G. Pisier. *Introduction to operator space theory*, volume 294 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2003.
[24] 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.
[25] 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.
[26] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.
[27] T. Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.
[28] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8:1–230, 2015.
[29] 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.
[30] J. A. Tropp. Second-order matrix concentration inequalities. *Appl. Comput. Harmon. Anal.*, 44(3):700–736, 2018.

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