A NEW APPROACH TO STRONG CONVERGENCE II. THE CLASSICAL ENSEMBLES

CHI-FANG CHEN, JORGE GARZA-VARGAS, AND RAMON VAN HANDEL

Abstract. The first paper in this series introduced a new approach to strong convergence of random matrices that is based primarily on soft arguments. This method was applied to achieve a refined qualitative and quantitative understanding of strong convergence of random permutation matrices and of more general representations of the symmetric group. In this paper, we introduce new ideas that make it possible to achieve stronger quantitative results and that facilitate the application of the method to new models.

When applied to the Gaussian GUE/GOE/GSE ensembles of dimension N, these methods achieve strong convergence for noncommutative polynomials with matrix coefficients of dimension $\exp(o(N))$. This provides a sharp form of a result of Pisier on strong convergence with coefficients in a subexponential operator space. Analogous results up to logarithmic factors are obtained for Haar-distributed random matrices in $\mathrm{U}(N)/\mathrm{O}(N)/\mathrm{Sp}(N)$. We further illustrate the methods of this paper in the following applications.

1. We obtain improved rates for strong convergence of random permutations.
2. We obtain a quantitative form of strong convergence of the model introduced by Hayes for the solution of the Peterson-Thom conjecture.
3. We prove strong convergence of tensor GUE models of Γ-independence.
4. We prove strong convergence of all nontrivial representations of $\mathrm{SU}(N)$ of dimension up to $\exp(N^{1/3-\delta})$, improving a result of Magee and de la Salle.

1. Introduction

A sequence of r-tuples $\boldsymbol{X}^N = (X_1^N, \ldots, X_r^N)$ of random matrices is said to *converge strongly* to an r-tuple $\boldsymbol{x} = (x_1, \ldots, x_r)$ of elements of a C^*-algebra if

$$\lim_{N\to\infty} \|P(X_1^N, \ldots, X_r^N, X_1^{N*}, \ldots, X_r^{N*})\| = \|P(x_1, \ldots, x_r, x_1^*, \ldots, x_r^*)\|$$

in probability for every noncommutative polynomial P. In recent years, this notion has proved to have powerful consequences for problems of random graphs, random surfaces, and operator algebras, resulting in major breakthroughs in these areas; see, for example, [7, 8, 15] for discussion and references.

Our previous paper [15] introduced a new approach to strong convergence that is based primarily on soft arguments and requires limited problem-specific inputs, in contrast to earlier approaches that were heavily dependent on problem-specific analytic methods and/or delicate combinatorial estimates. The replacement of hard analysis by soft arguments has made it possible to establish strong convergence in new situations, and to access quantitative information about the strong convergence phenomenon that was previously out of reach. Both features were illustrated in [15]

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

in the context of random permutation matrices and of more general representations of the symmetric group. Another illustration is provided by the remarkable results of Magee and de la Salle [37] and Cassidy [13] that establish strong convergence of extremely high-dimensional representations of $\mathrm{U}(N)$ and S_N.

In this paper, we develop new ideas that advance the method of [15] in two directions: they achieve considerably stronger (and in some respects nearly optimal) quantitative results; and they further eliminate the need for problem-specific computations in many situations, which facilitates the application of the method to new models. These new ingredients, which will be discussed in section 1.2 below, enable a particularly transparent and nearly parallel treatment of the classical invariant ensembles of random matrix theory that yields new quantitative information on the strong convergence phenomenon for these models. Beyond the classical ensembles, we will further illustrate these methods in several additional applications.

1.1. **Main results.**

1.1.1. *The Gaussian ensembles.* In this paper, a GUE/GOE/GSE random matrix of dimension N is an $N \times N$ self-adjoint random matrix G^N whose entries above the diagonal are independent complex/real/quaternionic Gaussian variables with variance $\frac{1}{N}$. The limiting operators associated to independent random matrices $G_1^N, \ldots, G_r^N$ from these ensembles form a free semicircular family $s_1, \ldots, s_r$.[1] Precise definitions of these notions will be recalled in section 2.3.

The following is the main result on this paper on the Gaussian ensembles. A more explicit form of the constant c appears in the proof.

Theorem 1.1. *Let $\boldsymbol{G}^N = (G_1^N, \ldots, G_r^N)$ be i.i.d. GUE/GOE/GSE random matrices of dimension N, and let $\boldsymbol{s} = (s_1, \ldots, s_r)$ be a free semicircular family. Let $\varepsilon \in (0, 1]$ and $q_0 \in \mathbb{N}$. Then for every noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{s} \rangle$ of degree q_0 with matrix coefficients of dimension $D \le e^{cN\varepsilon^2}$, we have*

$$\mathbf{P}\big[\|P(\boldsymbol{G}^N)\| \ge (1+\varepsilon)\|P(\boldsymbol{s})\|\big] \le \frac{N}{c\varepsilon}\, e^{-cN\varepsilon^2},$$

where c is a constant that depends only on q_0 and r.

Theorem 1.1 settles a question that has motivated many recent works on strong convergence. In its basic form, strong convergence implies that

$$\|P(\boldsymbol{G}^N)\| = (1 + o(1))\|P(\boldsymbol{s})\| \quad \text{as} \quad N \to \infty$$

when the polynomial P and thus the coefficient dimension D is fixed. However, much stronger implications could be obtained if D is allowed to grow sufficiently rapidly with N (a considerable strengthening of the strong convergence property). For example, Hayes [30] shows that the case $D = N$ already suffices to prove the Peterson-Thom conjecture in the theory of von Neumann algebras. The latter was settled in [4, 7, 37], and Theorem 1.1 provides yet another proof of this conjecture. On the other hand, in his study of subexponential operator spaces, Pisier [50] shows that strong convergence holds *up to a factor* 2 even when $D = e^{o(N)}$.

[1] The reader is warned that this notation differs from [15]. In the present paper, a free semicircular family is denoted as $\boldsymbol{s} = (s_1, \ldots, s_r)$, while free Haar unitaries are denoted as $\boldsymbol{u} = (u_1, \ldots, u_r)$.

Theorem 1.1 closes the gap between these two extremes by showing that strong convergence holds whenever $D = e^{o(N)}$, providing a sharp form of Pisier's theorem.

Corollary 1.2. *Let $\boldsymbol{G}^N$ and $\boldsymbol{s}$ be defined as in Theorem 1.1. For any sequence of noncommutative polynomials $P_N \in \mathrm{M}_{D_N}(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{s}\rangle$ of degree $O(1)$ and matrix coefficients of dimension $D_N = e^{o(N)}$, we have*

$$\|P_N(\boldsymbol{G}^N)\| = (1 + o(1))\|P_N(\boldsymbol{s})\| \quad \text{a.s.} \quad \text{as} \quad N \to \infty.$$

Consequently, for any noncommutative polynomial $Q \in \mathbf{W} \otimes \mathbb{C}\langle \boldsymbol{s}\rangle$ with coefficients in a subexponential operator space $\mathbf{W}$, we have

$$\frac{1}{C(\mathbf{W})}\|Q(\boldsymbol{s})\|_{\min} \le \liminf_{N\to\infty} \|Q(\boldsymbol{G}^N)\| \le \limsup_{N\to\infty} \|Q(\boldsymbol{G}^N)\| \le C(\mathbf{W})\|Q(\boldsymbol{s})\|_{\min} \ \text{a.s.},$$

where $C(\mathbf{W})$ denotes the subexponential constant of $\mathbf{W}$.[2]

Whether strong convergence could hold even beyond the subexponential regime is a tantalizing question. While we do not resolve this question, our results are optimal in the sense that they achieve the largest regime that is accessible by trace statistics, as will be discussed in section 1.3.1 below.

Remark 1.3 (Previous bounds)**.** Prior to the present work, strong convergence of Gaussian ensembles was established only for $D = o(N/\log^3 N)$ [2] (following earlier works that achieved $D = o(N^{1/4})$ [50] and $D = o(N^{1/3})$ [17]). Thus even the linear dimension regime had remained out of reach in this setting.

For Haar unitary matrices (see section 1.1.2), analogous sublinear bounds appear in [45, 46]. In this setting, a major step forward [7] achieved strong convergence for $D \le \exp(N^{1/(32r+160)})$, breaking the linear dimension barrier. A significant improvement $D \le \exp(N^{1/2-o(1)})$ was obtained in [37]. Most recently, in work concurrent with the present paper, the GUE case was revisited in [48] where strong convergence was proved for $D = \exp(o(N^{2/3}))$. The methods of the present paper finally make it possible to reach the entire subexponential regime.

In the complementary regime where $D = N^{O(1)}$ is polynomial, Theorem 1.1 yields a universal bound on the rate of strong convergence: a direct application of Theorem 1.1 with $\varepsilon = C\sqrt{\log(N)/N}$ yields the following.

Corollary 1.4. *Let $\boldsymbol{G}^N$ and $\boldsymbol{s}$ be defined as in Theorem 1.1. For any sequence of noncommutative polynomials $P_N \in \mathrm{M}_{D_N}(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{s}\rangle$ with $D_N = N^{O(1)}$, we have*[3]

$$\|P_N(\boldsymbol{G}^N)\| \le \left(1 + O_{\mathrm{P}}\left(\sqrt{\frac{\log N}{N}}\right)\right)\|P_N(\boldsymbol{s})\| \quad \text{as} \quad N \to \infty.$$

A similar rate was obtained by Parraud [47] for GUE (and Haar unitaries [46]) with $D = 1$, but the method used there does not extend to large D.

The $N^{-1/2}$ rate is not expected to be optimal: when P is a linear polynomial with scalar coefficients, the optimal rate $N^{-2/3}$ follows from classical Tracy-Widom asymptotics. At present, however, Corollary 1.4 yields the best known rate for arbitrary polynomials P; see section 1.3.2 for further discussion.

[2]The definition of a subexponential operator space is recalled in section 9.1.

[3]The notation $Z_N = O_{\mathrm{P}}(z_N)$ denotes that $\{Z_N/z_N\}_{N\ge 1}$ is bounded in probability.

1.1.2. *The classical compact groups.* For Haar-distributed random matrices from the classical compact groups $\mathrm{U}(N)/\mathrm{O}(N)/\mathrm{Sp}(N)$, the methods of this paper yield nearly parallel results to those obtained for the Gaussian ensembles. Our main result in this setting differs from Theorem 1.1 by a logarithmic factor.

Theorem 1.5. *Let $\boldsymbol{U}^N = (U_1^N, \ldots, U_r^N)$ be i.i.d. Haar-distributed random matrices in $\mathrm{U}(N)/\mathrm{O}(N)/\mathrm{Sp}(N)$, and let $\boldsymbol{u} = (u_1, \ldots, u_r)$ be free Haar unitaries. Let $\varepsilon \in [\frac{1}{c\sqrt{N}}, 1]$ and $q_0 \in \mathbb{N}$. For every noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{u}, \boldsymbol{u}^* \rangle$ of degree q_0 with matrix coefficients of dimension $D \le e^{cN\varepsilon^2/\log^2(N\varepsilon^2)}$, we have*

$$\mathbf{P}\big[\|P(\boldsymbol{U}^N, \boldsymbol{U}^{N*})\| \ge (1+\varepsilon)\|P(\boldsymbol{u}, \boldsymbol{u}^*)\|\big] \le \frac{N}{c\varepsilon}\, e^{-cN\varepsilon^2/\log^2(N\varepsilon^2)}.$$

Here c is a constant that depends only on q_0 and r.

Theorem 1.5 yields strong convergence whenever $D = e^{o(N/\log^2 N)}$, and yields a universal rate $O_{\mathrm{P}}((\frac{\log N}{N})^{1/2} \log\log N)$ when $D = N^{O(1)}$.

Remark 1.6. Theorem 1.5 falls short by a logarithmic factor of reaching the full subexponential regime. However, as was pointed out to us by Mikael de la Salle, Corollary 1.2 extends *verbatim* to the setting of Theorem 1.5 by using a coupling between the Gaussian and Haar-distributed ensembles due to Collins and Male [18]. This argument achieves the full subexponential regime for the $\mathrm{U}(N)/\mathrm{O}(N)/\mathrm{Sp}(N)$ models, but does not provide any quantitative information.

This suggests that the logarithmic factor in Theorem 1.5 is likely an artefact of the proof and should not be necessary. The logarithmic factor arises from a single point in the proof that is explained in Remark 7.2 below.

1.1.3. *Further applications.* While the new ingredients developed in this paper enable a particularly sharp treatment of the classical Gaussian and Haar-distributed ensembles, they are by no means restricted to this setting. To further illustrate the utility of these methods, we will develop four additional applications.

1. It was shown in [15] that strong convergence of random permutation matrices holds with a universal rate $O_{\mathrm{P}}((\frac{\log N}{N})^{1/8})$. We will obtain $O_{\mathrm{P}}((\frac{\log N}{N})^{1/6})$ with no additional effort, improving the best known rate for this problem.
2. In an influential paper that motivated much recent work on strong convergence, Hayes [30] proved a reduction of the Peterson-Thom conjecture to the statement that the family of N^2-dimensional random matrices

   $$G_1^N \otimes \mathbf{1}_N,\ \ldots,\ G_r^N \otimes \mathbf{1}_N,\ \mathbf{1}_N \otimes \tilde{G}_1^N,\ \ldots,\ \mathbf{1}_N \otimes \tilde{G}_r^N$$

   converges strongly as $N \to \infty$ to

   $$s_1 \otimes \mathbf{1},\ \ldots,\ s_r \otimes \mathbf{1},\ \mathbf{1} \otimes s_1,\ \ldots,\ \mathbf{1} \otimes s_r,$$

   where $G_i^N, \tilde{G}_i^N$ are independent GUE matrices of dimension N and s_i are free semicircular variables. This strong convergence property follows readily from Corollary 1.2 using exactness of the C^*-algebra generated by a free semicircular family, but this argument provides no quantitative information. We will develop a quantitative form of strong convergence of Hayes' model.

3. In contrast to the Hayes model, where GUE matrices act on distinct factors of a tensor product, models where GUE matrices act on overlapping factors of a tensor product arise naturally in the study of quantum many-body systems and in random geometry. The limiting model in this setting is described by the notion of Γ-independence [53, 14]. We will prove strong convergence of general tensor GUE models to a Γ-independent semicircular family, settling open problems formulated in [40, Problem 1.6] and [21]. Quantitative bounds on strong convergence for the Hayes model play a key role in the proof.
4. Let $g_1, \ldots, g_r$ be i.i.d. Haar distributed elements of $\mathrm{SU}(N)$. Given a unitary representation π_N of $\mathrm{SU}(N)$, we can define random matrices $U_i^{\pi_N} := \pi_N(g_i)$ of dimension $\dim(\pi_N)$. It was shown by Magee and de la Salle [37] that the random matrices $U_1^{\pi_N}, \ldots, U_r^{\pi_N}$ converge strongly to free Haar unitaries $u_1, \ldots, u_r$ uniformly over all nontrivial representations π_N with $\dim(\pi_N) \le \exp(N^{1/24-\delta})$, achieving the first strong convergence result for representations of quasiexponential dimension (much lower dimensional representations were considered in earlier work of Bordenave and Collins [8]). We will improve this conclusion to representations of dimension $\dim(\pi_N) \le \exp(N^{1/3-\delta})$.

We postpone precise mathematical statements of these results to section 9, where the above applications will be developed.

1.2. **New ingredients.** A detailed overview of the soft approach to strong convergence introduced in our previous paper is given in [15, §2.2]. Here we summarily recall only the most basic steps of this method in order to enable the discussion of the new ideas developed in this paper. The reader who is new to the method is encouraged to review *ibid.* prior to proceeding.

1.2.1. *Review of the basic method.* Let X^N (e.g., $P(\boldsymbol{U}^N, \boldsymbol{U}^{N*})$) be a self-adjoint random matrix of dimension M_N, and let X_{F} (e.g., $P(\boldsymbol{u}, \boldsymbol{u}^*)$) be its limiting model in a C^*-probability space $(\mathcal{A}, \tau)$. For the present discussion, assume for simplicity that $\|X^N\| \le K$ a.s. To control the norm of X^N, we bound

$$\mathbf{P}\big[\|X^N\| \ge \|X_{\mathrm{F}}\| + \varepsilon\big] \le \mathbf{E}[\mathrm{Tr}\,\chi(X^N)] = M_N\,\mathbf{E}[\mathrm{tr}\,\chi(X^N)], \tag{1.1}$$

where tr denotes the normalized trace and $\chi \ge 0$ is a smooth test function so that $\chi(x)$ vanishes for $|x| \le \|X_{\mathrm{F}}\| + \frac{\varepsilon}{2}$ and equals one for $|x| \ge \|X_{\mathrm{F}}\| + \varepsilon$. Our aim is to show that the right-hand side of this bound is $o(1)$. While we are ultimately interested in smooth test functions χ, our approach is based on the availability of powerful tools in the analytic theory of polynomials; the application of these tools to polynomial test functions will yield quantitative bounds that are so strong that they can be lifted to smooth test functions a posteriori.

The basic input for our approach is that for many models (including all those considered in this paper), any polynomial h of X^N satisfies

$$\mathbf{E}[\mathrm{tr}\,h(X^N)] = \Phi_h\big(\tfrac{1}{N}\big),$$

where Φ_h is a rational function whose degree is bounded in terms of that of h. Using only this fact and the trivial bound $|\Phi_h(\frac{1}{N})| \le \|h\|_{[-K,K]} := \sup_{|x|\le K} |h(x)|$, we can apply classical polynomial inequalities due to A. and V. Markov to obtain

an asymptotic expansion of $\mathbf{E}[\operatorname{tr} h(X^N)]$ of the form

$$\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \tau(h(X_{\mathrm{F}})) - \sum_{k=1}^{m-1} \frac{\nu_k(h)}{N^k}\right| \le \frac{C(m)}{N^m}\, q^{\beta m} \|h\|_{[-K,K]} \tag{1.2}$$

for all $N, m, q \in \mathbb{N}$ and real polynomials h of degree at most q. Here $C(m)$ is a constant that depends only on m, β is a universal constant, and ν_k are linear functionals on the space of real univariate polynomials.

The key feature of (1.2) is that the error bound is sufficiently strong that it can efficiently control the expansion of any smooth function into Chebyshev polynomials. In particular, a Fourier-analytic argument shows that the linear functionals ν_k extend to compactly supported distributions, and that the expansion (1.2) remains valid for any smooth function h when $q^{\beta m}\|h\|_{[-K,K]}$ is replaced by $\|h\|_{C^{\lceil \beta m\rceil+1}[-K,K]}$ on the right-hand side. If we could furthermore show that

$$\operatorname{supp} \nu_k \subseteq [-\|X_{\mathrm{F}}\|, \|X_{\mathrm{F}}\|] \quad \text{for } k = 1, \ldots, m-1, \tag{1.3}$$

then χ in (1.1) would satisfy $\nu_k(\chi) = 0$ for $k \le m-1$ and the expansion yields

$$M_N\, \mathbf{E}[\operatorname{tr} \chi(X^N)] = O\left(\frac{C'(m) M_N}{N^m}\right).$$

Thus we achieve strong convergence provided that (1.3) can be established for m sufficiently large that the above bound is $o(1)$. It is shown in [15] how the problem of bounding the support of ν_k can be reduced to a moment computation; the latter is the main part of the method that relies on a problem-specific analysis.

Remark 1.7. In the context of strong convergence, asymptotic expansions for smooth test functions were first used by Schultz [52], and were systematically developed by Parraud [47, 46, 48] for GUE and Haar unitary matrices. These works rely on specialized analytic tools and explicit computations that are available for these models. A key feature of the polynomial method is that it applies to models for which such specialized tools are not available. At the same time, our main results yield stronger quantitative information even for the classical ensembles.

The basic approach described above achieves not only strong convergence, but also strong quantitative bounds that yield much stronger implications. The quantitative features of the method are controlled by three parameters:

1. The value of β in (1.2);
2. The dependence of $C(m)$ in (1.2) on m;
3. The largest m for which (1.3) can be established.

The main contribution of this paper are several (independent) new ingredients that yield significant improvements to the method of [15] in each of these parameters. We describe these new ingredients in the remainder of this section. The combination of all these ingredients is key to achieving our main results.

1.2.2. *Optimal polynomial interpolation.* The proof of (1.2) is based on the observation that its left-hand side is merely the Taylor expansion to order $m-1$ of the rational function $\Phi_h(\frac{1}{N})$, so that it is bounded by the remainder term $\frac{1}{m!N^m}\|\Phi_h^{(m)}\|_{[0,\frac{1}{N}]}$. The problem with this bound is that it depends on $\Phi_h(x)$ for x not of the form

$\frac{1}{N}$, so that the connection with random matrices is lost. We surmount this using the classical fact that bounding a polynomial on a sufficiently dense discrete set already suffices to achieve uniform control of its derivatives.

A key step in the argument is that we must bound the rational function Φ_h in between the points $\frac{1}{N}$ by interpolating its values at these points. To this end, [15] relies on a classical result on polynomial interpolation, which states that for any real polynomial h of degree q, we have $\|h\|_{[0,\delta]} \lesssim \max_{x\in I} |h(x)|$ for any set $I \subseteq [0,\delta]$ with spacing at most $O(\frac{\delta}{q^2})$ between its points. The latter condition is optimal for a general set I [22]. When applied to the set $I_M := \{\frac{1}{N} : N \geq M\}$ that is of interest in the present setting, the spacing condition limits us to considering only $N \gtrsim q^2$, which results in a quantitative loss in the analysis.

Surprisingly, this restriction turns out to be suboptimal in the present setting due to the special structure of the set I_M: even though $O(\frac{\delta}{q^2})$ spacing is necessary for general I, we will prove in section 3 that $O(\frac{\delta}{q})$ spacing suffices (and is optimal) for I_M, so that we can in fact work with $N \gtrsim q$ in the analysis. While this is in itself purely a statement about polynomials that is unrelated to random matrices, it yields a crucial improvement of the constant β in (1.2) in essentially every application of the polynomial method in the random matrix context.

The special feature of the set I_M is that is becomes increasingly dense near zero. This enables us to exploit a result of Rakhmanov [51], which states that $O(\frac{\delta}{q})$ uniformly spaced points suffice to interpolate a real polynomial of degree q strictly in the interior of the interval $[0,\delta]$, in a multiscale manner.

1.2.3. *High-order expansions.* The strong quantitative results of this paper rely on asymptotic expansion to very large order m (e.g., $m \propto N$ for Gaussian ensembles), which requires an essentially optimal constant $C(m)$ in (1.2). To this end, we must overcome two distinct obstacles that arise in different models.

For Haar-distributed models, we aim to apply inequalities for polynomials to the rational function Φ_h. This is accomplished in [15] by applying the chain rule to express $\Phi_h^{(m)}$ in terms of the derivatives of the numerator and denominator and bounding each term separately, resulting in lossy estimates. In section 7.2, we show instead that the rational function Φ_h can be approximated to very high precision by a polynomial of nearly the same degree, which enables us to apply polynomial inequalities directly without incurring a quantitative loss.

For Gaussian ensembles, the function Φ_h is itself a polynomial (known as the genus expansion), so that the above issue does not arise. In this setting, however, the random matrices are not uniformly bounded, so that the assumption $\|X^N\| \leq K$ a.s. that was made for simplicity in section 1.2.1 does not apply. Surmounting this issue requires a truncation argument. The challenge in implementing such an argument is that this must be done without incurring any quantitative loss in the final bounds. The methods to do so will be developed in section 4.

1.2.4. *Support and concentration.* The quantitative features of our bounds are controlled not only by the asymptotic expansion (1.2), but also by the number of distributions ν_k whose supports can be bounded as in (1.3). In general, (1.3) need not hold for arbitrarily large m; for example, this is the case for models based on random permutation matrices. The latter phenomenon has a precise probabilistic

interpretation: $x \in \operatorname{supp}\nu_k$ for some $|x| > \|X_{\mathrm{F}}\|$ detects the presence of an outlier in the spectrum of X^N with probability $\sim N^{1-k}$, cf. [15, §3.2.2].

Unlike random permutation matrices, however, the norms of random matrices constructed from Gaussians or the classical compact groups are subject to the concentration of measure phenomenon [35], which ensures that the probability that the norm deviates from its median by a fixed amount is *exponentially* small in N. Thus if strong convergence holds in a qualitative sense $\operatorname{med}(\|X^N\|) = \|X_F\| + o(1)$, then the presence of an outlier in the spectrum with probability N^{-c} is automatically ruled out. In other words, whenever (1.2) holds, we have the formal implication

$$\text{concentration of measure} + \text{qualitative strong convergence} \implies \operatorname{supp}\nu_k \subseteq [-\|X_{\mathrm{F}}\|, \|X_{\mathrm{F}}\|] \quad \text{for all } k \ge 1,$$

cf. section 5.2. When applicable, this simple observation controls the supports of ν_k in a soft manner, avoiding the need for problem-specific moment computations such as those used in [15] for random permutations. Let us note that a variant of this idea appears in the work of Parraud [47, pp. 285–286].

It should be emphasized that the above argument cannot in itself prove strong convergence, as it requires strong convergence as input. However, establishing strong convergence generally only requires the validity of (1.3) for small m: for example, when X^N has dimension $M_N = N$, we must only understand ν_1 ($m = 2$) to achieve strong convergence. The concentration argument then automatically extends the conclusion to ν_k for all $k > 1$, resulting in far stronger quantitative bounds. Thus the concentration method serves as a powerful bootstrapping argument to deduce strong quantitative bounds from weak ones.

1.2.5. *Supersymmetric duality.* In contrast to the above techniques that are broadly applicable, we finally discuss an idea that is special to the classical random matrix ensembles. Despite its limited range of applicability, a special property of these models can be very fruitfully exploited when it is present.

A remarkable property of the classical ensembles is that the rational function $\mathbf{E}[\operatorname{tr} h(X^N)] = \Phi_h(\frac{1}{N})$ still has a spectral interpretation if we replace N by $-N$: there exists a "dual" random matrix model Y^N so that

$$\mathbf{E}[\operatorname{tr} h(Y^N)] = \Phi_h(-\tfrac{1}{N}).$$

In particular, GUE and $\mathrm{U}(N)$ models are self-dual, while GOE and $\mathrm{O}(N)$ models are dual to GSE and $\mathrm{Sp}(N)$ models [12, 39]. We presently explain how this yields stronger bounds using a classical property of polynomials [10].

Recall (cf. section 1.2.2) that the proof of (1.2) aims to bound the remainder term $\|\Phi_h^{(m)}\|_{[0,\frac{1}{N}]}$ in the Taylor expansion of Φ_h, while polynomial interpolation yields a uniform bound on Φ_h itself. This is achieved using the Markov inequality: if a real polynomial h of degree q is uniformly bounded on an interval, its derivative is bounded by $O(q^2)$ everywhere on that interval.

While the Markov inequality is optimal if we aim to bound the derivative of a polynomial everywhere on the interval, a much better $O(q)$ bound holds strictly in the interior of the interval by the Bernstein inequality. Unfortunately, this is not applicable in our setting, as we can only control Φ_h in a positive interval $[0, \delta]$ and

we aim to control its derivatives near a boundary point 0 of this interval. However, the existence of a dual random matrix model enables us to bound Φ_h on a symmetric interval $[-\delta,\delta]$. Thus in this case we can apply the Bernstein inequality to achieve a crucial improvement to the constant β in (1.2).

Beside this quantitative improvement, we will also exploit the duality property in an entirely different manner: by an elementary observation, the existence of a dual random matrix model automatically implies the validity of (1.3) for $m = 2$, cf. section 6.3. Somewhat surprisingly, this completely eliminates the need for any problem-specific moment estimates from the proofs of our main results.

1.3. **Discussion.**

1.3.1. *The optimal dimension of matrix coefficients.* Corollary 1.2 states that the Gaussian ensembles exhibit strong convergence for polynomials with matrix coefficients of subexponential dimension $D = e^{o(N)}$. Whether or not this conclusion is the best possible is a tantalizing question of Pisier [50]. Pisier shows[4] that strong convergence can fail in the subgaussian regime $D = e^{O(N^2)}$; what happens in between the subexponential and subgaussian regimes remains open.

However, the results of the present paper are already optimal in a weaker sense: the subexponential regime is the largest one that is accessible by trace methods. This phenomenon is best illustrated by means of a simple example. Let G^N be a GUE matrix of dimension N, and consider the random matrix

$$X^N = \mathbf{1}_D \otimes G^N.$$

We aim to show that $\|X^N\| = 2+o(1)$, which is obvious due to the special structure of the model. However, in general we have no way of reasoning directly about the norm; instead, we bound the norm by a trace statistic such as $\mathbf{E}[\operatorname{Tr}\chi(X^N)]$ in (1.1), which is amenable to computation. But in the present example,

$$\mathbf{E}[\operatorname{Tr}\chi(X^N)] \ge \mathbf{E}[\#\{\text{eigenvalues of } X^N \text{ in } [-(2+\varepsilon),2+\varepsilon]^c\}] \ge De^{-cN}$$

for a universal constant c, where we used $\mathbf{P}[\|G^N\| \ge 2+\varepsilon] \ge \mathbf{P}[G^N_{11} \ge 2+\varepsilon] \ge e^{-cN}$ and that any eigenvalue of G^N gives rise to an eigenvalue of X^N of multiplicity D. Thus $\mathbf{E}[\operatorname{Tr}\chi(X^N)] = o(1)$ can only occur when $D = e^{O(N)}$.

This example shows that the main results of this paper are the best possible in the sense that they capture the optimal regime where the expected number of outlier eigenvalues is $o(1)$. The reason for this, however, is that when matrix coefficients are very high dimensional, outlier eigenvalues can appear with very large multiplicity. This does not rule out the possibility that strong convergence holds in the superexponential regime, but presents a fundamental obstacle to the application of (1.1) or of any other standard trace method (e.g., the moment method).

1.3.2. *The optimal rate of convergence.* While the new ideas of this paper have made it possible to implement most of the ingredients of the basic method of [15] in a nearly optimal manner, one significant inefficiency remains: the $N^{-1/2+o(1)}$ rate of strong convergence achieved by Corollary 1.4 is not expected to be optimal. For linear polynomials P, the optimal rate $N^{-2/3}$ follows from classical Tracy-Widom

[4] As this is not stated explicitly in [50], we include a short self-contained proof in Appendix A.

asymptotics, and heuristic universality principles of random matrix theory suggest that the same rate should extend to arbitrary P.

The source of this inefficiency is the parameter β in the asymptotic expansion (1.2). For both Gaussian and Haar-distributed ensembles, the methods of this paper yield $\beta = 2$. An $N^{-2/3+o(1)}$ rate would follow if this could be improved to $\beta = \frac{3}{2}$. In the very special case of linear polynomials P of GUE matrices, such an expansion has in fact been established by Haagerup and Thorbjørnsen [29] by using explicit differential equations satisfied by the GUE density, lending credence to the validity of such an expansion in the more general setting.

In our approach, β is ultimately controlled by the Bernstein inequality from the analytic theory of polynomials [10], which is optimal for arbitrary polynomials. The conjectured validity of an improved β therefore suggests that the polynomials that arise from the function Φ_h in random matrix models are somewhat better behaved than the worst case polynomials. It is unclear, however, how the latter may be captured. There are known improvements of the Bernstein inequality for special classes of polynomials (e.g., those with no roots in the unit disc [24]) that would imply $\beta = \frac{3}{2}$ if they were applicable, but numerical evidence suggests that the polynomials that arise here do not satisfy the requisite assumptions.

1.3.3. *Beyond the classical ensembles.* An unexpected feature of the present paper is that the entire analysis of the classical Gaussian and Haar-distributed ensembles uses only qualitative properties of the rational function Φ_h and general tools such as concentration of measure. Beyond these basic ingredients, no problem-specific arguments are used in the analysis. It should be emphasized, however, that this simple analysis is enabled by the serendipitous coincidence of two special properties of the classical ensembles: they are of dimension N, so that (1.3) need only be established for $m = 2$ to achieve strong convergence; and they admit a dual model as in section 1.2.5, which yields the latter property automatically.

These properties are by no means needed for the application of our approach, and both the methods of [15] and of the present paper are much more broadly applicable. In general, however, it should be expected that (1.3) must be established at least for small m by means of a problem-specific moment computation as in [15]. Such moment estimates are greatly facilitated by a technique developed by Magee and de la Salle [37, §6.2], which shows that it often possible to reduce such estimates to the special case of polynomials P with nonnegative coefficients.

Let us finally note that while the unexpected appearance of dual random matrix models for the classical ensembles might raise the hope that this phenomenon arises more generally, that does not appear to be the case. For example, it is unlikely that random permutation models admit dual random matrix models, as the Deligne category $\mathrm{Rep}(\mathrm{S}_t)$ is semisimple abelian for all $t < 0$ [23, Theorem 2.18 and §9.5].

1.4. **Organization of this paper.** The rest of this paper is organized as follows.

In section 2, we recall some basic definitions and analytic tools that will be used throughout the paper. Section 3 develops an optimal polynomial interpolation bound for $\frac{1}{N}$ samples that will be used in all results in this paper.

In section 4, we implement the polynomial method to achieve an asymptotic expansion for smooth spectral statistics of GUE. This is combined in section 5 with

a bootstrapping argument to prove Theorem 1.1 in the GUE case. The analysis is extended to GOE/GSE matrices in section 6, concluding the proof of Theorem 1.1. The proof of Theorem 1.5 is contained in sections 7 and 8, which develop the corresponding arguments for the $\mathrm{U}(N)$ and $\mathrm{O}(N)/\mathrm{Sp}(N)$ models, respectively.

Section 9 develops applications to subexponential operator spaces (Corollary 1.2), random permutation models, Hayes' model of the Peterson-Thom conjecture, tensor GUE models, and high-dimensional representations of $\mathrm{SU}(N)$.

The paper concludes with three appendices. Appendix A discusses Pisier's upper bound on the dimension of matrix coefficients for which strong convergence can hold. Appendix B develops an approximation result for Γ-independent semicircular families that is used in the treatment of tensor models. Appendix C contains a result of Magee on duality of stable representations of $\mathrm{U}(N)$.

1.5. **Notation.** Throughout this paper, $a \lesssim b$ denotes that $a \le Cb$ for a universal constant $C > 0$. Unless otherwise specified, $C, c > 0$ denote universal constants that may change from line to line in proofs. We write $[r] := \{1, \ldots, r\}$ for $r \in \mathbb{N}$.

We denote by $\mathcal{P}$ to denote the space of all real univariate polynomials, and by $\mathcal{P}_q \subset \mathcal{P}$ the polynomials of degree at most q. We denote by $h^{(m)}$ the mth derivative of a univariate function h, and we will write $\|h\|_I := \sup_{x \in I} |h(x)|$ for $I \subseteq \mathbb{R}$.

We denote by $\mathrm{M}_N(\mathcal{A})$ the space of $N \times N$ matrices with entries in $\mathcal{A}$. The unnormalized and normalized traces of $M \in \mathrm{M}_N(\mathbb{C})$ are denoted as $\operatorname{Tr} M$ and $\operatorname{tr} M := \frac{1}{N} \operatorname{Tr} M$, respectively. The identity matrix or operator is denoted as $\mathbf{1}$.

2. Preliminaries

The aim of this section is to recall a number of basic tools that will be used throughout this paper, as well as to recall the precise definitions of the classical random matrix ensembles and their limiting models.

2.1. **Polynomial inequalities.** If a real polynomial is bounded in a finite interval, we can control its derivatives inside that interval and its growth outside the interval. The first property is captured by the Bernstein inequality, which plays a fundamental role throughout this paper; it replaces the use of the Markov brothers inequality in [15]. We recall here a version for higher derivatives.

Lemma 2.1 (Bernstein inequality)**.** *For any $h \in \mathcal{P}_q$ and $\delta > 0$, we have*

$$|h^{(m)}(x)| \le \left(\frac{2q}{\delta\sqrt{1-(x/\delta)^2}}\right)^m \|h\|_{[-\delta,\delta]} \quad \text{for all } x \in (-\delta, \delta).$$

Proof. The statement is given for $\delta = 1$ in [10, p. 260], and follows for arbitrary $\delta > 0$ by a straightforward scaling argument. □

The second property is captured by the following extrapolation lemma.

Lemma 2.2. *For any $h \in \mathcal{P}_q$ and $K > 0$, we have*

$$|h(x)| \le \left(\frac{2|x|}{K}\right)^q \|h\|_{[-K,K]} \quad \text{for all } x \in \mathbb{R}\backslash[-K, K].$$

The proof can be found in [10, p. 247].

2.2. **Some analytic tools.**

2.2.1. *Chebyshev expansions.* Let $h \in \mathcal{P}_q$ and fix $K > 0$. Then we can express

$$h(x) = \sum_{j=0}^{q} a_j T_j(K^{-1}x) \tag{2.1}$$

for some real coefficients a_j, where T_j denotes the Chebyshev polynomial of the first kind of degree j defined by $T_j(\cos\theta) = \cos(j\theta)$. The following is classical.

Lemma 2.3. *Let h be as in (2.1) and define $f(\theta) := h(K\cos(\theta))$. Then*

$$|a_0| \le \|h\|_{[-K,K]},$$

and for every $m \in \mathbb{Z}_+$

$$\sum_{j=1}^{q} j^m |a_j| \lesssim \|f^{(m+1)}\|_{[0,2\pi]}.$$

Proof. Note that a_j in (2.1) are the Fourier coefficients of f. Thus the first inequality follows from $a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(\theta)\,d\theta$. The second inequality follows as

$$\sum_{j=1}^{q} j^m |a_j| \le \left(\sum_{j=1}^{q} \frac{1}{j^2}\right)^{1/2} \left(\sum_{j=1}^{q} j^{2(m+1)} |a_j|^2\right)^{1/2} \lesssim \|f^{(m+1)}\|_{L^2[0,2\pi]}$$

by Cauchy-Schwarz and Parseval, and using that $\|g\|_{L^2[0,2\pi]} \le \sqrt{2\pi}\|g\|_{[0,2\pi]}$. □

As the bounds of Lemma 2.3 are independent of q, it follows that every $C^{m,1}$ function $h : [-K, K] \to \mathbb{R}$ has a uniformly convergent Chebyshev expansion (2.1) (with $q = \infty$). The conclusion of Lemma 2.3 extends to such h by continuity.

2.2.2. *Taylor expansions.* While Chebyshev expansions are useful for the analysis of smooth functions, Taylor expansions are often more convenient for the analysis of analytic functions due to the following standard estimate.

Lemma 2.4. *Let $f : \mathbb{C} \to \mathbb{C}$ be holomorphic in a neighborhood of $\{z \in \mathbb{C} : |z| \le r\}$. Then $f(z) = \sum_{k=0}^{\infty} a_k z^k$ is absolutely convergent for $|z| < r$ with*

$$|a_k| \le r^{-k} \max_{|z|=r} |f(z)|.$$

Proof. The conclusion follows readily by estimating the integrand in the Cauchy integral formula $a_k = \frac{1}{k!} f^{(k)}(0) = \frac{1}{2\pi i} \oint_{\{|y|=r\}} f(y)\, y^{-(k+1)}\, dy$. □

2.2.3. *Test functions.* The following nearly optimal construction of smooth test functions will be used repeatedly throughout this paper.

Lemma 2.5. *Fix $m \in \mathbb{Z}_+$ and $K, \rho, \varepsilon > 0$ so that $\rho + \varepsilon < K$. Then there exists a function $\chi : \mathbb{R} \to [0, 1]$ with the following properties.*

1. *$\chi(x) = 0$ for $|x| \le \rho + \frac{\varepsilon}{2}$, and $\chi(x) = 1$ for $|x| \ge \rho + \varepsilon$.*
2. *Let $f(\theta) := \chi(K\cos\theta)$. Then for every $k \le m$, we have*

$$\|f^{(k+1)}\|_{[0,2\pi]} \le 8^{k+1} m^k \left(\frac{K}{\varepsilon}\right)^{k+1}.$$

Proof. The result follows readily from the proof of [15, Lemma 4.10]. □

2.2.4. *Distributions.* Throughout this paper, we only consider distributions on $\mathbb{R}$. We adopt the following definition; see, e.g., [31, §2.2–2.3].

Definition 2.6. A linear functional ν on $C^\infty(\mathbb{R})$ is called a *compactly supported distribution* if there exist $C, K \geq 0$ and $m \in \mathbb{Z}_+$ so that

$$|\nu(f)| \leq C \max_{0\leq k\leq m} \|f^{(k)}\|_{[-K,K]} \quad \text{for all } f \in C^\infty(\mathbb{R}).$$

The *support* $\operatorname{supp}\nu$ of a compactly supported distribution ν is the smallest closed set $A \subseteq \mathbb{R}$ so that $\nu(f) = 0$ for all $f \in C^\infty(\mathbb{R})$ that vanish in a neighborhood of A.

The linear functionals that arise in this paper are defined *a priori* only on the space $\mathcal{P}$ of real polynomials. The following criterion enables us to extend these functionals to compactly supported distributions, cf. [15, Lemma 4.7].

Lemma 2.7. *Let ν be a linear functional on $\mathcal{P}$. If there exist $C, K, m \geq 0$ so that*

$$|\nu(h)| \leq Cq^m \|h\|_{[-K,K]} \quad \textit{for all } h \in \mathcal{P}_q,\ q \in \mathbb{N},$$

then ν extends to a compactly supported distribution with $|\nu(h)| \lesssim \|h\|_{C^{m+1}[-K,K]}$.

2.3. Random matrices and asymptotic freeness.

2.3.1. *Unitary and orthogonal invariant ensembles.* We begin by recalling the definitions of the standard complex and real Gaussian ensembles. We denote the real Gaussian distribution as $N(0,\sigma^2)$, and define the complex Gaussian distribution $N_{\mathbb{C}}(0,\sigma^2)$ as the distribution of $\xi_1 + i\xi_2$ where ξ_1, ξ_2 are i.i.d. $N(0, \frac{\sigma^2}{2})$.

Definition 2.8. Let X be an $N \times N$ self-adjoint random matrix with independent entries $(X_{ij})_{i\geq j}$ on and above the diagonal.

a. X is called a *GUE matrix* if $X_{ij} \sim N_{\mathbb{C}}(0, \frac{1}{N})$ for $i \neq j$ and $X_{ii} \sim N(0, \frac{1}{N})$.

b. X is called a *GOE matrix* if $X_{ij} \sim N(0, \frac{1}{N})$ for $i \neq j$ and $X_{ii} \sim N(0, \frac{2}{N})$.

The defining property of GUE and GOE models is that they are the Gaussian ensembles whose distributions are invariant under conjugation by unitary and orthogonal matrices, respectively. Beside the Gaussian ensembles, we will develop parallel results for random unitary and orthogonal matrices drawn from the normalized Haar measure on $\mathrm{U}(N)$ and $\mathrm{O}(N)$, respectively.

2.3.2. *Symplectic invariant ensembles.* To define the symplectic analogues of the above ensembles, we must first recall some basic facts.

We denote by $\mathbb{H}$ the skew-field of quaternions. Recall that $z \in \mathbb{H}$ is represented as $z = z_0\mathbf{1} + z_1\mathbf{i} + z_2\mathbf{j} + z_3\mathbf{k}$, where $z_i \in \mathbb{R}$ and $\mathbf{i}, \mathbf{j}, \mathbf{k}$ satisfy the relations

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1}.$$

The conjugate is $\bar{z} = z_0\mathbf{1} - z_1\mathbf{i} - z_2\mathbf{j} - z_3\mathbf{k}$, and the real part is $\operatorname{Re} z = z_0$.

For a quaternionic matrix $A \in \mathrm{M}_N(\mathbb{H})$, the adjoint A^* is defined as the conjugate transpose as for complex matrices. We denote by

$$\mathrm{Sp}(N) := \{U \in \mathrm{M}_N(\mathbb{H}) : UU^* = U^*U = \mathbf{1}\}$$

the group of $N \times N$ symplectic (i.e., quaternionic unitary) matrices.

We define the quaternionic Gaussian distribution $N_{\mathbb{H}}(0, \sigma^2)$ as the distribution of $\xi_0 \mathbf{1} + \xi_1 \mathbf{i} + \xi_2 \mathbf{j} + \xi_3 \mathbf{k}$ where $\xi_0, \ldots, \xi_3$ are i.i.d. $N(0, \frac{\sigma^2}{4})$. We can now recall the definition of the standard quaternionic Gaussian ensemble.

Definition 2.9. An $N \times N$ self-adjoint random matrix with independent entries $(X_{ij})_{i \geq j}$ is called a *GSE matrix* if $X_{ij} \sim N_{\mathbb{H}}(0, \frac{1}{N})$ for $i \neq j$ and $X_{ii} \sim N(0, \frac{1}{2N})$.

The defining property of the GSE model is that it is the Gaussian ensemble whose distribution is invariant under conjugation by symplectic matrices. We will develop parallel results for random symplectic matrices drawn from the normalized Haar measure on the compact group $\mathrm{Sp}(N)$.

For the purposes of linear algebra, working directly with quaternions is somewhat awkward; for example, we cannot apply noncommutative polynomials with complex coefficients to them, as the quaternions form an algebra over the reals. Instead, we will identify $\mathbb{H}$ with the subring of $\mathrm{M}_2(\mathbb{C})$ generated by

$$\mathbf{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{i} = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, \qquad \mathbf{j} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \qquad \mathbf{k} = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$

In this manner, $\mathrm{M}_N(\mathbb{H})$ is naturally identified with a subring of $\mathrm{M}_{2N}(\mathbb{C})$. In this paper, we will always interpret linear algebra operations on $M \in \mathrm{M}_N(\mathbb{H})$ as being applied to the associated complex representations; for example, $\operatorname{Tr} M$ will denote the trace of the $2N$-dimensional complex representation of M.

2.3.3. *Asymptotic freeness.* For the purposes of this paper, a C^*-probability space $(\mathcal{A}, \tau)$ is defined by a unital C^*-algebra $\mathcal{A}$ and a faithful trace τ.

Definition 2.10. Let $(\mathcal{A}, \tau)$ be a C^*-probability space.

a. $s_1, \ldots, s_r \in \mathcal{A}$ form a *free semicircular family* if the spectral distribution of each s_i is the standard semicircle distribution, and $s_1, \ldots, s_r$ are freely independent.
b. $u_1, \ldots, u_r \in \mathcal{A}$ are *free Haar unitaries* if the spectral distribution of each u_i is uniformly distributed on the unit circle, and $u_1, \ldots, u_r$ are freely independent.

We do not recall here the definition of free independence, which is discussed in detail in the excellent text [44]. The significance of the above definition is that it provides a limiting model as $N \to \infty$ for many random matrix models.

Lemma 2.11 (Weak asymptotic freeness)**.** *Let $\boldsymbol{G}^N = (G_1^N, \ldots, G_r^N)$ be i.i.d. GUE/GOE/GSE random matrices of dimension N, and let $\boldsymbol{s} = (s_1, \ldots, s_r)$ be a free semicircular family. Then*

$$\lim_{N \to \infty} \mathbf{E}[\operatorname{tr} P(\boldsymbol{G}^N)] = (\operatorname{tr} \otimes \tau)[P(\boldsymbol{s})]$$

for every noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{s} \rangle$.

Similarly, let $\boldsymbol{U}^N = (U_1^N, \ldots, U_r^N)$ be i.i.d. Haar-distributed random matrices in $\mathrm{U}(N)/\mathrm{O}(N)/\mathrm{Sp}(N)$, and let $\boldsymbol{u} = (u_1, \ldots, u_r)$ be free Haar unitaries. Then

$$\lim_{N \to \infty} \mathbf{E}[\operatorname{tr} P(\boldsymbol{U}^N, \boldsymbol{U}^{N*})] = (\operatorname{tr} \otimes \tau)[P(\boldsymbol{u}, \boldsymbol{u}^*)]$$

for every noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{u}, \boldsymbol{u}^* \rangle$.

Lemma 2.11 is a special case of a celebrated result of Voiculescu [55] (Voiculescu does not consider the symplectic ensembles, but the proofs are entirely analogous).

3. Polynomial interpolation from $\frac{1}{N}$ samples

It is classical [22] that for $h \in \mathcal{P}_q$, we have $\|h\|_{[0,\delta]} \lesssim \max_{x\in I} |h(x)|$ for any set $I \subseteq [0,\delta]$ with spacing at most $O(\frac{\delta}{q^2})$ between its points; this is optimal for general sets I. When applied to the set $I_M := \{\frac{1}{N} : N \geq M\}$ that arises in random matrix problems, this enables us to bound $\|h\|_{[0,\delta]}$ only for $\delta = O(\frac{1}{q^2})$. The aim of this section is to prove that a much better bound can be achieved in this case.

Proposition 3.1 (Interpolation from $\frac{1}{N}$ samples)**.** *We have*

$$\|h\|_{[0,\delta]} \leq C \sup_{\frac{1}{N}\leq 2\delta} |h(\tfrac{1}{N})|$$

for every $q \in \mathbb{N}$, $h \in \mathcal{P}_q$, and $0 \leq \delta \leq \frac{1}{24q}$, where C is a universal constant.

This is optimal up to the values of the constants.

Example 3.2 (Optimality)**.** Let $h_q(x) = T_q(qx)\prod_{j=1}^q(1-jx)$, where T_q is the Chebyshev polynomial of degree q. Then $h_q \in \mathcal{P}_{2q}$ and $|h_q(\frac{1}{N})| \leq 1$ for all $N \geq 1$. On the other hand, for $x = \frac{1}{m+1/2}$ with $m \in \{1,\dots,q-1\}$, we can estimate

$$|h_q(x)| = \frac{\left|T_q\left(\frac{q}{m+1/2}\right)\right|}{(4m+2)^q}\frac{(2q)!}{q!}\frac{\binom{q}{m}}{\binom{2q}{2m}} \geq \left(\frac{Cq}{m}\right)^q$$

for a universal constant C, where we used $|T_q(x)| \geq \frac{1}{2}x^q$ for $x \geq 1$. Thus there is a universal constant c so that $\|h_q\|_{[0,\frac{c}{q}]} \geq 2^q$ for all q. This shows that the conclusion of Proposition 3.1 must fail if the assumption $\delta \leq \frac{1}{24q}$ is replaced by $\delta \leq \frac{2c}{q}$.

Proposition 3.1 is based on a powerful result of Rakhmanov [51]: if $h \in \mathcal{P}_q$ is bounded at equispaced points I in the interval $[-1,1]$, then $O(\frac{1}{q})$ spacing suffices to achieve a uniform bound on h strictly in the interior of the interval, even though $O(\frac{1}{q^2})$ spacing is necessary for a uniform bound in the entire interval. The distinct behavior in the interior and near the edges of the interval is reminiscent of the distinction between the Bernstein and Markov inequalities.

Theorem 3.3 (Rakhmanov)**.** *Let $M, q \in \mathbb{N}$ with $q \leq M$, and let $h \in \mathcal{P}_q$. Then*

$$\|h\|_{[-\frac{1}{2},\frac{1}{2}]} \leq C \max_{k=1,\dots,2M} \left|h\left(-1+\tfrac{2k-1}{2M}\right)\right|,$$

where C is a universal constant.

Proof. Apply [51, eq. (1.6)] with $\delta = 1$ and $N = 2M$. □

The setting of Proposition 3.1 differs considerably from that of Rakhmanov's theorem: the $\frac{1}{N}$ samples are highly nonuniform in the interval $[0,\delta]$, while we aim to bound h near the endpoint 0 of the interval rather than strictly in its interior. However, the fact that the $\frac{1}{N}$ samples become increasingly dense near 0 will enable us to apply Rakhmanov's theorem in a multiscale manner: we can cover $[0,\delta]$ by a sequence of intervals so that the $\frac{1}{N}$ samples in each interval are approximately uniform, and apply Theorem 3.3 to each interval.

The main difficulty in the proof is that the samples in each interval must be mapped to equispaced samples in order to apply Theorem 3.3. Concretely, suppose we aim to bound $h \in \mathcal{P}_q$ based on its values at the $2M$ sample points

$$\left\{\tfrac{1}{2M+2k-1} : k = 1, \ldots, 2M\right\}.$$

As we have

$$h\big(\tfrac{1}{2M+2k-1}\big) = r\big(-1 + \tfrac{2k-1}{2M}\big) \quad \text{with} \quad r(x) := h\big(\tfrac{1}{2M(2+x)}\big),$$

the problem is equivalent to bounding the rational function r at equispaced points as in Theorem 3.3. However r is no longer a polynomial, so that Theorem 3.3 does not apply. To surmount this issue, we will use that r can be approximated by a polynomial while only losing a constant factor in its magnitude and degree.

Lemma 3.4. *Let $M, q \in \mathbb{N}$, let $h \in \mathcal{P}_q$, and define $r(x) := h\big(\frac{1}{2M(2+x)}\big)$. Then there exists a polynomial $g \in \mathcal{P}_{8q}$ so that*

$$\tfrac{4}{7}|g(x)| \le |r(x)| \le 4|g(x)| \quad \textit{for all } x \in [-1, 1].$$

Proof. We can clearly write $r(x) = \frac{u(x)}{(2+x)^q}$ for a polynomial $u \in \mathcal{P}_q$. Note that

$$\big|\tfrac{1}{(2+z)^q}\big| \le 2^q \quad \text{for all } z \in \mathbb{C},\ |z| \le \tfrac{3}{2}.$$

Thus $\frac{1}{(2+z)^q} = \sum_{k=0}^{\infty} a_k z^k$ with $|a_k| \le 2^q(\frac{3}{2})^{-k}$ by Lemma 2.4. Therefore

$$\big|\tfrac{1}{(2+x)^q} - t(x)\big| \le \sum_{k=7q+1}^{\infty} |a_k| \le 4^{-q} \le \tfrac{3}{4}\tfrac{1}{(2+x)^q} \quad \text{for } x \in [-1, 1],$$

where we defined $t(x) := \sum_{k=0}^{7q} a_k x^k$ and we used that $3^{-q} \le \frac{1}{(2+x)^q}$ for $|x| \le 1$. In particular, we have shown that $\frac{4}{7}t(x) \le \frac{1}{(2+x)^q} \le 4t(x)$ for all $x \in [-1, 1]$, and the conclusion follows readily by choosing $g(x) = u(x)t(x)$. □

We can now complete the proof of Proposition 3.1.

Proof of Proposition 3.1. Fix $q \in \mathbb{N}$ and $h \in \mathcal{P}_q$ throughout the proof. We first apply Theorem 3.3 to the polynomial g in Lemma 3.4 to estimate

$$\|h\|_{[\frac{1}{5M}, \frac{1}{3M}]} \le C \max_{k=1,\ldots,2M} \big|h\big(\tfrac{1}{2M+2k-1}\big)\big|$$

for any $M \in \mathbb{N}$ with $M \ge 8q$, where C is a universal constant. Now let $m = \lfloor \frac{1}{3\delta} \rfloor$. Then $m \ge 8q$ by the assumption $\delta \le \frac{1}{24q}$, and it is readily verified that

$$(0, \delta] \subseteq (0, \tfrac{1}{3m}] = \bigcup_{M \ge m} [\tfrac{1}{5M}, \tfrac{1}{3M}]$$

as there are no gaps between the intervals $[\frac{1}{5M}, \frac{1}{3M}]$ for $M \ge 2$. We therefore obtain

$$\|h\|_{[0,\delta]} \le \sup_{M \ge m} \|h\|_{[\frac{1}{5M}, \frac{1}{3M}]} \le C \sup_{N \ge 2m+1} \big|h\big(\tfrac{1}{N}\big)\big|,$$

and the conclusion follows as $\frac{1}{2m+1} \le 2\delta$ (because $2m + 1 \ge \frac{2}{3\delta} - 1 \ge \frac{1}{2\delta}$). □

4. Asymptotic expansion for GUE

The aim of this section is to establish an asymptotic expansion of smooth trace statistics of polynomials of GUE matrices. This expansion will be used in section 5 to prove Theorem 1.1 in the GUE case, while the requisite modifications in the case of GOE/GSE matrices will be developed in section 6.

The following will be fixed throughout this section. Let $\boldsymbol{G}^N = (G_1^N, \ldots, G_r^N)$ be independent GUE matrices of dimension N, and let $\boldsymbol{s} = (s_1, \ldots, s_r)$ be a free semicircular family. We will further fix a self-adjoint noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle x_1, \ldots, x_r\rangle$ of degree q_0 with matrix coefficients of dimension D. For simplicity of notation, we will denote by

$$X^N := P(\boldsymbol{G}^N), \qquad X_{\mathrm{F}} := P(\boldsymbol{s})$$

the random matrix of interest and its limiting model.

The main result of this section is as follows.

Theorem 4.1 (Smooth asymptotic expansion for GUE)**.** *There exist universal constants $C, c > 0$, and a compactly supported distribution ν_k for every $k \in \mathbb{Z}_+$, such that the following hold. Fix any bounded $h \in C^\infty(\mathbb{R})$, and define*

$$f(\theta) := h(K\cos\theta) \qquad \textit{with} \qquad K := (Cr)^{q_0}\|X_{\mathrm{F}}\|.$$

Then for every $m, N \in \mathbb{N}$ with $m \le \frac{N}{2}$, we have

$$\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1} \frac{\nu_k(h)}{N^k}\right| \le \frac{(Cq_0)^{2m}}{m!N^m}\|f^{(2m+1)}\|_{[0,2\pi]} + Cre^{-cN}\big(\|h\|_{(-\infty,\infty)} + \|f^{(1)}\|_{[0,2\pi]}\big).$$

Remark 4.2. The conclusion of Theorem 4.1 extends readily by continuity to any test function $h \in C_b(\mathbb{R})$ so that $\|f^{(2m+1)}\|_{[0,2\pi]} < \infty$. In particular, the proof of Theorem 1.1 will apply this theorem to the test functions provided by Lemma 2.5. This observation will be used without further comment in the sequel.

The remainder of this section is devoted to the proof of Theorem 4.1.

4.1. A priori bounds.

The general approach to Theorem 4.1 follows the basic method outlined in section 1.2.1. However, for the Gaussian ensembles, a significant complication arises from the unboundedness of the Gaussian distribution. To surmount this issue, we begin by proving a priori bounds that will be used to truncate the model. The challenge in the remainder of the proof will be to apply these bounds without incurring any quantitative loss.

Lemma 4.3 (A priori bounds)**.** *There exist universal constants $C, c > 0$ such that*

$$\mathbf{P}[\|X^N\| > K] \le Cre^{-cN}, \tag{4.1}$$

where $K := (Cr)^{q_0}\|X_{\mathrm{F}}\|$. Moreover, we have

$$|\mathbf{E}[\operatorname{tr} h(X^N)]| \le 2\|h\|_{[-K,K]} \tag{4.2}$$

and

$$|\mathbf{E}[\operatorname{tr} h(X^N)\cdot 1_{\{\|X^N\|>K\}}]| \le C\sqrt{r}e^{-cN}\|h\|_{[-K,K]} \tag{4.3}$$

for every $h \in \mathcal{P}_q$ *with* $q \le \frac{N}{q_0}$.

The remainder of this section is devoted to the proof of this result. We begin by recalling a crude tail bound on the norm of GUE matrices.

Lemma 4.4. *There exist universal constants* $C, c, \kappa > 0$ *such that for any GUE matrix* G^N *of dimension* N*, we have*

$$\mathbf{P}[\|G^N\| \ge \kappa + t] \le Ce^{-cNt^2} \quad \text{for all } t \ge 0,$$

and

$$\mathbf{E}[\|G^N\|^p] \le \left(\kappa + C\sqrt{\frac{p}{N}}\right)^p \quad \text{for all } p \in \mathbb{N}.$$

Proof. This first inequality follows from a simple ε-net argument [54, §2.3.1]. The second inequality follows directly using $\mathbf{E}[\|G^N\|^p]^{\frac{1}{p}} \le \kappa + \mathbf{E}[(\|G^N\| - \kappa)_+^p]^{\frac{1}{p}}$ and integrating the first inequality (see, e.g., [35, Proposition 1.10]). □

To proceed, it will be useful to choose a convenient representation of the noncommutative polynomial P. To this end, denote by U_j the Chebyshev polynomial of the second kind of degree j defined by $U_j(\cos\theta)\sin\theta = \sin((j+1)\theta)$. Moreover, define the noncommutative polynomial $U_{\boldsymbol{i},\boldsymbol{j}} \in \mathbb{C}\langle x_1, \ldots, x_r\rangle$ as

$$U_{\boldsymbol{i},\boldsymbol{j}}(x_1, \ldots, x_r) := U_{j_1}(\tfrac{1}{2}x_{i_1})U_{j_2}(\tfrac{1}{2}x_{i_2})\cdots U_{j_k}(\tfrac{1}{2}x_{i_k})$$

for every $k \ge 0$, $\boldsymbol{i} = (i_1, \ldots, i_k)$, and $\boldsymbol{j} = (j_1, \ldots, j_k)$ such that $j_1, \ldots, j_k \in \mathbb{N}$ and $i_1, \ldots, i_k \in [r]$ with $i_1 \ne i_2, i_2 \ne i_3, \ldots, i_{k-1} \ne i_k$ (where $U_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{x}) := \mathbf{1}$ for $k = 0$). Then we can represent P uniquely as

$$P(x_1, \ldots, x_r) = \sum_{\boldsymbol{i},\boldsymbol{j}} A_{\boldsymbol{i},\boldsymbol{j}} \otimes U_{\boldsymbol{i},\boldsymbol{j}}(x_1, \ldots, x_r),$$

where $A_{\boldsymbol{i},\boldsymbol{j}} \in \mathrm{M}_D(\mathbb{C})$ are matrix coefficients and the sum ranges over all $\boldsymbol{i}, \boldsymbol{j}$ as above with $0 \le k \le q_0$ and $j_1 + \cdots + j_k \le q_0$. The significance of this representation is that when $U_{\boldsymbol{i},\boldsymbol{j}}$ are applied to a free semicircular family $\boldsymbol{s}$, the operators $\{U_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{s})\}$ form an orthonormal system in $L^2(\tau)$, cf. [5, §5.1]. The latter enables us to bound the norms of the coefficients $\|A_{\boldsymbol{i},\boldsymbol{j}}\|$ by $\|X_{\mathrm{F}}\|$.

Lemma 4.5. *For every* $\boldsymbol{i}, \boldsymbol{j}$ *as above and self-adjoint operators* $x_1, \ldots, x_r$*, we have*

$$\|U_{\boldsymbol{i},\boldsymbol{j}}(x_1, \ldots, x_r)\| \le 2^k(\|x_{i_1}\|^{j_1} \vee 1)(\|x_{i_2}\|^{j_2} \vee 1)\cdots(\|x_{i_k}\|^{j_k} \vee 1).$$

Moreover, there is a universal constant C *so that*

$$\sum_{\boldsymbol{i},\boldsymbol{j}} \|A_{\boldsymbol{i},\boldsymbol{j}}\| \le (Cr)^{q_0}\|X_{\mathrm{F}}\|.$$

Proof. Note that $|T_j(\frac{1}{2}x)| \le |x|^j \vee 1$ for all $x \in \mathbb{R}$ by Lemma 2.2, where T_j is the Chebyshev polynomial of the first kind. As $U_j(x) = \sum_{a=0}^j x^a T_{j-a}(x)$ [10, p. 37],

$$|U_j(\tfrac{1}{2}x)| \le \sum_{a=0}^j 2^{-a}(|x|^j \vee |x|^a) \le 2(|x|^j \vee 1)$$

for all $x \in \mathbb{R}$. The first inequality follows directly.

Next, note that as $X_{\rm F} := P(\boldsymbol{s})$ and $\{U_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{s})\}$ are orthonormal in $L^2(\tau)$, we have

$$\|X_{\rm F}\|^2 \ge \|(\mathrm{id}\otimes\tau)(X_{\rm F}^* X_{\rm F})\| = \left\|\sum_{\boldsymbol{i},\boldsymbol{j}} A_{\boldsymbol{i},\boldsymbol{j}}^* A_{\boldsymbol{i},\boldsymbol{j}}\right\| \ge \max_{\boldsymbol{i},\boldsymbol{j}} \|A_{\boldsymbol{i},\boldsymbol{j}}\|^2.$$

The second inequality follows as there are at most $(Cr)^{q_0}$ terms in the sum. □

We are now ready to prove Lemma 4.3.

Proof of Lemma 4.3. If $\|G_i^N\| \le \kappa+1$ for all $i\in[r]$, then we can estimate

$$\|X^N\| \le 2^{q_0}(\kappa+1)^{q_0}\sum_{\boldsymbol{i},\boldsymbol{j}}\|A_{\boldsymbol{i},\boldsymbol{j}}\| \le (Cr)^{q_0}\|X_{\rm F}\| =: K$$

for a universal constant C by the triangle inequality and Lemma 4.5. As

$$\mathbf{P}[\|G_i^N\| > \kappa+1 \text{ for some } i\in[r]] \le Cre^{-cN}$$

by the union bound and the first inequality of Lemma 4.4, we proved (4.1).

To proceed, fix $h\in\mathcal{P}_q$ with $q\le\frac{2N}{q_0}$. Note first that

$$\max_{\boldsymbol{i},\boldsymbol{j}}\mathbf{E}[\|U_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{G}^N)\|^q] \le 2^{qq_0}\mathbf{E}[\|G^N\|^{qq_0}\vee 1] \le C_1^{qq_0}$$

for a universal constant C_1 using Lemma 4.5, Hölder's inequality, and the second inequality of Lemma 4.4 (here we used that $q\le\frac{2N}{q_0}$). Now define the constant $L := 2C_1^{q_0}\sum_{\boldsymbol{i},\boldsymbol{j}}\|A_{\boldsymbol{i},\boldsymbol{j}}\|$, and apply Lemma 2.2 to estimate

$$|\mathbf{E}[\mathrm{tr}\,h(X^N)]| \le \mathbf{E}\left[\sup_{|x|\le\|X^N\|}|h(x)|\right] \le \left(1+\mathbf{E}\left[\left(\frac{2\|X^N\|}{L}\right)^q\right]\right)\|h\|_{[-L,L]}.$$

As

$$\mathbf{E}\left[\left(\frac{2\|X^N\|}{L}\right)^q\right] \le \frac{1}{C_1^{qq_0}}\,\mathbf{E}\left[\left(\frac{\sum_{\boldsymbol{i},\boldsymbol{j}}\|A_{\boldsymbol{i},\boldsymbol{j}}\|\,\|U_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{G}^N)\|}{\sum_{\boldsymbol{i},\boldsymbol{j}}\|A_{\boldsymbol{i},\boldsymbol{j}}\|}\right)^q\right] \le 1$$

by Jensen's inequality, we have shown that $|\mathbf{E}[\mathrm{tr}\,h(X^N)]|\le 2\|h\|_{[-L,L]}$. To prove (4.2), it remains to note that $L\le K$ by Lemma 4.5, provided that the universal constant C in the definition of K is chosen sufficiently large.

Finally, we now suppose $h\in\mathcal{P}_q$ and $q\le\frac{N}{q_0}$. Then we can estimate

$$\begin{aligned}|\mathbf{E}[\mathrm{tr}\,h(X^N)\cdot 1_{\{\|X^N\|>K\}}]| &\le \mathbf{E}[\mathrm{tr}\;h(X^N)^2]^{\frac12}\,\mathbf{P}[\|X^N\|>K]^{\frac12}\\ &\le \sqrt{2Cr}\,e^{-cN/2}\|h\|_{[-K,K]}\end{aligned}$$

by Cauchy-Schwarz, (4.1), and (4.2), where we used that $h^2\in\mathcal{P}_{2q}$ and $2q\le\frac{2N}{q_0}$. Redefining the constants concludes the proof of (4.3). □

4.2. **The master inequality.** Our next aim is to prove a form of the asymptotic expansion of Theorem 4.1 for polynomial test functions h.

Lemma 4.6 (Master inequality)**.** *There exists a linear functional ν_m on $\mathcal{P}$ for every $m\in\mathbb{Z}_+$ such that for every $q\in\mathbb{N}$ and $h\in\mathcal{P}_q$*

$$|\nu_m(h)| \le \frac{(Cqq_0)^{2m}}{m!}\|h\|_{[-K,K]}, \tag{4.4}$$

and such that if in addition $q \le \frac{N}{Cq_0}$*, we have*

$$\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1} \frac{\nu_k(h)}{N^k}\right| \le \frac{(Cqq_0)^{2m}}{m!N^m}\|h\|_{[-K,K]}. \tag{4.5}$$

Here C *is a universal constant and* $K := (Cr)^{q_0}\|X_{\mathrm{F}}\|$.

The proof is a straightforward application of the polynomial method of [15]. We exploit the well-known fact that $\mathbf{E}[\operatorname{tr} h(X^N)]$ is a polynomial of $\frac{1}{N}$.

Lemma 4.7 (Polynomial encoding)**.** *For any* $h \in \mathcal{P}_q$*, there is* $\Phi_h \in \mathcal{P}_{qq_0}$ *so that*

$$\mathbf{E}[\operatorname{tr}(h(X^N))] = \Phi_h(\tfrac{1}{N}) = \Phi_h(-\tfrac{1}{N}).$$

Proof. This is an immediate consequence of the genus expansion for GUE, which states that $\mathbf{E}[\operatorname{tr} G^N_{i_1}\cdots G^N_{i_k}]$ is a polynomial of $\frac{1}{N^2}$ of degree at most $\frac{k}{4}$ for every $i_1,\ldots,i_k \in [r]$; for example, see the proof of [42, §1.10, Lemma 9]. □

We can now prove Lemma 4.6.

Proof of Lemma 4.6. Fix $q \in \mathbb{N}$ and $h \in \mathcal{P}_q$. Throughout the proof, we adopt the notation of Lemma 4.7 without further comment. Lemma 4.3 implies that

$$|\Phi_h(\tfrac{1}{N})| \le 2\|h\|_{[-K,K]} \quad \text{for } N \ge qq_0.$$

Thus Proposition 3.1 yields

$$\|\Phi_h\|_{[0,\delta]} \le C\|h\|_{[-K,K]}$$

with $\delta := \frac{1}{24qq_0}$. Since $\Phi_h(-\frac{1}{N}) = \Phi_h(\frac{1}{N})$, we obtain the same bound for $\|\Phi_h\|_{[-\delta,0]}$. We can therefore apply Bernstein's inequality (Lemma 2.1) to estimate

$$\|\Phi_h^{(m)}\|_{[-\frac{\delta}{2},\frac{\delta}{2}]} \le (Cqq_0)^{2m}\|h\|_{[-K,K]} \tag{4.6}$$

for all $m \in \mathbb{Z}_+$, where C is a universal constant. Now define

$$\nu_m(h) := \frac{\Phi_h^{(m)}(0)}{m!},$$

so that

$$\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1} \frac{\nu_k(h)}{N^k}\right| = \left|\Phi_h(\tfrac{1}{N}) - \sum_{k=0}^{m-1} \frac{\Phi_h^{(k)}(0)}{k!N^k}\right| \le \frac{\|\Phi^{(m)}\|_{[0,\frac{\delta}{2}]}}{m!N^m}$$

whenever $\frac{1}{N} \le \frac{\delta}{2} = \frac{1}{48qq_0}$ by Taylor's theorem. Both parts of the lemma now follow immediately from (4.6), concluding the proof. □

4.3. **Extension to smooth functions.** To complete the proof of Theorem 4.1, it remains to extend the expansion of Lemma 4.6 from polynomial to smooth test functions h. The difficulty here is that, unlike in [15], the expansion (4.5) cannot hold for arbitrarily large $q \in \mathbb{N}$ due to the unboundedness of the Gaussian distribution. To surmount this problem, we must provide a separate treatment of the low- and high-degree terms in the Chebyshev expansion of h.

Proof of Theorem 4.1. We first note that the linear functional ν_m of Lemma 4.6 extends to a compactly supported distribution with $|\nu_m(h)| \lesssim \|h\|_{C^{2m+1}[-K,K]}$ for every $m \in \mathbb{Z}_+$ by Lemma 2.7 and the inequality (4.4).

In the following, we fix any bounded $h \in C^\infty(\mathbb{R})$ and denote by

$$h(x) = \sum_{q=0}^{\infty} a_q\, T_q(K^{-1}x) \quad \text{for } x \in [-K,K] \tag{4.7}$$

its Chebyshev expansion on the interval $[-K,K]$ (cf. section 2.2.1). Note that as $\mathbf{E}[\operatorname{tr} h(X^N)] = c$ is independent of N whenever $h \equiv c$ is a constant function, we have $\nu_0(c) = c$ and $\nu_k(c) = 0$ for all $k \geq 1$. This implies that the theorem statement is invariant under the replacement $h \leftarrow h - a_0$. We will therefore assume without loss of generality in the rest of the proof that $a_0 = 0$.

Let $B \leq \frac{N}{q_0}$ be the largest integer q for which (4.5) has been established, and let

$$h_0(x) := \sum_{q=1}^{B} a_q\, T_q(K^{-1}x).$$

Then we can estimate

$$\begin{aligned}\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1} \frac{\nu_k(h)}{N^k}\right| &\leq \left|\mathbf{E}[\operatorname{tr} h_0(X^N)] - \sum_{k=0}^{m-1} \frac{\nu_k(h_0)}{N^k}\right| \\ &\quad + \sum_{k=0}^{m-1} \frac{|\nu_k(h-h_0)|}{N^k} + |\mathbf{E}[\operatorname{tr}(h(X^N) - h_0(X^N))]|.\end{aligned}$$

We now bound each term on the right-hand side.

First term. Using (4.5), we readily estimate

$$\begin{aligned}\left|\mathbf{E}[\operatorname{tr} h_0(X^N)] - \sum_{k=0}^{m-1} \frac{\nu_k(h_0)}{N^k}\right| &\leq \sum_{q=1}^{B} |a_q| \left|\mathbf{E}[\operatorname{tr} T_q(K^{-1}X^N)] - \sum_{k=0}^{m-1} \frac{\nu_k(T_q(K^{-1}x))}{N^k}\right| \\ &\leq \frac{(Cq_0)^{2m}}{m!N^m} \sum_{q=1}^{B} q^{2m}|a_q|.\end{aligned}$$

Second term. Because $|\nu_k(h)| \lesssim \|h\|_{C^{2k+1}[-K,K]}$ for all k, we are able to substitute the Chebyshev expansion (4.7) into $\nu_k(h-h_0)$. This yields

$$\sum_{k=0}^{m-1} \frac{|\nu_k(h-h_0)|}{N^k} \leq \sum_{k=0}^{m-1} \frac{1}{N^k} \sum_{q>B} |a_q|\, |\nu_k(T_q(K^{-1}x))| \leq \sum_{k=0}^{m-1} \frac{(Cq_0)^{2k}}{k!N^k} \sum_{q>B} q^{2k}|a_q|$$

using (4.4). Now note that $q^{2k} \leq \frac{q^{2m}}{B^{2(m-k)}} \leq (\frac{Cq_0}{N})^{2(m-k)} q^{2m}$ for any $k < m$ and $q > B$, where we used that $B \gtrsim \frac{N}{q_0}$ by Lemma 4.6. We therefore obtain

$$\sum_{k=0}^{m-1} \frac{|\nu_k(h-h_0)|}{N^k} \leq \left(\sum_{k=0}^{m-1} \frac{1}{k!N^{m-k}}\right) \frac{(Cq_0)^{2m}}{N^m} \sum_{q>B} q^{2m}|a_q|.$$

Third term. We estimate

$$\begin{aligned}|\mathbf{E}[\mathrm{tr}(h(X^N) - h_0(X^N))]| &\le |\mathbf{E}[\mathrm{tr}(h(X^N) - h_0(X^N)) \cdot 1_{\{\|X^N\| \le K\}}]| \\ &\quad + |\mathbf{E}[\mathrm{tr}\, h(X^N) \cdot 1_{\{\|X^N\| > K\}}]| + |\mathbf{E}[\mathrm{tr}\, h_0(X^N) \cdot 1_{\{\|X^N\| > K\}}]| \\ &\le \frac{(Cq_0)^{2m}}{N^{2m}} \sum_{q>B} q^{2m}|a_q| + Cre^{-cN}\left(\|h\|_{(-\infty,\infty)} + \sum_{q\le B}|a_q|\right)\end{aligned}$$

using (4.7) and $1 \le (\frac{Cq_0}{N})^{2m} q^{2m}$, (4.1), and (4.3), respectively.

Combining the above estimates yields

$$\begin{aligned}&\left|\mathbf{E}[\mathrm{tr}\, h(X^N)] - \sum_{k=0}^{m-1} \frac{\nu_k(h)}{N^k}\right| \\ &\quad \le \left(\sum_{k=0}^{m} \frac{1}{k!N^{m-k}}\right)\frac{(Cq_0)^{2m}}{N^m} \sum_{q=1}^{\infty} q^{2m}|a_q| + Cre^{-cN}\left(\|h\|_{(-\infty,\infty)} + \sum_{q=1}^{\infty}|a_q|\right).\end{aligned}$$

To conclude the proof, it remains to apply Lemma 2.3 and to note that we can estimate $\sum_{k=0}^m \frac{1}{k!N^{m-k}} \le \frac{1}{m!}\sum_{k=0}^m (\frac{m}{N})^{m-k} \le \frac{2}{m!}$ for $m \le \frac{N}{2}$. □

5. Strong convergence for GUE

The aim of this section is to complete the proof of Theorem 1.1 for the case that $\boldsymbol{G}^N = (G_1^N, \ldots, G_r^N)$ are GUE matrices. With the asymptotic expansion of Theorem 4.1 in hand, it remains to control the supports of the infinitesimal distributions ν_k as in (1.3). To this end, we will first prove a basic form of strong convergence in section 5.1, which only requires control of ν_1. We will then apply a bootstrapping argument in section 5.2 to extend the conclusion to all ν_k. Finally, we complete the proof of Theorem 1.1 for GUE in section 5.3.

Throughout this section, we will always assume without further comment that the setting and notations of section 4 are in force.

5.1. Strong convergence.

Establishing strong convergence for GUE matrices is especially simple due to the following observation.

Lemma 5.1. *In the setting of Theorem 4.1, we have*

$$\nu_0(h) = (\mathrm{tr}\otimes\tau)(h(X_{\mathrm{F}})) \quad \text{and} \quad \nu_1(h) = 0 \quad \text{for all } h \in C^\infty(\mathbb{R}).$$

Proof. It suffices to consider $h \in \mathcal{P}$. Then Lemma 2.11 yields

$$\nu_0(h) = \lim_{N\to\infty} \mathbf{E}[\mathrm{tr}\, h(X^N)] = (\mathrm{tr}\otimes\tau)(h(X_{\mathrm{F}})).$$

Now note that as $\Phi_h(-\frac{1}{N}) = \Phi_h(\frac{1}{N})$ in Lemma 4.7, the polynomial $\Phi_h(x)$ can contain only monomials of even degree. Thus $\nu_1(h) = \Phi_h'(0) = 0$. □

This yields the following.

Corollary 5.2 (Strong convergence)**.** *$\|X^N\| \to \|X_{\mathrm{F}}\|$ in probability as $N \to \infty$.*

Proof. Fix $\varepsilon > 0$ sufficiently small, and let h be the test function of Lemma 2.5 with $m = 4$, $K = (Cr)^{q_0}\|X_{\rm F}\|$, and $\rho = \|X_{\rm F}\|$. Then $\nu_0(h) = \nu_1(h) = 0$ by Lemma 5.1 and as h vanishes on $[-\|X_{\rm F}\| - \frac{\varepsilon}{2}, \|X_{\rm F}\| + \frac{\varepsilon}{2}]$. Thus Theorem 4.1 with $m = 2$ yields

$$\mathbf{P}[\|X^N\| > \|X_{\rm F}\| + \varepsilon] \le \mathbf{E}[\operatorname{Tr} h(X^N)] = DN\,\mathbf{E}[\operatorname{tr} h(X^N)] = O\left(\frac{D}{N}\right),$$

where we used that $h(x) \in [0,1]$ for all x and $h(x) = 1$ for $|x| > \|X_{\rm F}\| + \varepsilon$ in the first inequality. As ε may be chosen arbitrarily small, the conclusion follows. □

5.2. **Bootstrapping.** While Corollary 5.2 has been stated in a qualitative form, its proof shows that strong convergence remains valid for sequences of noncommutative polynomials P with matrix coefficients of dimension $D = o(N)$. The information obtained so far does not suffice, however, to capture coefficients of exponential dimension; to this end, the higher-order infinitesimal distributions ν_m must be controlled as well. The aim of this section is to prove the following.

Proposition 5.3. *In the setting of Theorem 4.1, we have*

$$\operatorname{supp}\nu_m \subseteq [-\|X_{\rm F}\|, \|X_{\rm F}\|] \quad \textit{for all } m \in \mathbb{Z}_+.$$

As was explained in section 1.2.4, we will prove this theorem by combining Corollary 5.2 with concentration of measure. This enables a bootstrapping argument that uses only information on ν_0, ν_1 in Lemma 5.1 to achieve control of ν_k for all k.

Remark 5.4. Proposition 5.3 for GUE recovers a result of Parraud [47] in the special case $D = 1$ (that is, without matrix coefficients). The argument given here holds for any $D \in \mathbb{N}$ and will extend almost verbatim beyond the GUE case.

5.2.1. *Concentration of measure.* Before we develop the bootstrapping argument, we recall the requisite concentration property in a convenient form for our purposes. Similar results are well known, see, e.g., [50, Lemma 7.6]. We include the proof as it is short and carries over to GOE and GSE matrices without any changes. Here and in the sequel, $\operatorname{med}(Z)$ denotes the median of a random variable Z.

Lemma 5.5. *For any $\varepsilon > 0$ and $N \in \mathbb{N}$, we have*

$$\mathbf{P}[|\|X^N\| - \operatorname{med}(\|X^N\|)| > \varepsilon] \le Cre^{-cN} + Ce^{-c_P N\varepsilon^2}$$

for a constant $c_P > 0$ that depends only on P and universal constants $C, c > 0$.

The proof uses a Gaussian concentration inequality for non-Lipschitz functions.

Lemma 5.6 (Gaussian concentration)**.** *Let $Z \sim N(0, \mathbf{1}_d)$ be a d-dimensional standard Gaussian vector. Let $\Omega \subseteq \mathbb{R}^d$ be a measurable set with $\mathbf{P}[Z \in \Omega] \ge \frac{3}{4}$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be a function whose restriction to Ω is L-Lipschitz. Then*

$$\mathbf{P}[|f(Z) - \operatorname{med}(f(Z))| > \varepsilon] \le \mathbf{P}[Z \notin \Omega] + Ce^{-c\varepsilon^2/L^2}$$

for any $\varepsilon > 0$, where $C, c > 0$ are universal constants.

Proof. The result is stated in [1, Lemma 2.2] for random vectors on the unit sphere. The Gaussian case follows immediately by the Poincaré limit [9, eq. (3.3)]. □

Lemma 5.5 is then a direct consequence of the above result.

Proof of Lemma 5.5. When restricted to the set $\Omega = \{\|G_i^N\| \le \kappa+1 \text{ for all } i \in [r]\}$, it is clear that $\|X^N\| = \|P(G_1^N, \ldots, G_r^N)\|$ is an L_P-Lipschitz function of the real and imaginary parts of the entries of $G_1^N, \ldots, G_r^N$ on and above the diagonal, where L_P depends only on P. As these entries are independent Gaussians with variance $\frac{1}{N}$, the conclusion follows readily from Lemma 5.6 and Lemma 4.4. □

5.2.2. *Proof of Proposition 5.3.* The key observation behind the proof is that strong convergence and concentration of measure imply that any spectral statistic that vanishes in a neighborhood of $[-\|X_{\mathrm{F}}\|, \|X_{\mathrm{F}}\|]$ is exponentially small.

Corollary 5.7. *Fix $\varepsilon > 0$ and any bounded function $h \in C^\infty(\mathbb{R})$ that vanishes on the interval $[-\|X_{\mathrm{F}}\| - \varepsilon, \|X_{\mathrm{F}}\| + \varepsilon]$. Then we have*

$$|\mathbf{E}[\operatorname{tr} h(X^N)]| = O(e^{-cN}) \quad \text{as} \quad N \to \infty,$$

where $c > 0$ may depend on P and ε.

Proof. By the assumption on h, we have

$$|\mathbb{E}[\operatorname{tr} h(X^N)]| = |\mathbb{E}[\operatorname{tr} h(X^N) 1_{\|X^N\| > \|X_{\mathrm{F}}\| + \varepsilon}]| \le \|h\|_{(-\infty,\infty)} \mathbf{P}[\|X^N\| > \|X_{\mathrm{F}}\| + \varepsilon].$$

Now note that $\operatorname{med}(\|X^N\|) \le \|X_{\mathrm{F}}\| + \frac{\varepsilon}{2}$ for all N sufficiently large by Corollary 5.2. It follows that for all N sufficiently large

$$|\mathbb{E}[\operatorname{tr} h(X^N)]| \le \|h\|_{(-\infty,\infty)} \mathbf{P}[\|X^N\| - \operatorname{med}(\|X^N\|) > \tfrac{\varepsilon}{2}],$$

and the conclusion follows from Lemma 5.5. □

We can now prove Proposition 5.3.

Proof of Proposition 5.3. Fix any $\varepsilon > 0$ and bounded $h \in C^\infty(\mathbb{R})$ that vanishes on $[-\|X_{\mathrm{F}}\| - \varepsilon, \|X_{\mathrm{F}}\| + \varepsilon]$. We show by induction that $\nu_m(h) = 0$ for all m.

That $\nu_0(h) = 0$ follows immediately from Lemma 5.1. Now let $m \ge 1$ and assume that we have shown $\nu_0(h) = \cdots = \nu_{m-1}(h) = 0$. Then Theorem 4.1 yields

$$\left| \mathbf{E}[\operatorname{tr} h(X^N)] - \frac{\nu_m(h)}{N^m} \right| = O\left(\frac{1}{N^{m+1}}\right)$$

as $N \to \infty$. Thus

$$|\nu_m(h)| \le N^m |\mathbf{E}[\operatorname{tr} h(X^N)]| + O\left(\frac{1}{N}\right) = O\left(\frac{1}{N}\right)$$

as $N \to \infty$ by the triangle inequality and Corollary 5.7. As the left-hand side is independent of N, it follows that $\nu_m(h) = 0$. □

5.3. **Proof of Theorem 1.1: GUE case.** We can now prove Theorem 1.1 in the case that $\boldsymbol{G}^N$ are GUE matrices by combining Theorem 4.1 and Proposition 5.3.

Proof of Theorem 1.1: GUE case. We may assume without loss of generality that P is self-adjoint (see Remark 5.8 below). Fix $\varepsilon \in (0, 1]$, $K = (Cr)^{q_0}\|X_{\mathrm{F}}\|$, and $m \in \mathbb{N}$ with $m \le \frac{N}{2}$ that will be chosen at the end of the proof.

Let h be the test function provided by Lemma 2.5 with $m \leftarrow 2m$, $\rho \leftarrow \|X_{\rm F}\|$, and $\varepsilon \leftarrow \varepsilon\|X_{\rm F}\|$. Then $\nu_k(h) = 0$ for all $k \in \mathbb{Z}_+$ by Proposition 5.3, and thus

$$\begin{aligned}\mathbf{P}[\|X^N\| \geq (1+\varepsilon)\|X_{\rm F}\|] &\leq \mathbf{E}[\operatorname{Tr} h(X^N)] = DN\,\mathbf{E}[\operatorname{tr} h(X^N)] \\ &\leq DN\left[\frac{(Cq_0m)^{2m}}{m!N^m}\left(\frac{(Cr)^{q_0}}{\varepsilon}\right)^{2m+1} + Cre^{-cN}\left(1+\frac{(Cr)^{q_0}}{\varepsilon}\right)\right]\end{aligned}$$

for a universal constant C by Theorem 4.1 and Lemma 2.5. In particular, if we assume that $D \leq e^m$, then we can further estimate

$$\mathbf{P}[\|X^N\| \geq (1+\varepsilon)\|X_{\rm F}\|] \leq \frac{(Cr)^{q_0+1}N}{\varepsilon}\left[\left(\frac{(Cr)^{2q_0}q_0^2m}{N\varepsilon^2}\right)^m + e^{m-cN}\right]$$

for a universal constant C, where we used that $\frac{1}{m!} \leq (\frac{e}{m})^m$.

Now choose $m = \lfloor \frac{N\varepsilon^2}{L} \rfloor$ with $L := \max\{e(Cr)^{2q_0}q_0^2, \frac{2}{c}\}$. If $m \geq 2$, then we have $\frac{N\varepsilon^2}{2L} \leq m \leq \frac{N\varepsilon^2}{L}$, and the above estimate yields

$$\mathbf{P}[\|X^N\| \geq (1+\varepsilon)\|X_{\rm F}\|] \leq \frac{(Cr)^{q_0+1}N}{\varepsilon}\big(e^{-N\varepsilon^2/2L} + e^{-cN\varepsilon^2/2}\big),$$

when $D \leq e^{N\varepsilon^2/2L}$, where we used that $m - cN \leq (\frac{1}{L} - c)N\varepsilon^2 \leq -\frac{cN\varepsilon^2}{2}$ as $\varepsilon \leq 1$.

On the other hand, if $m < 2$, then $\frac{N\varepsilon^2}{L} < 2$ and thus

$$\mathbf{P}[\|X^N\| \geq (1+\varepsilon)\|X_{\rm F}\|] \leq 1 \leq e^2e^{-N\varepsilon^2/L} \leq \frac{e^2N}{\varepsilon}e^{-N\varepsilon^2/L},$$

where we used that $\frac{N}{\varepsilon} \geq 1$. This concludes the proof. □

Remark 5.8. Throughout this section, we assumed that P is a *self-adjoint* noncommutative polynomial, and proved Theorem 1.1 in this case. However, the conclusion extends immediately to arbitrary P by applying the self-adjoint case to P^*P. Thus the restriction to self-adjoint P entails no loss of generality. We apply the same observation in the remainder of the paper without further comment.

6. Strong convergence for GOE and GSE

The aim of this section is to complete the proof of Theorem 1.1 for the GOE and GSE ensembles. Most of the proof in the GUE case extends *verbatim* to the present setting, so that we will focus attention only on the necessary modifications. The key difference in the case of GOE and GSE is that it is no longer true that $\Phi_h(x) = \Phi_h(-x)$ as in Lemma 4.7, and thus that $\nu_1 = 0$ as in Lemma 5.1. Instead the arguments in the proof where these properties were used will be adapted to the GOE and GSE cases by exploiting supersymmetric duality.

The following setting and notations will be fixed throughout this section. Let $\boldsymbol{G}^N = (G_1^N, \ldots, G_r^N)$ and $\boldsymbol{H}^N = (H_1^N, \ldots, H_r^N)$ be independent GOE and GSE matrices of dimension N, respectively, and $\boldsymbol{s} = (s_1, \ldots, s_r)$ be a free semicircular family. We fix a self-adjoint noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle x_1, \ldots, x_r\rangle$ of degree q_0 with matrix coefficients of dimension D, and denote by

$$X^N := P(\boldsymbol{G}^N), \qquad Y^N := P(\boldsymbol{H}^N), \qquad X_{\rm F} := P(\boldsymbol{s})$$

the random matrices of interest and the limiting model.

6.1. **Supersymmetric duality.** The key fact we will use is that the GOE and GSE ensembles are dual in the sense that their moments are encoded by the same polynomial at positive and negative values. This is captured by the following result, which replaces Lemma 4.7 in the present setting.

Lemma 6.1 (Polynomial encoding)**.** *For any $h \in \mathcal{P}_q$, there is $\Phi_h \in \mathcal{P}_{qq_0}$ so that*

$$\begin{aligned}\mathbf{E}[\operatorname{tr} h(X^N)] &= \Phi_h(\tfrac{1}{N}),\\ \mathbf{E}[\operatorname{tr} h(Y^N)] &= \Phi_h(-\tfrac{1}{2N}).\end{aligned}$$

Proof. This is a direct consequence of the genus expansions for GOE and GSE established in [12]. In the present notation, [12, Theorem B] states that

$$\mathbf{E}[\operatorname{tr} G^N_{i_1}\cdots G^N_{i_{2n}}] = \sum_\Gamma N^{\chi(\Gamma)-2}$$

for every $n \in \mathbb{N}$ and $i_1,\ldots,i_{2n} \in [r]$, while [12, Theorem 4.1] states that

$$\mathbf{E}[\operatorname{tr} H^N_{i_1}\cdots H^N_{i_{2n}}] = \sum_\Gamma (-2N)^{\chi(\Gamma)-2}$$

for every $n \in \mathbb{N}$ and $i_1,\ldots,i_{2n} \in [r]$.[5] Here the sums range over a certain family (that is determined by the choice of $i_1,\ldots,i_{2n}$) of 2-dimensional CW-complexes Γ with 1 vertex and n edges, and $\chi(\Gamma)$ denotes the Euler characteristic. As each such Γ is connected and has at least one face, we have $-n \le \chi(\Gamma) - 2 \le 0$ for all such Γ, and thus the right-hand sides of the above equations are defined by the same polynomial of degree at most n applied to $\frac{1}{N}$ and $-\frac{1}{2N}$, respectively.

Now note that, by linearity, $\mathbf{E}[\operatorname{tr} h(X^N)]$ and $\mathbf{E}[\operatorname{tr} h(Y^N)]$ are linear combinations of expected traces of words of length at most qq_0 in $\boldsymbol{G}^N$ and $\boldsymbol{H}^N$, respectively. We have shown that words of even length yield a polynomial as in the statement, while words of odd length vanish as the Gaussian distribution is symmetric. Finally, note that $\mathbf{E}[\operatorname{tr} h(X^N)]$ is real as P is self-adjoint, so Φ_h is a real polynomial. □

6.2. **Asymptotic expansion.** With Lemma 6.1 in hand, we can now repeat the proof of Theorem 4.1 with only trivial modifications.

Theorem 6.2 (Smooth asymptotic expansion for GOE/GSE)**.** *There exist universal constants $C, c > 0$, and a compactly supported distribution ν_k for every $k \in \mathbb{Z}_+$, such that the following hold. Fix any bounded $h \in C^\infty(\mathbb{R})$, and define*

$$f(\theta) := h(K\cos\theta) \qquad \text{with} \qquad K := (Cr)^{q_0}\|X_{\mathrm{F}}\|.$$

Then for every $m, N \in \mathbb{N}$ with $m \le \frac{N}{2}$, we have

$$\begin{aligned}&\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1}\frac{\nu_k(h)}{N^k}\right| \vee \left|\mathbf{E}[\operatorname{tr} h(Y^N)] - \sum_{k=0}^{m-1}\frac{\nu_k(h)}{(-2N)^k}\right|\\ &\qquad\le \frac{(Cq_0)^{2m}}{m!N^m}\|f^{(2m+1)}\|_{[0,2\pi]} + Cre^{-cN}\big(\|h\|_{(-\infty,\infty)} + \|f^{(1)}\|_{[0,2\pi]}\big).\end{aligned}$$

[5] The reader should beware that the notation of [12] differs from that of the present paper: in [12], GOE and GSE matrices are normalized so that their off-diagonal elements have variance 1 and 4, respectively, and tr denotes the sum of the diagonal entries viewed as elements of $\mathbb{H}$.

Proof. The only modification that must be made to the arguments of section 4 is that we apply Lemma 6.1 instead of Lemma 4.7 in the proof of Lemma 4.6; the master inequalities for GOE and GSE are then obtained by Taylor expanding $\Phi_h(x)$ at the points $x = \frac{1}{N}$ and $x = -\frac{1}{2N}$, respectively. The remaining results and proofs in section 4.1 extend *verbatim* to the case of GOE and GSE. □

6.3. **The first-order distribution.** While Lemma 2.11 directly yields

$$\nu_0(h) = \lim_{N\to\infty} \mathbf{E}[\operatorname{tr} h(X^N)] = (\operatorname{tr}\otimes\tau)(h(X_{\mathrm{F}}))$$

as for GUE, it is no longer true in the present setting that we have $\nu_1(h) = 0$ as in Lemma 5.1. Nonetheless, we can exploit the duality between the GOE and GSE ensembles to give a very simple proof of the following fact.

Lemma 6.3. *In the setting of Theorem 6.2, we have*

$$\operatorname{supp}\nu_1 \subseteq [-\|X_{\mathrm{F}}\|, \|X_{\mathrm{F}}\|].$$

In the proof we will need the following basic fact about distributions.

Lemma 6.4. *Let ν be a compactly supported distribution and $A \subseteq \mathbb{R}$ be a closed set. If $\nu(h) = 0$ for all* nonnegative *bounded functions $h \in C^\infty(\mathbb{R})$ that vanish in a neighborhood of A, then we have* $\operatorname{supp}\nu \subseteq A$.

Proof. The assumption implies *a fortiori* that $\nu(h) \geq 0$ for all nonnegative functions $h \in C_0^\infty(\mathbb{R}\backslash A)$. The restricted distribution $\nu|_{\mathbb{R}\backslash A}$ (cf. [31, §2.2]) is therefore a positive measure by [31, Theorem 2.1.7]. Thus $\nu(h) = 0$ for all nonnegative functions $h \in C_0^\infty(\mathbb{R}\backslash A)$ then implies that $\nu|_{\mathbb{R}\backslash A} = 0$, which yields the conclusion. □

We can now prove Lemma 6.3.

Proof of Lemma 6.3. Let h be a bounded nonnegative function $h \in C^\infty(\mathbb{R})$, $h \geq 0$ that vanishes in a neighborhood of $[-\|X_{\mathrm{F}}\|, \|X_{\mathrm{F}}\|]$. Then Theorem 6.2 yields

$$0 \leq \mathbf{E}[\operatorname{tr} h(X^N)] = \frac{\nu_1(h)}{N} + O\left(\frac{1}{N^2}\right),$$
$$0 \leq \mathbf{E}[\operatorname{tr} h(Y^N)] = -\frac{\nu_1(h)}{2N} + O\left(\frac{1}{N^2}\right),$$

where we used that $\nu_0(h) = (\operatorname{tr}\otimes\tau)(h(X_{\mathrm{F}})) = 0$. Taking N sufficiently large then yields both $\nu_1(h) \geq 0$ and $-\nu_1(h) \geq 0$, which implies that we must have $\nu_1(h) = 0$. The conclusion now follows from Lemma 6.4. □

6.4. **Proof of Theorem 1.1: GOE/GSE case.** The remainder of the proof of Theorem 1.1 is now essentially identical to the proof of the GUE case.

Proof of Theorem 1.1: GOE/GSE case. The results and proofs of section 5 extend *verbatim* to the case of GOE and GSE, provided that Theorem 4.1 and Lemma 5.1 are replaced by Theorem 6.2 and Lemma 6.3, respectively. □

7. Strong convergence for $\mathrm{U}(N)$

We now turn our attention to Haar-distributed random matrices. While the structure of the proofs is similar to those for the Gaussian ensembles, there are distinct complications that arise in the two settings that must be surmounted to reach matrix coefficients of (nearly) exponential dimension. Unlike in the Gaussian case, no truncation is needed in the Haar setting as the random matrix model is already bounded. However, in the Haar setting polynomial spectral statistics are no longer polynomials, but rather rational functions, of $\frac{1}{N}$. These require a careful analysis in order not to incur any quantitative loss in the final result.

The aim of this section is to give a complete proof of Theorem 1.5 for Haar-distributed matrices in $\mathrm{U}(N)$. The requisite modifications for the $\mathrm{O}(N)/\mathrm{Sp}(N)$ models will subsequently be developed in section 8.

The following will be fixed throughout this section. Let $\boldsymbol{U}^N = (U_1^N, \ldots, U_r^N)$ be independent Haar-distributed random matrices in $\mathrm{U}(N)$, and let $\boldsymbol{u} = (u_1, \ldots, u_r)$ be free Haar unitaries. We further fix a self-adjoint noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle x_1, \ldots, x_r, x_1^*, \ldots, x_r^* \rangle$ of degree q_0 with matrix coefficients of dimension D. We denote the random matrix of interest and its limiting model as

$$X^N := P(\boldsymbol{U}^N, \boldsymbol{U}^{N*}), \qquad X_{\mathrm{F}} := P(\boldsymbol{u}, \boldsymbol{u}^*).$$

Finally, we will denote by

$$K := \|P\|_{\mathrm{M}_D(\mathbb{C}) \otimes C^*(\mathbf{F}_r)} = \sup_{n \in \mathbb{N}} \sup_{W_1, \ldots, W_r \in \mathrm{U}(n)} \|P(W_1, \ldots, W_r, W_1^*, \ldots, W_r^*)\|$$

the norm of P in the full C^*-algebra of the free group $\mathbf{F}_r$ with r free generators. The significance of this definition is that $\|X^N\| \le K$ a.s. for every N.

7.1. Polynomial encoding.

For GUE matrices, Lemma 4.7 showed that polynomial spectral statistics can be expressed as polynomials of $\frac{1}{N}$. The aim of this section is to prove an analogue of this property for $\mathrm{U}(N)$ matrices. In this case, however, we obtain rational functions rather than polynomials.

In the sequel, we always fix the following special polynomial

$$g_q(x) := \prod_{j=1}^{q} (1 - (jx)^2)^{\lfloor \frac{q}{j} \rfloor} \tag{7.1}$$

that will arise as the denominator in the rational expressions that appear for Haar-distributed matrices. The main result of this section is the following.

Lemma 7.1 (Polynomial encoding)**.** *For every $h \in \mathcal{P}_q$, there is a rational function of the form $\Psi_h := \frac{f_h}{g_{qq_0}}$ with $f_h, g_{qq_0} \in \mathcal{P}_{\lfloor 3qq_0(1+\log qq_0) \rfloor}$ so that*

$$\mathbf{E}[\operatorname{tr} h(X^N)] = \Psi_h(\tfrac{1}{N}) = \Psi_h(-\tfrac{1}{N})$$

for all $N \in \mathbb{N}$ such that $N > qq_0$.

Remark 7.2. Lemma 7.1 and its counterpart for $\mathrm{O}(N)/\mathrm{Sp}(N)$ models are solely responsible for the loss of the logarithmic factor in Theorem 1.5 as compared to Theorem 1.1, which arises from the logarithmic factor in the degree of the numerator and denominator f_h, g_{qq_0} of the rational function Ψ_h. It is unclear whether

this logarithmic factor is necessary, as there could be cancellations between the numerator and denominator that are not captured by the crude analysis below.

Lemma 7.1 for $\mathrm{U}(N)$ is a special case of [37, Theorem 3.1]. However, as the argument will be needed below also for $\mathrm{O}(N)/\mathrm{Sp}(N)$, we spell out the proof here. We begin with some definitions. For every $L \in \mathbb{N}$, we denote by S_L the symmetric group on L letters. In view of [20, Theorem 4.3], we introduce the following.

Definition 7.3 (Unitary Weingarten functions)**.** For any $L \in \mathbb{N}$ and $\alpha \in \mathrm{S}_L$

$$\mathrm{Wg}_L(\alpha, N) := \frac{1}{L!} \sum_{\lambda \vdash L} \frac{d_\lambda \chi^\lambda(\alpha)}{\prod_{\square \in \lambda}(N + c(\square))} \tag{7.2}$$

where the product runs over all boxes $\square = (i, j)$ in the Young diagram associated to λ and $c(\square) := i - j$. Here d_λ denotes the dimension and χ^λ denotes the character of the irreducible representation of S_L associated to λ.

For our purposes, the only relevant feature of the function $\mathrm{Wg}_L(\alpha, N)$ is that it is a rational function of N and that we know its poles.

Lemma 7.4. *For every $\alpha \in \mathrm{S}_L$, there exists $a_\alpha \in \mathcal{P}$ so that*

$$\mathrm{Wg}_L(\alpha, N) = \frac{a_\alpha(N)}{N^L \prod_{k=1}^L (N^2 - k^2)^{\lfloor \frac{L}{k} \rfloor}} \quad \textit{for all } N > L.$$

Proof. Fix a partition $\lambda \vdash L$. Since λ has at most L columns and at most L rows, we clearly have $-L \le c(\square) \le L$ for all $\square \in \lambda$. Thus

$$\prod_{\square \in \lambda} (N + c(\square)) = \prod_{k=-L}^{L} (N + k)^{\omega_k(\lambda)},$$

where $\omega_k(\lambda)$ denotes the number of boxes $\square = (i, j)$ in the Young diagram associated to λ with $i - j = k$. As for any such box, the Young diagram must contain the rectangle with side lengths i, j, respectively, we have $ij \le L$ and thus

$$\begin{aligned} \omega_k(\lambda) &\le \#\{(i, j) \in [L]^2 : ij \le L,\ i - j = k\} \\ &\le \#\{j \in [L] : (|k| + j)j \le L\} \le \lfloor \tfrac{L}{|k|+1} \rfloor. \end{aligned}$$

Thus the numerator of

$$\mathrm{Wg}_L(\alpha, N) = \frac{\sum_{\lambda \vdash L} \frac{d_\lambda \chi^\lambda(\alpha)}{L!} N^{L - \omega_0(\lambda)} \prod_{k=1}^L (N - k)^{\lfloor \frac{L}{k} \rfloor - \omega_{-k}(\lambda)} (N + k)^{\lfloor \frac{L}{k} \rfloor - \omega_k(\lambda)}}{N^L \prod_{k=1}^L (N^2 - k^2)^{\lfloor \frac{L}{k} \rfloor}}$$

is polynomial in N. It remains to note that the numerator is in fact a real polynomial, as all characters χ^λ of the symmetric group are real-valued. □

We can now prove Lemma 7.1.

Proof of Lemma 7.1. Let $w(U_1^N, \ldots, U_r^N)$ be a reduced word of length at most L in the Haar unitary matrices U_i^N and their adjoints U_i^{N*}. Suppose that $w \neq \mathbf{1}$ and that it is balanced, that is, U_i^N and U_i^{N*} appear an equal number of times. Let L_i

be the number of appearances of U_i^N in w. Then by [38, Theorem 2.8]

$$\mathbf{E}[\operatorname{tr} w(U_1^N,\ldots,U_r^N)] = \sum_{(\alpha_1,\beta_1)\in\mathrm{S}^2_{L_1},\ldots,(\alpha_r,\beta_r)\in\mathrm{S}^2_{L_r}} \left(\prod_{i=1}^r \mathrm{Wg}_{L_i}(\alpha_i^{-1}\beta_i,N)\right) \ell(N;\alpha_1,\beta_1,\ldots,\alpha_r,\beta_r)$$

for all $N > \max_i L_i$, where each $\ell(N;\alpha_1,\beta_1,\ldots,\alpha_r,\beta_r)$ is either identically zero or a non-negative integer power of N. Thus Lemma 7.4 yields

$$\mathbf{E}[\operatorname{tr} w(U_1^N,\ldots,U_r^N)] = \frac{b_w(N)}{N^L\prod_{k=1}^L (N^2-k^2)^{\lfloor \frac{L}{k}\rfloor}}$$

for some $b_w \in \mathcal{P}$, where we used that $\sum_{i=1}^r L_i \le L$ and therefore $\sum_{i=1}^r \lfloor \frac{L_i}{j}\rfloor \le \lfloor \frac{L}{j}\rfloor$. Now note that as $|\mathbf{E}[\operatorname{tr} w(U_1^N,\ldots,U_r^N)]| \le 1$ for all N, the degree of the numerator is at most the degree $\Sigma = L + 2\sum_{k=1}^L \lfloor \frac{L}{k}\rfloor \le 3L(1+\log L)$ of the denominator. Dividing both numerator and denominator by N^Σ therefore yields

$$\mathbf{E}[\operatorname{tr} w(U_1^N,\ldots,U_r^N)] = \frac{f_w(\frac{1}{N})}{g_L(\frac{1}{N})} \tag{7.3}$$

for all $N > L$, where g_L was defined in (7.1) and $f_w, g_L \in \mathcal{P}_\Sigma$.

If w is not balanced, it is readily seen that $\mathbf{E}[\operatorname{tr} w(U_1^N,\ldots,U_r^N)] = 0$ as U_i^N has the same distribution as $e^{i\theta}U_i^N$ for every θ, while clearly $\mathbf{E}[\operatorname{tr} w(U_1^N,\ldots,U_r^N)] = 1$ for $w = \mathbf{1}$. Thus (7.3) remains valid for all reduced words w.

Now note that $\mathbf{E}[\operatorname{tr} h(X^N)]$ is a linear combination of terms $\mathbf{E}[\operatorname{tr} w(U_1^N,\ldots,U_r^N)]$ for words w of length at most $L = qq_0$. As we have shown that each term is real and as $\mathbf{E}[\operatorname{tr} h(X^N)]$ is also real, the representation $\mathbf{E}[\operatorname{tr} h(X^N)] = \Psi_h(\frac{1}{N})$ follows from (7.3). That $\Psi_h(\frac{1}{N}) = \Psi_h(-\frac{1}{N})$ follows from [38, Remark 1.9], which implies that the power series expansion of Ψ_h contains only even powers of $\frac{1}{N}$. □

7.2. **A rational Bernstein inequality.** As Ψ_h in Lemma 7.1 is a rational function, we can no longer apply Bernstein's inequality directly to control the remainder term in its Taylor expansion as we did in section 4, where the analogous expression in Lemma 4.7 was polynomial. This issue could be surmounted by applying Bernstein's inequality to the numerator and denominator separately using the chain rule; see, e.g., [15, Lemma 4.3]. However, this naive approach turns out to be lossy: it results in a multipicative factor $m!$ in the bound on the mth derivative, which prevents us from reaching coefficients with exponential dimension.

Instead, we develop here a more specialized argument that exploits the fact that we have strong control over the denominator of the rational expressions.

Lemma 7.5 (Rational Bernstein inequality)**.** *Let $p,q \in \mathbb{N}$ with $p \ge q$, let $f \in \mathcal{P}_p$, and define the rational function $r := \frac{f}{g_q}$ where g_q is as defined in (7.1). Then*

$$\frac{1}{m!}\|r^{(m)}\|_{[-\frac{1}{cp},\frac{1}{cp}]} \le \left(e^{-p}(Cp)^m + \frac{(Cp)^{2m}}{m!}\right)\|r\|_{I_q}$$

for all $m \ge 1$, where c, C are universal constants and $I_q := \{\frac{1}{N} : N \in \mathbb{Z}, |N| > q\}$.

The key idea behind the proof is that the function $\frac{1}{g_q}$ and its derivatives can be approximated very precisely by a polynomial.

Lemma 7.6. *For every $b \in \mathbb{N}$, there is a polynomial $s \in \mathcal{P}_{2bq}$ so that*

$$\frac{1}{k!}\|(\tfrac{1}{g_q} - s)^{(k)}\|_{[-\frac{1}{8q},\frac{1}{8q}]} \le 2^{-bq}(8q)^k$$

for all $k \ge 0$. In particular, $\frac{1}{2}\frac{1}{g_q(x)} \le s(x) \le \frac{3}{2}\frac{1}{g_q(x)}$ for all $|x| \le \frac{1}{8q}$.

Proof. Let $\frac{1}{g_q(z)} = \sum_{i=0}^{\infty} a_i z^i$ be the power series expansion of $\frac{1}{g_q}$ around zero. As $|1 - z^2| \ge 1 - |z|^2 \ge e^{-2|z|^2}$ for $|z| \le \frac{1}{2}$, we can estimate using (7.1)

$$\left|\tfrac{1}{g_q(z)}\right| \le e^{\sum_{j=1}^{q} 2j^2|z|^2\lfloor \frac{q}{j}\rfloor} \le e^{q^2(q+1)|z|^2} \le e^{q/2}$$

for all $z \in \mathbb{C}$ with $|z| \le \frac{1}{2q}$. Thus Lemma 2.4 yields $|a_i| \le e^{q/2}(2q)^i$ for all i.

Now let $s(x) := \sum_{i=0}^{2bq} a_i x^i$. Then we can estimate

$$\frac{|(\frac{1}{g_q} - s)^{(k)}(x)|}{k!} \le e^{q/2}(8q)^k \sum_{i=\max\{2bq+1,k\}}^{\infty} \binom{i}{k}\frac{1}{4^i}$$

for $|x| \le \frac{1}{8q}$, where we used $(\frac{1}{g_q} - s)^{(k)}(x) = \sum_{i=\max\{2bq+1,k\}}^{\infty} a_i \frac{i!}{(i-k)!}x^{i-k}$. The first part of the lemma follows as $\binom{i}{k} \le 2^i$ and $2^{-2}e^{1/2} \le \frac{1}{2}$. The second part follows from the first using that $|\frac{1}{g_q(x)} - s(x)| \le \frac{1}{2} \le \frac{1}{2}\frac{1}{g_q(x)}$ for $|x| \le \frac{1}{8q}$ as $g_q(x) \le 1$ ☐

We can now complete the proof of Lemma 7.5.

Proof of Lemma 7.5. Let s be the polynomial of Lemma 7.6 (we will choose $b \in \mathbb{N}$ at the end of the proof). The product formula yields

$$\frac{r^{(m)}}{m!} = \frac{(fs)^{(m)}}{m!} + \sum_{k=0}^{m} \frac{f^{(k)}}{k!}\frac{(\frac{1}{g_q} - s)^{(m-k)}}{(m-k)!}.$$

As fs has degree $p' := p + 2bq$, applying Lemma 2.1 and Proposition 3.1 yields

$$\|f^{(k)}\|_{[-\frac{1}{cp},\frac{1}{cp}]} \le C(Cp)^{2k}\|f\|_{I_q} \le C(Cp)^{2k}\|r\|_{I_q},$$
$$\|(fs)^{(m)}\|_{[-\frac{1}{cp'},\frac{1}{cp'}]} \le (Cp')^{2m}\|fs\|_{I_q} \le (Cp')^{2m}\|r\|_{I_q}.$$

Here we used $|f| \le |r|$ in the last inequality on the first line, and that $|fs| \le \frac{3}{2}|r|$ by the second part of Lemma 7.6 in the last inequality on the second line.

To conclude the proof, note that

$$\sum_{k=0}^{m} \frac{\|f^{(k)}\|_{[-\frac{1}{cp'},\frac{1}{cp'}]}}{k!}\frac{\|(\frac{1}{g_q} - s)^{(m-k)}\|_{[-\frac{1}{cp'},\frac{1}{cp'}]}}{(m-k)!}$$
$$\le 2^{-bq}(Cp)^m\|r\|_{I_q}\sum_{k=0}^{m}\frac{p^k}{k!} \le e^p 2^{-bq}(Cp)^m\|r\|_{I_q},$$

where we used the first part of Lemma 7.6. If we now choose $b = \lceil \frac{2}{\log 2}\frac{p}{q}\rceil$, then $e^p 2^{-bq} \le e^{-p}$ and $p' \le Cp$, and the proof is readily completed. ☐

7.3. **The master inequality.** We now have the necessary ingredients to prove the analogue of Lemma 4.6 for the U(N) model. Note that as no truncation is needed, the result holds for all q, N and not merely for $q \lesssim N$ as in the Gaussian setting.

Lemma 7.7 (Master inequality)**.** *There exists a linear functional μ_m on $\mathcal{P}$ for every $m \in \mathbb{Z}_+$ such that for every $q \in \mathbb{N}$ and $h \in \mathcal{P}_q$*

$$|\mu_m(h)| \le \left((C\tilde{q}\tilde{q}_0)^m + \frac{(C\tilde{q}\tilde{q}_0)^{2m}}{m!}\right)\|h\|_{[-K,K]}, \tag{7.4}$$

and such that for all $N \in \mathbb{N}$

$$\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1}\frac{\mu_k(h)}{N^k}\right| \le \frac{1}{N^m}\left((C\tilde{q}\tilde{q}_0)^m + \frac{(C\tilde{q}\tilde{q}_0)^{2m}}{m!}\right)\|h\|_{[-K,K]}. \tag{7.5}$$

Here C is a universal constant, $\tilde{q} := q(1+\log q)$, and $\tilde{q}_0 := q_0(1+\log q_0)$.

Proof. Let Ψ_h be as in Lemma 7.1. Then the degree of the numerator of Ψ_h is bounded by $p = \lfloor 3\tilde{q}\tilde{q}_0 \rfloor$, and as $\|X^N\| \le K$ a.s. we have

$$|\Psi_h(\tfrac{1}{N})| = |\mathbf{E}[\operatorname{tr} h(X^N)]| \le \|h\|_{[-K,K]}.$$

Applying Lemma 7.5 therefore yields

$$\frac{1}{m!}\|\Psi_h^{(m)}\|_{[-\frac{1}{c\tilde{q}\tilde{q}_0},\frac{1}{c\tilde{q}\tilde{q}_0}]} \le \left((C\tilde{q}\tilde{q}_0)^m + \frac{(C\tilde{q}\tilde{q}_0)^{2m}}{m!}\right)\|h\|_{[-K,K]},$$

for every $m, N \in \mathbb{N}$, where $C, c > 0$ are universal constants. Now define

$$\mu_m(h) := \frac{\Psi_h^{(m)}(0)}{m!}.$$

Then (7.4) follows immediately from the previous equation display for $m \ge 1$, while for $m = 0$ we have $|\mu_0(h)| = \lim_{N\to\infty}|\mathbf{E}[\operatorname{tr} h(X^N)]| \le \|h\|_{[K,K]}$. Moreover, (7.5) follows for $\frac{1}{N} \le \frac{1}{c\tilde{q}\tilde{q}_0}$ by Taylor expanding Ψ_h as in the proof of Lemma 4.6.

On the other hand, in the case $\frac{1}{c\tilde{q}\tilde{q}_0} < \frac{1}{N}$, we first estimate

$$\begin{aligned}\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1}\frac{\mu_k(h)}{N^k}\right| &\le \sum_{k=0}^{m-1}\frac{1}{N^k}\left((C\tilde{q}\tilde{q}_0)^k + \frac{(C\tilde{q}\tilde{q}_0)^{2k}}{k!}\right)\|h\|_{[-K,K]} \\ &\le \frac{(C\tilde{q}\tilde{q}_0)^m}{N^m}\sum_{k=0}^{m-1}\left(1 + \frac{(C\tilde{q}\tilde{q}_0)^k}{k!}\right)\|h\|_{[-K,K]}\end{aligned}$$

using the triangle inequality and (7.4) on the first line, and $1 < \frac{c\tilde{q}\tilde{q}_0}{N}$ on the second line. We now consider two cases. If $C\tilde{q}\tilde{q}_0 < m$, the sum on the second line can be estimated by $m + e^{C\tilde{q}\tilde{q}_0} \le C^m$. If $C\tilde{q}\tilde{q}_0 \ge m$, we use that $\frac{a^k}{k!} \le \frac{a^m}{m!}$ for $m \le a$ to estimate the sum by $m(1+\frac{(C\tilde{q}\tilde{q}_0)^m}{m!})$. In each case, the proof is readily concluded. □

7.4. **Extension to smooth functions.** We now proceed to prove an analogue of Theorem 4.1 for the U(N) model. The proof in the present setting is somewhat simpler as Lemma 7.7 holds without constraint on q.

Theorem 7.8 (Smooth asymptotic expansion for U(N))**.** *There is a universal constant $C > 0$, and a compactly supported distribution μ_k for every $k \in \mathbb{Z}_+$, such*

that the following hold. Fix any $h \in C^\infty(\mathbb{R})$, and let $f(\theta) := h(K\cos\theta)$. Then

$$\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1} \frac{\mu_k(h)}{N^k}\right| \le \frac{(C\tilde{q}_0)^m}{u^m N^m}\|f^{(m'+1)}\|_{[0,2\pi]} + \frac{(C\tilde{q}_0)^{2m}}{m!\,u^{2m}N^m}\|f^{(2m'+1)}\|_{[0,2\pi]}$$

for every $m, N \in \mathbb{N}$ and $u \in (0,1)$, where $m' := \lceil (1+u)m \rceil$ and $\tilde{q}_0 := q_0(1+\log q_0)$.

We will need a variant of Lemma 2.3 with logarithmic terms.

Lemma 7.9. *Fix $k, l \in \mathbb{Z}_+$ and $u > 0$, and let $s := \lceil k + ul \rceil$. Let $h \in \mathcal{P}_q$ with Chebyshev expansion $h(x) = \sum_{j=0}^q a_j T_j(K^{-1}x)$, and let $f(\theta) := h(K\cos(\theta))$. Then*

$$\sum_{j=1}^q j^k (1+\log j)^l |a_j| \lesssim \left(\frac{e^u}{u}\right)^l \|f^{(s+1)}\|_{[0,2\pi]}.$$

Proof. This follows from Lemma 2.3 using $1 + \log j = \frac{1}{u}\log((ej)^u) \le \frac{e^u}{u} j^u$. □

We can now prove Theorem 7.8.

Proof of Theorem 7.8. We first note that the linear functional μ_m of Lemma 7.7 extends to a compactly supported distribution for every $m \in \mathbb{Z}_+$ by Lemma 2.7 and the inequality (7.4). Now fix $h \in \mathcal{P}_q$ with Chebyshev expansion $h(x) = \sum_{j=0}^q a_j T_j(K^{-1}x)$. Then (7.5) and the triangle inequality yield

$$\begin{aligned}\left|\mathbf{E}[\operatorname{tr} h(X^N)] - \sum_{k=0}^{m-1} \frac{\mu_k(h)}{N^k}\right| \\ \le \frac{(C\tilde{q}_0)^m}{N^m}\sum_{j=0}^q j^m(1+\log j)^m|a_j| + \frac{(C\tilde{q}_0)^{2m}}{m!N^m}\sum_{j=0}^q j^{2m}(1+\log j)^{2m}|a_j|.\end{aligned}$$

Applying Lemma 7.9 and noting that $e^u \le C$ for $u \le 1$ yields the conclusion for any $h \in \mathcal{P}$. The general result follows as polynomials are dene in $h \in C^\infty(\mathbb{R})$. □

7.5. **Bootstrapping.** We now aim to prove the following.

Proposition 7.10. *In the setting of Theorem 7.8, we have*

$$\operatorname{supp}\mu_m \subseteq [-\|X_{\mathrm{F}}\|, \|X_{\mathrm{F}}\|] \quad \text{for all } m \in \mathbb{Z}_+.$$

The proof of the analogous Gaussian result in Proposition 5.3 transfers to the present setting with minimal modifications. The main input we need is an appropriate concentration inequality for Haar unitary matrices.

Lemma 7.11. *For any $\varepsilon > 0$ and $N \in \mathbb{N}$, we have*

$$\mathbf{P}[|\|X^N\| - \operatorname{med}(\|X^N\|)| > \varepsilon] \le Ce^{-c_P N\varepsilon^2}$$

for a constant $c_P > 0$ that depends only on P and a universal constant $C > 0$.

Proof. As $\|X^N\| = \|P(\boldsymbol{U}^N, \boldsymbol{U}^{N*})\|$ is L_P-Lipschitz as a function of $U_1^N, \ldots, U_r^N$ where L_P depends only on P, concentration around the mean follows directly from [41, Theorem 5.17]. It is a standard fact that the concentration around the mean is equivalent to concentration around the median [35, Proposition 1.8]. □

The proof of Proposition 7.10 now follows exactly as in the Gaussian case.

Proof of Proposition 7.10. All the proofs in sections 5.1 and 5.2.2 extend directly to the present setting, provided that we replace Theorem 4.1 and Lemmas 4.7 and 5.5 by Theorem 7.8 and Lemmas 7.1 and 7.11, respectively. □

7.6. **Proof of Theorem 1.5: $\mathrm{U}(N)$ case.** We can now prove Theorem 1.5 in the case that $\boldsymbol{U}^N$ are Haar-distributed in $\mathrm{U}(N)$. We first note the following.

Lemma 7.12. *We can estimate* $K := \|P\|_{\mathrm{M}_D(\mathbb{C})\otimes C^*(\mathbf{F}_r)} \le (Cr)^{q_0}\|X_{\mathrm{F}}\|$.

Proof. We can represent the noncommutative polynomial P as

$$P(\boldsymbol{u},\boldsymbol{u}^*) = \sum_w A_w \otimes w(u_1,\ldots,u_r),$$

where $A_w \in \mathrm{M}_D(\mathbb{C})$ are matrix coefficients and where the sum is over all reduced words in $\boldsymbol{u},\boldsymbol{u}^*$ of length at most q_0. As there are at most $(Cr)^{q_0}$ such words, we immediately obtain $K \le (Cr)^{q_0}\max_w\|A_w\|$ by the triangle inequality. The conclusion follows as in the second part of the proof of Lemma 4.5 using that the reduced words $\{w(u_1,\ldots,u_r)\}$ are orthornormal in $L^2(\tau)$. □

We now complete the proof. The argument is very similar to the one used in section 5.3, except that we must take care of the logarithmic terms.

Proof of Theorem 1.5: $\mathrm{U}(N)$ case. Let $\varepsilon \in [\frac{2}{\sqrt{N}},1]$, and fix $m\in\mathbb{N}$ and $u\in[\frac{1}{m},1]$ that will be chosen at the end of the proof. Define $m' := \lceil(1+u)m\rceil$.

Let h be the test function provided by Lemma 2.5 with $m\leftarrow 2m'$, $\rho\leftarrow\|X_{\mathrm{F}}\|$, and $\varepsilon\leftarrow\varepsilon\|X_{\mathrm{F}}\|$. Then $\mu_k(h)=0$ for all $k\in\mathbb{Z}_+$ by Proposition 7.10. Thus

$$\begin{aligned}\mathbf{P}[\|X^N\|\ge(1+\varepsilon)\|X_{\mathrm{F}}\|] &\le \mathbf{E}[\mathrm{Tr}\,h(X^N)] = DN\,\mathbf{E}[\mathrm{tr}\,h(X^N)]\\ &\le \frac{(Cr)^{q_0}DN}{\varepsilon}\left[\left(\frac{(Cr)^{3q_0}\tilde q_0 m^{1+2u}}{\varepsilon^{1+2u}uN}\right)^m + \left(\frac{(Cr)^{6q_0}\tilde q_0^2 m^{1+4u}}{\varepsilon^{2(1+2u)}u^2N}\right)^m\right]\end{aligned}$$

for a universal constant C by Theorem 7.8, Lemma 2.5, and Lemma 7.12, where $\tilde q_0 := q_0(1+\log q_0)$ and we used that $m'\le(1+2u)m\le 3m$ and $\frac{1}{m!}\le(\frac{e}{m})^m$.

Now assume that $D\le e^m$, and choose $u=\frac{1}{1+\log m}$ and $m=\lfloor\frac{N\varepsilon^2}{L\log^2(N\varepsilon^2)}\rfloor$. Then $m(1+\log m)^2\le\frac{N\varepsilon^2}{L}$ and $\varepsilon^u\ge\frac{1}{C}$ (as $\varepsilon\ge\frac{2}{\sqrt{N}}$), so the inequality simplifies to

$$\mathbf{P}[\|X^N\|\ge(1+\varepsilon)\|X_{\mathrm{F}}\|] \le \frac{(Cr)^{q_0}N}{\varepsilon}\left(\frac{(Cr)^{6q_0}\tilde q_0^2}{L}\right)^m.$$

The conclusion follows by choosing $L=e(Cr)^{6q_0}\tilde q_0^2$ provided that $m\ge 2$, which is ensured by assuming that $\varepsilon\ge\frac{1}{c\sqrt{N}}$ for a constant c that depends only on L. □

8. Strong convergence for $\mathrm{O}(N)$ and $\mathrm{Sp}(N)$

The aim of this section is to complete the proof of Theorem 1.5 for Haar-distributed random matrices in $\mathrm{O}(N)$ and $\mathrm{Sp}(N)$. As in the Gaussian case (section 6), most of the proof in the $\mathrm{U}(N)$ case extends *verbatim* to the present setting, so that we will focus attention here only on the necessary modifications.

The following setting and notations will be fixed throughout this section. Let $\boldsymbol{U}^N = (U_1^N, \ldots, U_r^N)$ and $\boldsymbol{V}^N = (V_1^N, \ldots, V_r^N)$ be independent Haar-distributed random matrices in $\mathrm{O}(N)$ and $\mathrm{Sp}(N)$, respectively, and let $\boldsymbol{u} = (u_1, \ldots, u_r)$ be free Haar unitaries. We further fix a self-adjoint noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle x_1, \ldots, x_r, x_1^*, \ldots, x_r^* \rangle$ of degree q_0 with matrix coefficients of dimension D, and denote the random matrices of interest and their limiting model as

$$X^N := P(\boldsymbol{U}^N, \boldsymbol{U}^{N*}), \qquad Y^N := P(\boldsymbol{V}^N, \boldsymbol{V}^{N*}), \qquad X_{\mathrm{F}} := P(\boldsymbol{u}, \boldsymbol{u}^*).$$

Finally, we define K as in section 7.

8.1. **Polynomial encoding and supersymmetric duality.** We begin by proving an analogue of Lemmas 6.1 and 7.1 in the present setting. We will always denote by g_q the polynomial defined in (7.1) without further comment.

Lemma 8.1 (Polynomial encoding)**.** *For every $h \in \mathcal{P}_q$, there is a rational function of the form $\Psi_h := \frac{f_h}{g_{qq_0}}$ with $f_h, g_{qq_0} \in \mathcal{P}_{\lfloor 6qq_0(1+\log qq_0)\rfloor}$ so that*

$$\begin{aligned}\mathbf{E}[\operatorname{tr} h(X^N)] &= \Psi_h(\tfrac{1}{N}),\\ \mathbf{E}[\operatorname{tr} h(Y^N)] &= \Psi_h(-\tfrac{1}{2N})\end{aligned}$$

for all $N \in \mathbb{N}$ such that $N > qq_0$.

In the proof, we require the orthogonal counterparts of the Weingarten functions of Definition 7.3. It follows from [19, Theorem 3.1] that these are given by

$$\widetilde{\mathrm{Wg}}_L(m_1, m_2, N) = \sum_{\lambda \vdash L} \frac{C_{\lambda, m_1, m_2}}{\prod_{(i,j)\in\lambda}(N + 2j - i - 1)} \tag{8.1}$$

for all $L, N \in \mathbb{N}$ with $L \le N$ and $m_1, m_2 \in \mathcal{M}_{2L}$. Here $\mathcal{M}_{2L}$ denotes the set of perfect matchings (i.e., pair partitions) of $[2L]$, and C_{λ,m_1,m_2} is a real constant that depends only on λ, m_1, m_2 whose precise form is irrelevant for our analysis. The following lemma is the counterpart of Lemma 7.4 in the present setting.

Lemma 8.2. *For every $m_1, m_2 \in \mathcal{M}_{2L}$, there exists $a_{m_1,m_2} \in \mathcal{P}$ so that*

$$\widetilde{\mathrm{Wg}}_L(m_1, m_2, N) = \frac{a_{m_1,m_2}(N)}{N^L \prod_{k=1}^{2L}(N^2 - k^2)^{\lfloor \frac{2L}{k} \rfloor}} \quad \text{for all } N > 2L.$$

Proof. Fix a partition $\lambda \vdash L$. Since λ has at most L columns and at most L rows, we clearly have $-L \le 2j - i - 1 \le 2L$ for all $(i,j) \in \lambda$. Thus

$$\prod_{(i,j)\in\lambda}(N + 2j - i - 1) = \prod_{k=-2L}^{2L}(N + k)^{\omega_k(\lambda)},$$

where $\omega_k(\lambda)$ denotes the number of $(i,j) \in \lambda$ with $2j - i - 1 = k$. As any $(i,j) \in \lambda$ must satisfy $ij \le L$ (cf. the proof of Lemma 7.4), we can estimate

$$\omega_k(\lambda) \le \#\{(i,j) \in [L]^2 : ij \le L,\ 2j - i - 1 = k\}.$$

If $k > 0$, we can further estimate

$$\omega_k(\lambda) \le \#\{i \in [L] : i(k + i + 1) \le 2L\} \le \lfloor \tfrac{2L}{k+2} \rfloor.$$

Similarly, if $k \le 0$ we can estimate

$$\omega_k(\lambda) \le \#\{j \in [L] : (2j-1+|k|)j \le L\} \le \lfloor \tfrac{L}{|k|+1} \rfloor.$$

Therefore $\omega_0(\lambda) \le L$ and $\omega_k(\lambda) \le \lfloor \frac{2L}{|k|} \rfloor$ for all $k \ne 0$.

It follows from the above observations and (8.1) that the numerator of

$$\widetilde{\mathrm{Wg}}_L(m_1,m_2,N) = \frac{\sum_{\lambda \vdash L} C_{\lambda,m_1,m_2} N^{L-\omega_0(\lambda)} \prod_{k=1}^{2L} (N-k)^{\lfloor \frac{2L}{k} \rfloor - \omega_{-k}(\lambda)} (N+k)^{\lfloor \frac{2L}{k} \rfloor - \omega_k(\lambda)}}{N^L \prod_{k=1}^{2L} (N^2-k^2)^{\lfloor \frac{2L}{k} \rfloor}}$$

is a real polynomial N, concluding the proof. □

We can now complete the proof of Lemma 8.1.

Proof of Lemma 8.1. The proof for the $\mathrm{O}(N)$ case is nearly identical to the proof of Lemma 7.1. Specifically, let $w(U_1^N, \dots, U_r^N)$ be a reduced word of length $L \le qq_0$ in the Haar orthogonal matrices U_i^N and their adjoints U_i^{N*}. If $w \ne \mathbf{1}$ and w is even—that is, each U_i^N appears an even number of times (with or without adjoint)—then a rational reperesentation of $\mathbf{E}[\mathrm{tr}\, w(U_1^N, \dots, U_r^N)]$ as in the statement of the lemma follows by the identical argument as in the proof of Lemma 7.1 by using [39, Theorem 3.4] and Lemma 8.2 instead of [38, Theorem 2.8] and Lemma 7.4, respectively. If w is not even, then $\mathbf{E}[\mathrm{tr}\, w(U_1^N, \dots, U_r^N)] = 0$ as U_i^N has the same distribution as $-U_i^N$. The proof is now readily completed.

The proof for the $\mathrm{Sp}(N)$ case follows directly from the $\mathrm{O}(N)$ case and the supersymmetric duality property given in [39, Theorem 1.2]. □

8.2. **Asymptotic expansion.** With Lemma 8.1 in hand, we can now repeat the proof of Theorem 7.8 with only trivial modifications.

Theorem 8.3 (Smooth asymptotic expansion for $\mathrm{O}(N)/\mathrm{Sp}(N)$)**.** *There is a universal constant $C > 0$, and a compactly supported distribution μ_k for every $k \in \mathbb{Z}_+$, so that the following hold. Fix any $h \in C^\infty(\mathbb{R})$, and let $f(\theta) := h(K\cos\theta)$. Then*

$$\left|\mathbf{E}[\mathrm{tr}\, h(X^N)] - \sum_{k=0}^{m-1} \frac{\mu_k(h)}{N^k}\right| \vee \left|\mathbf{E}[\mathrm{tr}\, h(Y^N)] - \sum_{k=0}^{m-1} \frac{\mu_k(h)}{(-2N)^k}\right| \le \frac{(C\tilde{q}_0)^m}{u^m N^m} \|f^{(m'+1)}\|_{[0,2\pi]} + \frac{(C\tilde{q}_0)^{2m}}{m!\, u^{2m} N^m} \|f^{(2m'+1)}\|_{[0,2\pi]}$$

for all $m, N \in \mathbb{N}$ and $u \in (0,1)$, where $m' := \lceil (1+u)m \rceil$ and $\tilde{q}_0 := q_0(1+\log q_0)$.

Proof. The only modification that must be made to the arguments of sections 7.3 and 7.4 is that we apply Lemma 8.1 instead of Lemma 7.1 in the proof of Lemma 7.7; the master inequalities for $\mathrm{O}(N)$ and $\mathrm{Sp}(N)$ are then obtained by Taylor expanding $\Psi_h(\frac{x}{2})$ at the points $x = \frac{2}{N}$ and $x = -\frac{1}{N}$, respectively. □

8.3. **Bootstrapping.** We now aim to prove the following.

Proposition 8.4. *In the setting of Theorem 8.3, we have*

$$\mathrm{supp}\, \mu_m \subseteq [-\|X_{\mathrm{F}}\|, \|X_{\mathrm{F}}\|] \quad \text{for all } m \in \mathbb{Z}_+.$$

Proof. For $m = 0, 1$, the proofs of section 6.3 extend *verbatim* to the present setting, provided that we use Theorem 8.3 instead of Theorem 6.2.

For $m \geq 2$, we may consider only the $\mathrm{Sp}(N)$ case without loss of generality, as the expansion in the $\mathrm{O}(N)$ case is defined by the same distributions μ_m. The proof is then identical to that of Proposition 7.10, provided that we replace Theorem 7.8 and Lemma 7.1 by Theorem 8.3 and Lemma 8.1, respectively, and we note that the proof of Lemma 7.11 extends *verbatim* to the $\mathrm{Sp}(N)$ model. □

Remark 8.5. Curiously, the proof of Lemma 7.11 does not extend directly to the $\mathrm{O}(N)$ model, as $\mathrm{O}(N)$ has two disjoint connected components and thus cannot exhibit a Lipschitz concentration principle. This minor issue is easily surmounted, but we need not do so as we can work with $\mathrm{Sp}(N)$ without loss of generality.

8.4. **Proof of Theorem 1.5: $\mathrm{O}(N)/\mathrm{Sp}(N)$ case.** The remainder of the proof of Theorem 1.5 is now essentially identical to the proof of the $\mathrm{U}(N)$ case.

Proof of Theorem 1.5: $\mathrm{O}(N)/\mathrm{Sp}(N)$ case. The proof in section 7.6 extends *verbatim* to the present setting, provided that Theorem 7.8 and Proposition 7.10 are replaced by Theorem 8.3 and Proposition 8.4, respectively. □

9. Applications

9.1. **Subexponential operator spaces.** The aim of this section is to prove Corollary 1.2. Let us begin by recalling some basic definitions [50, §4].

Definition 9.1 (Operator spaces)**.** An *operator space* is a closed subspace of a C^*-algebra. A finite-dimensional operator space $\mathbf{W}$ is called

- *exact* if for every $C > 1$, there exists $S \in \mathbb{N}$ and a linear embedding $u : \mathbf{W} \to \mathrm{M}_S(\mathbb{C})$ such that for every $N \in \mathbb{N}$ and $x \in \mathbf{W} \otimes \mathrm{M}_N(\mathbb{C})$, we have
$$\|(u \otimes \mathrm{id})(x)\| \leq \|x\| \leq C\|(u \otimes \mathrm{id})(x)\|;$$
- *C-subexponential* if there exists $D_N \in \mathbb{N}$ with $D_N = e^{o(N)}$ and a linear embedding $f_N : \mathbf{W} \to \mathrm{M}_{D_N}(\mathbb{C})$ such that for every $N \in \mathbb{N}$ and $x \in \mathbf{W} \otimes \mathrm{M}_N(\mathbb{C})$, we have
$$\|(f_N \otimes \mathrm{id})(x)\| \leq \|x\| \leq C\|(f_N \otimes \mathrm{id})(x)\|.$$

An operator space $\mathbf{W}$ is called exact or C-subexponential if every finite-dimensional subspace of $\mathbf{W}$ is exact or C-subexponential, respectively. The subexponential constant of an operator space $\mathbf{W}$ is $C(\mathbf{W}) := \inf\{C : \mathbf{W} \text{ is } C\text{-subexponential}\}$.

An important fact that will be used in the sequel is that the C^*-algebra $\mathcal{A}$ generated by a free semicircular family $\boldsymbol{s} = (s_1, \ldots, s_r)$ is exact (this follows, for example, from [49, Corollary 17.10] and the proof of [27, Theorem 2.4]). Examples of non-exact subexponential operator spaces are given in [50].

Remark 9.2. Given a noncommutative polynomial $Q \in \mathbf{W} \otimes \mathbb{C}\langle \boldsymbol{s} \rangle$ with coefficients in an operator space $\mathbf{W}$, we will always view $Q(\boldsymbol{s})$ as an element of the minimal tensor product $\mathbf{W} \otimes_{\min} \mathcal{A}$ whose norm is denoted as $\|\cdot\|_{\min}$; cf. [49, §2.1].

We now turn to the proof of Corollary 1.2. The main difficulty in the proof is in fact to prove the lower bound on $\|P_N(\boldsymbol{G}^N)\|$, as the upper bound is immediate

from Theorem 1.1. For completeness, we begin by recalling the standard argument for the case that $P_N = P$ is independent of N.

Lemma 9.3. *Let $\boldsymbol{G}^N$ and $\boldsymbol{s}$ be as in Theorem 1.1. For any $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{s}\rangle$, we have $\|P(\boldsymbol{G}^N)\| \ge (1-o(1))\|P(\boldsymbol{s})\|$ a.s. as $N \to \infty$.*

Proof. We may assume that P is self-adjoint (Remark 5.8). Lemma 2.11 yields

$$\mathbf{E}[\|P(\boldsymbol{G}^N)\|^{2p}] \ge \mathbf{E}[\mathrm{tr}\, P(\boldsymbol{G}^N)^{2p}] = (1+o(1))\,(\mathrm{tr}\otimes\tau)(P(\boldsymbol{s})^{2p})$$

as $N \to \infty$ for every $p \in \mathbb{N}$. Thus Lemma 5.5 and the Borel-Cantelli lemma yield

$$\|P(\boldsymbol{G}^N)\|^{2p} \ge (1+o(1))\,(\mathrm{tr}\otimes\tau)(P(\boldsymbol{s})^{2p}) \quad \text{a.s.}$$

as $N \to \infty$. It remains to note that $[(\mathrm{tr}\otimes\tau)(P(\boldsymbol{s})^{2p})]^{\frac{1}{2p}} \to \|P(\boldsymbol{s})\|$ as $p \to \infty$. □

We can now complete the proof of Corollary 1.2.

Proof of Corollary 1.2. Let P_N and $D_N = e^{o(N)}$ be as in the statement. Applying Theorem 1.1 with $P \leftarrow P_N$ and $\varepsilon \leftarrow \varepsilon_N := \max\{(\frac{\log D_N}{cN})^{1/2}, N^{-1/4}\} = o(1)$ yields

$$\mathbf{P}\big[\|P_N(\boldsymbol{G}^N)\| \ge (1+\varepsilon_N)\|P_N(\boldsymbol{s})\|\big] \le \frac{N^{5/4}}{c}\, e^{-cN^{1/2}}.$$

Thus $\|P_N(\boldsymbol{G}^N)\| \le (1+o(1))\|P_N(\boldsymbol{s})\|$ a.s. by the Borel-Cantelli lemma.

To prove the corresponding lower bound, note first that by a compactness argument, the conclusion of Lemma 9.3 extends readily to the case that P_N may depend on N but with both degree $O(1)$ and matrix coefficients of dimension $D_N = O(1)$. The lower bound for general D_N reduces to the case $D_N = O(1)$ by [37, Lemma 5.13] and the fact that the C^*-algebra generated by $\boldsymbol{s}$ is exact. We have therefore proved the lower bound $\|P_N(\boldsymbol{G}^N)\| \ge (1-o(1))\|P_N(\boldsymbol{s})\|$ a.s.

For the second part, let $\mathbf{W}$ be a subexponential operator space, and fix any $Q \in \mathbf{W} \otimes \mathbb{C}\langle \boldsymbol{s}\rangle$ and $C > C(\mathbf{W})$. Denote by $\mathbf{V} \subseteq \mathbf{W}$ the finite-dimensional operator space spanned by the coefficients of Q. For every $N \in \mathbb{N}$, let $f_N : \mathbf{V} \to \mathrm{M}_{D_N}(\mathbb{C})$ with $D_N = e^{o(N)}$ be the embedding provided by Definition 9.1. Then

$$(1-o(1))\|P_N(\boldsymbol{s})\| = \|P_N(\boldsymbol{G}^N)\| \le \|Q(\boldsymbol{G}^N)\| \le C\|P_N(\boldsymbol{G}^N)\| = C(1+o(1))\|P_N(\boldsymbol{s})\|$$

a.s., where $P_N = (f_N \otimes \mathrm{id})(Q) \in \mathrm{M}_{D_N}(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{s}\rangle$.

Now let $\mathbf{A}$ be the operator space spanned by the monomials of Q, and let $a > 1$. As the C^*-algebra generated by $\boldsymbol{s}$ is exact, so is $\mathbf{A}$. Let $S \in \mathbb{N}$ and the embedding $u : \mathbf{A} \to \mathrm{M}_S(\mathbb{C})$ be as in Definition 9.1 with $C \leftarrow a$. Then

$$\frac{1}{Ca}\|Q(\boldsymbol{s})\|_{\min} \le \frac{1}{C}\|(\mathrm{id}\otimes u)(Q(\boldsymbol{s}))\| \le \|P_N(\boldsymbol{s})\| \le a\|(\mathrm{id}\otimes u)(Q(\boldsymbol{s}))\| \le a\|Q(\boldsymbol{s})\|_{\min}$$

for all $N \ge S$, where we used [49, Proposition 2.1.1] for the first and last inequality. We conclude the proof by taking $a \downarrow 1$ and $C \downarrow C(\mathbf{W})$. □

9.2. **Improved rates for random permutations.** Let $\tilde\Pi_1^N, \dots, \tilde\Pi_r^N$ be i.i.d. uniformly distributed random permutation matrices of dimension N, and denote by $\Pi_i^N := \tilde\Pi_i^N|_{1^\perp}$ their restriction to the orthogonal complement of invariant vector 1. A breakthrough result of Bordenave and Collins [6] shows that $\boldsymbol{\Pi}^N = (\Pi_1^N, \dots, \Pi_r^N)$ converges strongly to free Haar unitaries $\boldsymbol{u} = (u_1, \dots, u_r)$.

As for the Gaussian and classical compact group ensembles, it is expected that the rate of convergence of $\|P(\mathbf{\Pi}^N, \mathbf{\Pi}^{N*})\|$ to $\|P(\boldsymbol{u}, \boldsymbol{u}^*)\|$ is of order $N^{-2/3}$, in agreement with Tracy-Widom asymptotics. This was proved up to a logarithmic factor in the recent work [32] in the special case $P(\boldsymbol{x}, \boldsymbol{x}^*) = x_1 + x_1^* + \cdots + x_r + x_r^*$, but remains well outside the reach of current methods for general P. In this setting, the best known rate of $(\frac{\log N}{N})^{1/8}$ was proved in [15], considerably improving the $\frac{\log\log N}{\log N}$ rate established by Bordenave and Collins in [7].

The main result of this section is a further quantitative improvement.

Theorem 9.4. *For any noncommutative polynomial $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{u}, \boldsymbol{u}^* \rangle$,*

$$\|P(\mathbf{\Pi}^N, \mathbf{\Pi}^{N*})\| \le \|P(\boldsymbol{u}, \boldsymbol{u}^*)\| + O_{\mathrm{P}}\Big(\big(\tfrac{\log N}{N}\big)^{1/6}\Big) \quad \text{as} \quad N \to \infty.$$

The result of Theorem 9.4 is only a modest improvement on that of [15, §3.3]. We include it here to illustrate that the methods of this paper yield quantitative improvements, essentially for free, even for models such as random permutations for which good concentration and duality properties are not available.

In the remainder of this section, we fix $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \boldsymbol{u}, \boldsymbol{u}^* \rangle$ of degree q_0, and write $X^N := P(\mathbf{\Pi}^N, \mathbf{\Pi}^{N*})$ and $X_{\mathrm{F}} := P(\boldsymbol{u}, \boldsymbol{u}^*)$. In this setting, the analogue of Lemma 7.1 is stated in [15, Lemma 5.1]: for every $h \in \mathcal{P}_q$, there is a rational function of the form $\Psi_h := \frac{f_h}{\tilde g_{qq_0}}$ with $f_h, \tilde g_{qq_0} \in \mathcal{P}_{\lfloor qq_0(1+\log r)\rfloor}$ so that

$$\mathbf{E}[\operatorname{tr} h(X^N)] = \Psi_h(\tfrac{1}{N}) \quad \text{for all } N \ge qq_0,$$

where

$$\tilde g_q(x) := \prod_{j=1}^{q-1} (1 - jx)^{\min\{r, \lfloor \frac{q}{j+1} \rfloor\}}.$$

We emphasize that we have no control of $\Psi_h(-x)$ here (cf. section 1.3.3). Thus we must use the Markov rather than Bernstein inequality in the proof.

The improved rate of Theorem 9.4 arises by replacing the classical polynomial interpolation argument used in [15] by the optimal interpolation inequality of Proposition 3.1, which enables us to interpolate Ψ_h between the $\frac{1}{N}$ samples for $N \gtrsim M = qq_0(1+\log r)$. The difficulty that then arises is that $\tilde g_{qq_0}$ is not uniformly bounded away from zero on the interval $[0, \frac{1}{M}]$, so that the elementary rational Markov inequality of [15, Lemma 4.3] cannot be applied. To surmount this issue, we prove a variant of Lemma 7.5 in the present setting.

Lemma 9.5 (Rational Markov inequality)**.** *Let $p, q \in \mathbb{N}$ with $p \ge q$, let $f \in \mathcal{P}_p$, and define the rational function $r := \frac{f}{\tilde g_q}$. Then*

$$\frac{1}{m!}\|r^{(m)}\|_{[0, \frac{1}{cp}]} \le \left(e^{-p}(Cp)^m + \frac{(Cp)^{3m}}{(m!)^2}\right) \sup_{N \ge q} |r(\tfrac{1}{N})|$$

for all $m \ge 1$, where c, C are universal constants.

Proof. We first note that the statement and proof of Lemma 7.6 extend readily to the setting where g_q is replaced by $\tilde g_q$. We proceed as in the proof of Lemma 7.5.

Applying the Markov inequality [15, Lemma 4.1] instead of Lemma 2.1 yields

$$k!\,\|f^{(k)}\|_{[-\frac{1}{cp},\frac{1}{cp}]} \le C(Cp)^{3k}\|f\|_{I_q} \le C(Cp)^{3k}\|r\|_{I_q},$$
$$m!\,\|(fs)^{(m)}\|_{[-\frac{1}{cp'},\frac{1}{cp'}]} \le (Cp')^{3m}\|fs\|_{I_q} \le (Cp')^{3m}\|r\|_{I_q},$$

where we now let $I_q = \{\frac{1}{N} : N \in \mathbb{N}, N \ge q\}$ and we used that $(2k-1)!! \ge k!$. The argument is readily concluded as in the proof of Lemma 7.5 by noting that

$$\sum_{k=0}^{m} \frac{\|f^{(k)}\|_{[-\frac{1}{cp'},\frac{1}{cp'}]}}{k!} \frac{\|(\frac{1}{\bar g_q} - s)^{(m-k)}\|_{[-\frac{1}{cp'},\frac{1}{cp'}]}}{(m-k)!}$$
$$\le 2^{-bq}(Cp)^m\|r\|_{I_q}\sum_{k=0}^{m}\frac{p^{2k}}{(k!)^2} \le e^{2p}2^{-bq}(Cp)^m\|r\|_{I_q},$$

where we used that $\sum_{k\ge0}\frac{p^{2k}}{(k!)^2} \le (\sum_{k\ge0}\frac{p^k}{k!})^2 = e^{2p}$. □

We can now complete the proof of Theorem 9.4.

Proof of Theorem 9.4. We can repeat the proof of [15, Theorem 3.9] *verbatim*, with the only modification that we use Lemma 9.5 to obtain the estimate

$$\frac{\|\Psi_h^{(m)}\|_{C^0[0,\frac{1}{M}]}}{m!} \le (Cqq_0(1+\log r))^{3m}\|h\|_{[-K,K]}$$

in the proof of [15, Theorem 6.1 and Corollary 6.2], instead of the corresponding estimate in [15] that has the exponent $4m$ rather than $3m$. □

9.3. **Hayes' model.** Fix $L, r \in \mathbb{N}$, and let $G_k^{N,i}$ be independent GUE matrices of dimension N for $i \in [L]$, $k \in [r]$. The aim of this section is to investigate whether the family of N^L-dimensional random matrices

$$\tilde{\boldsymbol{G}}^N := \big\{\mathbf{1}_N^{\otimes(i-1)} \otimes G_k^{N,i} \otimes \mathbf{1}_N^{\otimes(L-i)} : i \in [L],\ k \in [r]\big\}$$

converges strongly as $N \to \infty$ to

$$\tilde{\boldsymbol{s}} := \big\{\mathbf{1}^{\otimes(i-1)} \otimes s_k \otimes \mathbf{1}^{\otimes(L-i)} : i \in [L],\ k \in [r]\big\}$$

in $(\mathcal{A}^{\otimes_{\min}L}, \tau^{\otimes L})$, where $\mathcal{A}$ denotes the C^*-algebra generated by a free semicircular family $\boldsymbol{s} = (s_1, \dots, s_r)$ with respect to the trace τ. This question was considered in an influential paper of Hayes [30], whose main result states that strong convergence of this model in the case $L = 2$ implies an affirmative answer to a conjecture of Peterson and Thom in the theory of von Neumann algebras.

That Hayes' question does indeed have an affirmative answer has now been established by a variety of methods [4, 7, 37, 48], proving the Peterson-Thom conjecture. We provide yet another proof as a consequence of Corollary 1.2.

Lemma 9.6. *For any $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle\tilde{\boldsymbol{s}}\rangle$, we have*

$$\|P(\tilde{\boldsymbol{G}}^N)\| = (1+o(1))\|P(\tilde{\boldsymbol{s}})\|_{\min} \text{ a.s. as } N \to \infty.$$

Proof. Denote by $\tilde{\boldsymbol{G}}^N_{\le j}$ the subset of $\tilde{\boldsymbol{G}}^N$ with index $1 \le i \le j$, and denote by $\tilde{\boldsymbol{s}}_{>j}$ the subset of $\tilde{\boldsymbol{s}}$ with index $j < i \le L$. It clearly suffices to show that

$$\|P(\tilde{\boldsymbol{G}}^N_{\le j}, \tilde{\boldsymbol{s}}_{>j})\|_{\min} = (1+o(1))\|P(\tilde{\boldsymbol{G}}^N_{\le j-1}, \tilde{\boldsymbol{s}}_{>j-1})\|_{\min} \text{ a.s. as } N \to \infty$$

for $j = 1, \dots, L$. To this end, note that conditionally on $\boldsymbol{G}^{N,j} = (G_1^{N,j}, \dots, G_r^{N,j})$, we can interpret $P(\tilde{\boldsymbol{G}}^N_{\le j}, \tilde{\boldsymbol{s}}_{>j})$ as a noncommutative polynomial $P_N(\boldsymbol{G}^{N,j})$ with coefficients in $\mathrm{M}_N(\mathbb{C})^{\otimes(j-1)} \otimes \mathbf{A}$, where $\mathbf{A} \subseteq \mathcal{A}^{\otimes_{\min}(L-j)}$ denotes the operator space spanned by the monomials of $\tilde{\boldsymbol{s}}_{>j}$ that appear in P_N.

As exactness is stable under the minimal tensor product [49, p. 297], it follows that $\mathbf{A}$ is exact. Fix any $C > 1$, and let $u : \mathbf{A} \to \mathrm{M}_S(\mathbb{C})$ be the embedding provided by Definition 9.1. As the polynomial $(\mathrm{id} \otimes u)(P_N)$ has matrix coefficients of dimension $D_N = N^{j-1}S = e^{o(N)}$, applying Corollary 1.2 conditionally yields

$$\begin{aligned}\|P_N(\boldsymbol{G}^{N,j})\| &\le C\|(\mathrm{id}\otimes u)(P_N(\boldsymbol{G}^{N,j}))\| = (1+o(1))C\|(\mathrm{id}\otimes u)(P_N(\boldsymbol{s}))\| \\ &\le C(1+o(1))\|P_N(\boldsymbol{s})\|_{\min} = C(1+o(1))\|P(\tilde{\boldsymbol{G}}^N_{\le j-1}, \tilde{\boldsymbol{s}}_{>j-1})\|_{\min}\end{aligned}$$

a.s. as $N \to \infty$, where we used [49, Proposition 2.1.1] in the second inequality. Taking $C \downarrow 1$ yields an upper bound of the desired form. The corresponding lower bound follows in a completely analogous fashion, concluding the proof. □

Even though Lemma 9.6 provides a short proof of strong convergence of Hayes' model, the exactness argument provides no quantitative information. The main result of this section is the following quantitative form of Lemma 9.6.

Theorem 9.7. *Let $\varepsilon \in (0,1]$, and fix $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle\tilde{\boldsymbol{s}}\rangle$ of degree q_0 with matrix coefficients of dimension $D \le e^{cN\varepsilon^2}$. Then we have*

$$\mathbf{P}\big[\|P(\tilde{\boldsymbol{G}}^N)\| \ge (1+\varepsilon)\|P(\tilde{\boldsymbol{s}})\|_{\min}\big] \le \frac{N^L}{c\varepsilon}\, e^{-cN\varepsilon^2},$$

where $\frac{1}{c} = (CLr)^{2q_0} q_0^2$ for a universal constant C.

The proof of Theorem 9.7 is very similar to that of Theorem 1.1, with one twist. In our main results, we could deduce qualitative strong convergence merely from the fact that $\operatorname{supp}\nu_1 \subseteq [-\|X_{\mathrm{F}}\|, \|X_{\mathrm{F}}\|]$; this was then used as input to the bootstrapping argument (cf. section 1.2.4) to control the remaining ν_k. However, as the random matrices in the present section are N^L-dimensional rather than N-dimensional, we would need to control $\nu_1, \dots, \nu_L$ to prove strong convergence. In contrast to ν_1, control of $\nu_2, \dots, \nu_L$ cannot be achieved by supersymmetric duality, and would ordinarily require a problem-specific analysis as in [15].

Fortunately, as we already established strong convergence in a different manner in Lemma 9.6, we can use the latter as input to the bootstrapping argument and avoid any additional computations. The proof of Theorem 9.7 therefore illustrates the fact that the bootstrapping argument can be used to amplify a qualitative strong convergence result to a strong quantitative bound.

Proof of Theorem 9.7. Fix $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle\tilde{\boldsymbol{s}}\rangle$ as in the statement, and define $X^N = P(\tilde{\boldsymbol{G}}^N)$ and $X_{\mathrm{F}} = P(\tilde{\boldsymbol{s}})$. As any word $w(\tilde{\boldsymbol{G}}^N)$ of length q has the form

$$w(\tilde{\boldsymbol{G}}^N) = \big(G^{N,1}_{k_1}\cdots G^{N,1}_{k_{\ell_1}}\big) \otimes \big(G^{N,2}_{k_{\ell_1+1}}\cdots G^{N,2}_{k_{\ell_2}}\big) \otimes \cdots \otimes \big(G^{N,L}_{k_{\ell_{L-1}+1}}\cdots G^{N,L}_{k_q}\big)$$

for some $0 \le \ell_1 \le \cdots \le \ell_{L-1} \le q$ and $k_1, \ldots, k_q \in [r]$,

$$\mathbf{E}[\operatorname{tr} w(\tilde{\boldsymbol{G}}^N)] = \mathbf{E}[\operatorname{tr} G^{N,1}_{k_1} \cdots G^{N,1}_{k_{\ell_1}}]\, \mathbf{E}[\operatorname{tr} G^{N,2}_{k_{\ell_1+1}} \cdots G^{N,2}_{k_{\ell_2}}] \cdots \mathbf{E}[\operatorname{tr} G^{N,L}_{k_{\ell_{L-1}+1}} \cdots G^{N,L}_{k_q}]$$

is a polynomial of $\frac{1}{N^2}$ of degree at most $\frac{q}{4}$ as is noted in the proof of Lemma 4.7. In particular, the statement of Lemma 4.7 extends to the present setting.

With this observation in place, the entire proof of Theorem 1.1 for GUE matrices carries over directly to the present setting with two minor modifications: we replace the argument of section 5.1 by Lemma 9.6, and we note the correct normalization $\mathbf{E}[\operatorname{Tr} h(X^N)] = DN^L\, \mathbf{E}[\operatorname{tr} h(X^N)]$ in the present setting in section 5.3. □

We have considered the GUE form of the Hayes model in this section for concreteness. It is straightforward to repeat the above arguments to obtain the analogous result for the GOE/GSE or $\mathrm{U}(N)/\mathrm{O}(N)/\mathrm{Sp}(N)$ ensembles.

9.4. **Tensor GUE models.** In Hayes' model of the previous section, independent GUE matrices act on disjoint factors of a tensor product, that is, they are non-interacting. In this section, we investigate tensor GUE models that admit a general interaction pattern. Such models arise naturally in the study of quantum many-body systems [43, 21] and random geometry [40].

To this end, fix $L, V \in \mathbb{N}$ and nonempty subsets $K_1, \ldots, K_V \subseteq [L]$. For every $v \in [V]$, we define an independent N^L-dimensional random matrix

$$\hat{G}^N_v = H^N_v \otimes \mathbf{1}_{N^{|[L]\setminus K_v|}} \quad \text{in} \quad \bigotimes_{\ell\in[L]} \mathrm{M}_N(\mathbb{C}) \simeq \bigotimes_{\ell\in K_v} \mathrm{M}_N(\mathbb{C}) \otimes \bigotimes_{\ell\in[L]\setminus K_v} \mathrm{M}_N(\mathbb{C}),$$

where H^N_v is GUE of dimension $N^{|K_v|}$. Thus $\hat{\boldsymbol{G}}^N = (\hat{G}^N_v)_{v\in[V]}$ are independent GUE matrices that act on overlapping tensor factors.

To describe the limiting model, let $\Gamma = ([V], E)$ be a finite simple graph. A Γ-*independent semicircular family* is a family $\hat{\boldsymbol{s}} = (\hat{s}_v)_{v\in[V]}$ in a C^*-probability space with the following properties: each $\hat{s}_v$ is a semicircular variable; $\hat{s}_v, \hat{s}_w$ are classically independent if $\{v, w\} \in E$; and $\hat{s}_v, \hat{s}_w$ are freely independent if $v \neq w$, $\{v, w\} \notin E$. We refer to [53, §3] for the precise definition.

It was shown by Charlesworth and Collins [14, Theorem 4] that $\hat{\boldsymbol{G}}^N$ converges weakly to $\hat{\boldsymbol{s}}$ as $N \to \infty$ (in the sense of Lemma 2.11), where the graph Γ is defined by placing an edge $\{v, w\} \in E$ if and only if $K_v \cap K_w = \varnothing$. We will fix this graph Γ in the remainder of this section. It is readily seen that any finite simple graph Γ can be realized in this manner, as is noted in [14] and [40].

Whether $\hat{\boldsymbol{G}}^N$ converges strongly to $\hat{\boldsymbol{s}}$ as $N \to \infty$ has remained an open problem, cf. [40, Problem 1.6] and [21]. We resolve this problem here.

Theorem 9.8. *For every $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \hat{\boldsymbol{s}} \rangle$, we have*

$$\|P(\hat{\boldsymbol{G}}^N)\| = (1 + o(1))\|P(\hat{\boldsymbol{s}})\| \quad \textit{a.s.} \quad \textit{as} \quad N \to \infty.$$

As the random matrices $\hat{G}^N_v$ have dimension N^L, a direct application of the polynomial method to the present model would require us to control the supports of the infinitesimal distributions $\nu_1, \ldots, \nu_L$. Instead, we will take a different approach that circumvents the need for such an analysis.

The idea behind the proof is to use the central limit theorem to approximate the present model by the Hayes model of the previous section. To this end, we define in $\bigotimes_{\ell\in[L]} \mathrm{M}_N(\mathbb{C}) \simeq \bigotimes_{\ell\in K_v} \mathrm{M}_N(\mathbb{C}) \otimes \bigotimes_{\ell\in[L]\setminus K_v} \mathrm{M}_N(\mathbb{C})$ the random matrix

$$\hat{G}_v^{N,T} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \big(G_{v,t}^{N,1} \otimes G_{v,t}^{N,2} \otimes \cdots \otimes G_{v,t}^{N,|K_v|}\big) \otimes \mathbf{1}_{N^{|[L]\setminus K_v|}},$$

where $(G_{v,t}^{N,i})_{v,t,i}$ are independent GUE matrices of dimension N. Similarly, we define in[6] $\bigotimes_{\ell\in[L]} \mathcal{A} \simeq \bigotimes_{\ell\in K_v} \mathcal{A} \otimes \bigotimes_{\ell\in[L]\setminus K_v} \mathcal{A}$ the associated limiting model

$$\hat{s}_v^T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \big(s_{v,t} \otimes \cdots \otimes s_{v,t}\big) \otimes \mathbf{1},$$

where $(s_{v,t})_{v,t}$ is a free semicircular family in $\mathcal{A}$. In the following, we will write $\hat{\boldsymbol{G}}^{N,T} = (\hat{G}_v^{N,T})_{v\in[V]}$ and $\hat{\boldsymbol{s}}^T = (\hat{s}_v^T)_{v\in[V]}$.

Lemma 9.9. *The mean and covariance of the real and imaginary parts of the entries of $\hat{\boldsymbol{G}}^{N,T}$ coincide with those of $\hat{\boldsymbol{G}}^N$ for every $T \in \mathbb{N}$.*

Proof. Let $G = G^1 \otimes \cdots \otimes G^r$ where $G^1, \ldots, G^r$ are independent GUE matrices of dimension N, and let G' be a GUE matrix of dimension N^r acting on $(\mathbb{C}^N)^{\otimes r}$. Both G and G' have zero mean and are self-adjoint, and

$$\begin{aligned}\mathbf{E}\big[G_{(i_1,\ldots,i_r),(j_1,\ldots,j_r)} \bar{G}_{(k_1,\ldots,k_r),(l_1,\ldots,l_r)}\big] &= \mathbf{E}\big[G^1_{i_1,j_1}\bar{G}^1_{k_1,l_1}\big] \cdots \mathbf{E}\big[G^r_{i_r,j_r}\bar{G}^r_{k_r,l_r}\big] \\ &= \frac{1_{i_1=k_1} 1_{j_1=l_1} \cdots 1_{i_r=k_r} 1_{j_r=l_r}}{N^r} = \mathbf{E}\big[G'_{(i_1,\ldots,i_r),(j_1,\ldots,j_r)} \bar{G}'_{(k_1,\ldots,k_r),(l_1,\ldots,l_r)}\big].\end{aligned}$$

Thus the real and imaginary parts of the entries of G and G' have the same mean and covariance. The proof is readily completed. □

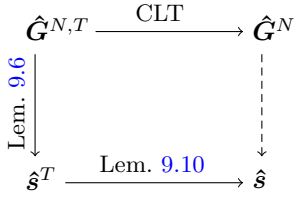
Lemma 9.9 ensures that $\hat{\boldsymbol{G}}^{N,T}$ converges in distribution to $\hat{\boldsymbol{G}}^N$ as $T \to \infty$ by the central limit theorem. The following lemma yields the corresponding property for the limiting models; its proof will be given in Appendix B.

Lemma 9.10. *For every $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \hat{\boldsymbol{s}} \rangle$, we have*

$$\|P(\hat{\boldsymbol{s}}^T)\| = (1 + o(1))\|P(\hat{\boldsymbol{s}})\| \quad \textit{as} \quad T \to \infty.$$

On the other hand, strong convergence of $\hat{\boldsymbol{G}}^{N,T}$ to $\hat{\boldsymbol{s}}^T$ as $N \to \infty$ follows from Lemma 9.6, as for any $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \hat{\boldsymbol{s}} \rangle$ of degree q_0, $P(\hat{\boldsymbol{G}}^{N,T})$ may be viewed as a noncommutative polynomial of degree at most Lq_0 of the Hayes model of the previous section with $r = VT$ free variables.

The above relations between the different models are summarized as follows:



[6] All tensor products of C^*-algebras in this section are minimal tensor products.

The key to the proof of Theorem 9.8 is that we have strong quantitative forms of the convergence of $\hat{\boldsymbol{G}}^{N,T}$ both as $N \to \infty$ (by Theorem 9.7) and as $T \to \infty$ (by the universality principle of [11]). This enables us to exchange the order of the limits $T \to \infty$ and $N \to \infty$ to deduce strong convergence of $\hat{\boldsymbol{G}}^N$ to $\hat{\boldsymbol{s}}$.

Proof of Theorem 9.8. The lower bound $\|P(\hat{\boldsymbol{G}}^N)\| \ge (1+o(1))\|P(\hat{\boldsymbol{s}})\|$ a.s. follows readily from weak convergence [14] and concentration of measure as in the proof of Lemma 9.3. It remains to prove the corresponding upper bound.

In order to apply the results of [11], it is convenient to use a classical linearization trick of Haagerup and Thorbjørnsen: by [28, Lemma 1], it suffices to show that

$$\mathrm{sp}(P(\hat{\boldsymbol{G}}^N)) \subseteq \mathrm{sp}(P(\hat{\boldsymbol{s}})) + [-\varepsilon,\varepsilon] \quad \text{eventually as } N \to \infty \quad \text{a.s.}$$

for every $\varepsilon > 0$, $D \in \mathbb{N}$, and self-adjoint $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \hat{\boldsymbol{s}} \rangle$ of degree $q_0 = 1$. We will fix such a polynomial P in the rest of the proof.

Now note that $P(\hat{\boldsymbol{G}}^{N,T}) = A_0 \otimes \mathbf{1}_{N^L} + \frac{1}{\sqrt{T}} \sum_{v,t} A_{v,t} \otimes Z_{v,t}$, where each $Z_{v,t}$ is an independent tensor product of at most L GUE matrices. In particular,

$$\mathbf{P}\Big[\max_{v,t} \|Z_{v,t}\| > C\Big] \le CTe^{-cN}, \quad \mathbf{E}\Big[\max_{v,t} \|Z_{v,t}\|^2\Big]^{1/2} \le C\bigg(1 + \sqrt{\frac{\log T}{N}}\bigg)^L$$

by Lemma 4.4, where the constant C may depend on L, V but not on N, T. We can therefore apply [11, Theorem 2.8] and Lemma 9.9 to obtain

$$\mathbf{P}\big[\mathrm{sp}(P(\hat{\boldsymbol{G}}^N)) \subseteq \mathrm{sp}(P(\hat{\boldsymbol{G}}^{N,T})) + [-\varepsilon(t),\varepsilon(t)]\big] \ge 1 - N^L e^{-t} - CTe^{-cN}$$

for all $t > 0$ and $T \le e^N$, where $\varepsilon(t) = C_P(N^{-1/2}t^{1/2} + T^{-1/12}t^{2/3} + T^{-1/4}t)$ and C_P is a constant that depends on P. Choosing $t = (L+2)\log N$ yields

$$\mathrm{sp}(P(\hat{\boldsymbol{G}}^N)) \subseteq \mathrm{sp}(P(\hat{\boldsymbol{G}}^{N,T_N})) + [-\varepsilon,\varepsilon] \quad \text{eventually as } N \to \infty \quad \text{a.s.}$$

for every $\varepsilon > 0$ by the Borel-Cantelli lemma, where $T_N := \lceil \log^9 N \rceil$.

On the other hand, for any $h \in \mathcal{P}_q$, we may view $h(P(\hat{\boldsymbol{G}}^{N,T_N}))$ as a noncommutative polynomial of degree at most Lq of the Hayes model of the previous section with $r = VT_N = O(\log^9 N)$ variables. Thus Theorem 9.7 and Borel-Cantelli yield

$$\|h(P(\hat{\boldsymbol{G}}^{N,T_N}))\| \le (1+o(1))\|h(P(\hat{\boldsymbol{s}}^{T_N}))\| \quad \text{a.s.} \quad \text{as} \quad N \to \infty$$

for every $h \in \mathcal{P}$. By Lemma 9.10 and [18, Proposition 2.1], this implies

$$\mathrm{sp}(P(\hat{\boldsymbol{G}}^{N,T_N})) \subseteq \mathrm{sp}(P(\hat{\boldsymbol{s}})) + [-\varepsilon,\varepsilon] \quad \text{eventually as } N \to \infty \quad \text{a.s.}$$

for every $\varepsilon > 0$. Combining the above estimates concludes the proof. □

We emphasize that the above argument relies fundamentally on the quantitative form of strong convergence for the Hayes model provided by Theorem 9.7. Indeed, $P(\hat{\boldsymbol{G}}^{N,T_N})$ defines a sequence of noncommutative polynomials in the Hayes model with an increasing number of variables $r = O(\log^9 N)$. We achieve uniform control of the error as the constant in Theorem 9.7 depends polynomially on r.

Now that strong convergence of the tensor GUE model has been established in a qualitative sense, Theorem 9.8 could be used as input to a bootstrapping argument

to obtain a quantitative strong convergence theorem along the lines of Theorem 9.7. As this does not require any new idea, we omit the details.

9.5. **High-dimensional representations.** When we considered Haar-distributed random matrices in the classical compact groups of dimension N, we implicitly identified elements of the group with their fundamental representation as N-dimensional matrices. It is of considerable interest to understand strong convergence of other representations whose dimension may be much larger than N. The aim of this section is to prove the following result in this direction.

Theorem 9.11. *Let $V_1^N,\ldots,V_r^N$ be i.i.d. Haar-distributed elements of $\mathrm{SU}(N)$ and let $\boldsymbol{u}=(u_1,\ldots,u_r)$ be free Haar unitaries. Fix $\delta>0$, and let π_N be any nontrivial[7] unitary representation of $\mathrm{SU}(N)$ with $\dim(\pi_N)\le\exp(N^{1/3-\delta})$ for all N. Then*

$$\|P(\pi_N(V_1^N),\ldots,\pi_N(V_r^N),\pi_N(V_1^N)^*,\ldots,\pi_N(V_r^N)^*)\| = (1+o(1))\|P(\boldsymbol{u},\boldsymbol{u}^*)\| \text{ a.s.}$$

as $N\to\infty$ for every $P\in\mathrm{M}_D(\mathbb{C})\otimes\mathbb{C}\langle\boldsymbol{u},\boldsymbol{u}^*\rangle$.

Theorem 9.11 improves a result of Magee and de la Salle [37], who prove the same statement for $\dim(\pi_N)\le\exp(N^{1/24-\delta})$. Our aim here is to showcase how the methods of this paper give rise to quantitative improvements.

Following [37], we will work with representations of $\mathrm{U}(N)$ rather than $\mathrm{SU}(N)$ in the proof. Recall that the distinct irreducible representations of $\mathrm{U}(N)$ are indexed by their highest weight vectors, which are N-tuples $\boldsymbol{z}=(z_1,\ldots,z_N)\in\mathbb{Z}^N$ such that $z_1\ge\cdots\ge z_N$. It will be convenient to parametrize $\boldsymbol{z}$ as

$$\boldsymbol{z}=(\lambda_1,\ldots,\lambda_{\ell(\lambda)},0,\ldots,0,-\mu_{\ell(\mu)},\ldots,-\mu_1),$$

where $\lambda\vdash L$ and $\mu\vdash M$ are two integer partitions with $\ell(\lambda)+\ell(\mu)\le N$; here and in the sequel, we denote by $\ell(\lambda)$ the number of elements of a partition λ. We denote the associated unitary representation of $\mathrm{U}(N)$ by $\pi_{\lambda,\mu}$, and we define

$$\pi_{L,M}:=\bigoplus_{\substack{\lambda\vdash L\\ \mu\vdash M}}\pi_{\lambda,\mu}.$$

By a variant of Schur-Weyl duality [36, §2.2], we have $\dim(\pi_{L,M})\le N^{L+M}$. The significance of this definition is that, on the one hand, every irreducible representation of $\mathrm{SU}(N)$ of dimension up to e^{cR} is contained in $\pi_{L,M}$ for some $L+M\le R$ [37, Proposition 2.1]; while on the other hand, $\pi_{L,M}$ is stable for $N\ge L+M$ and thus gives rise to rational expressions for spectral statistics.

In the following, fix $L,M,N\in\mathbb{Z}_+$ with $1\le L+M\le N$, i.i.d. Haar-distributed $\boldsymbol{U}^N=(U_1^N,\ldots,U_r^N)$ in $\mathrm{U}(N)$, and $P\in\mathrm{M}_D(\mathbb{C})\otimes\mathbb{C}\langle\boldsymbol{u},\boldsymbol{u}^*\rangle$ of degree q_0, and let

$$X^N:=P(\pi_{L,M}(\boldsymbol{U}^N),\pi_{L,M}(\boldsymbol{U}^N)^*),\qquad X_{\mathrm{F}}:=P(\boldsymbol{u},\boldsymbol{u}^*).$$

Then we have the following analogue of Lemma 7.1.

Lemma 9.12 (Polynomial encoding)**.** *Let $Q:=(L+M)q_0$. For every $h\in\mathcal{P}_q$, there is a rational function $\Psi_h:=\frac{f_h}{g_{Qq}}$ with $f_h,g_{Qq}\in\mathcal{P}_{\lfloor 4Qq(1+\log Qq)\rfloor}$ so that*

$$N^{-(L+M)}\mathbf{E}[\mathrm{Tr}\,h(X^N)]=\Psi_h(\tfrac{1}{N})=\Psi_h(-\tfrac{1}{N})$$

[7] We call a representation π nontrivial if it does not contain the trivial representation.

for all $N \in \mathbb{N}$ such that $N > Qq$. Here g_{Qq} is defined as in (7.1).

Proof. The rational representation is an immediate consequence of the first part of [37, Theorem 3.1]. That $\Psi_h(\frac{1}{N}) = \Psi_h(-\frac{1}{N})$ is proved in Appendix C. □

With Lemma 9.12 in hand, we can now repeat all the arguments in sections 7.3–7.6 identically in the present setting, except that we use the strong convergence result of [37] as input to the bootstrapping argument (as in section 9.3, the argument of section 5.1 fails here as X^N is not N-dimensional). This yields the following.

Theorem 9.13. *Let $\varepsilon \in [\frac{A}{c\sqrt{N}}, 1]$, and assume that $D \le e^{cN\varepsilon^2/A^2\log^2(N\varepsilon^2)}$. Then*

$$\mathbf{P}\big[\|P(\pi_{L,M}(\boldsymbol{U}^N), \pi_{L,M}(\boldsymbol{U}^N)^*)\| \ge (1+\varepsilon)\|P(\boldsymbol{u}, \boldsymbol{u}^*)\|\big] \le \frac{N^{L+M}}{c\varepsilon} e^{-cN\varepsilon^2/A^2\log^2(N\varepsilon^2)}.$$

Here c depends only on q_0 and r, and $A := (L+M)(1+\log(L+M))$.

We can now conclude the proof of Theorem 9.11.

Proof of Theorem 9.11. For any $1 \le R \le N^{1/3}$, we can estimate

$$\mathbf{P}\Big[\max_{1\le L+M\le R} \|P(\pi_{L,M}(\boldsymbol{U}^N), \pi_{L,M}(\boldsymbol{U}^N)^*)\| \ge (1+\varepsilon)\|P(\boldsymbol{u}, \boldsymbol{u}^*)\|\Big] \le \frac{R^2 N^R}{c\varepsilon} e^{-cN\varepsilon^2/R^2(1+\log R)^2\log^2(N\varepsilon^2)}$$

for $\varepsilon \in [\frac{R(1+\log R)}{c\sqrt{N}}, 1]$ using Theorem 9.13 and a union bound. In particular, if we choose $R \le N^{1/3-\delta}$ and $\varepsilon = N^{-\delta}$ for any $\delta > 0$ sufficiently small, the right-hand side of this inequality is bounded by $e^{-cN^{1/3}/\log^4 N}$ and thus

$$\max_{1\le L+M\le N^{1/3-\delta}} \|P(\pi_{L,M}(\boldsymbol{U}^N), \pi_{L,M}(\boldsymbol{U}^N)^*)\| \le (1+o(1))\|P(\boldsymbol{u}, \boldsymbol{u}^*)\| \quad \text{a.s.}$$

as $N \to \infty$ by the Borel-Cantelli lemma. The converse inequality then follows automatically by [37, Lemma 5.14]. The conclusion about strong convergence of representations of $\mathrm{SU}(N)$ now follows as in [37, §8]. □

Remark 9.14. The proof of Theorem 9.13 is not independent of [37], as we have used the main result of that paper as input to the bootstrapping argument. We note, however, that the bootstrapping argument only requires strong convergence of $\pi_{K,L}$ as $N \to \infty$ for *fixed* K, L. This is considerably simpler to achieve than the case of growing K, L; in particular, it does not require parts 2–3 of [37, Theorem 3.1] whose proof is based on delicate group-theoretic cancellations. (Alternatively, the main result of [8] would also have sufficed for this purpose.)

This considerable simplification was made possible here by exploiting the fact that $\mathrm{U}(N)$ has strong concentration of measure properties. We emphasize, however, that there are many interesting situations where this is not the case. This issue arises in particular in the analogue of Theorem 9.11 where $\mathrm{SU}(N)$ is replaced by the symmetric group S_N that was recently proved by Cassidy [13], whose analysis requires the full strength of the methods developed in [37].

Appendix A. An upper bound on the coefficient dimension

The aim of this Appendix is to explain the observation, due to Pisier [50], that strong convergence of the classical ensembles must fail for matrix coefficients of dimension $D = e^{O(N^2)}$. As this is not stated explicitly in [50], we provide a self-contained proof. We focus on GUE/GOE/GSE matrices for simplicity.

Lemma A.1 (Pisier)**.** *Let $\boldsymbol{G}^N$ and $\boldsymbol{s}$ be defined as in Theorem 1.1. Then there exists $r \in \mathbb{N}$ and matrices $A_1^N, \ldots, A_r^N \in \mathrm{M}_{D_N}(\mathbb{C})$ with $D_N = e^{O(N^2)}$ so that*

$$\left\| \sum_{i=1}^r A_i^N \otimes G_i^N \right\| \geq (2 + o(1)) \left\| \sum_{i=1}^r A_i^N \otimes s_i \right\| \quad \text{a.s.} \quad \text{as} \quad N \to \infty.$$

Proof. Let $\mathcal{B}_N$ be an optimal ε-net of $\{M \in \mathrm{M}_N(\mathbb{C})_{\mathrm{sa}} : \|M\| \leq 3\}$ with respect to the operator norm. It is classical that $\#\mathcal{B}_N \leq (\frac{C}{\varepsilon})^{N^2}$. We define $A_1^N, \ldots, A_r^N$ to be the block-diagonal matrices whose corresponding $N \times N$ blocks range over all r-tuples of elements of $\mathcal{B}_N$. Thus $D_N \leq N(\frac{C}{\varepsilon})^{rN^2}$.

Now note that $\|G_i^N\| = 2 + o(1)$ a.s. Therefore, a.s. for all sufficiently large N, there are $B_1^N, \ldots, B_r^N \in \mathcal{B}_N$ so that $\|\bar{G}_i^N - B_i^N\| \leq \varepsilon$, where $\bar{M}$ denotes the elementwise complex conjugate of M. By construction

$$\left\| \sum_{i=1}^r A_i^N \otimes G_i^N \right\| \geq \left\| \sum_{i=1}^r B_i^N \otimes G_i^N \right\| \geq \left\| \sum_{i=1}^r \bar{G}_i^N \otimes G_i^N \right\| - 3r\varepsilon$$

a.s. for all sufficiently large N. But note that

$$\left\| \sum_{i=1}^r \bar{G}_i^N \otimes G_i^N \right\| \geq \left\langle v_N, \left(\sum_{i=1}^r \bar{G}_i^N \otimes G_i^N \right) v_N \right\rangle = \sum_{i=1}^r \mathrm{tr}(G_i^{N*} G_i^N) = (1 + o(1))r$$

as $N \to \infty$ a.s. by Lemma 2.11 and concentration of measure, where we defined $v_N := \frac{1}{\sqrt{N}} \sum_{j=1}^N e_j \otimes e_j$. On the other hand, we can estimate

$$\left\| \sum_{i=1}^r A_i^N \otimes s_i \right\| \leq 2 \left\| \sum_{i=1}^r (A_i^N)^2 \right\|^{1/2} \leq 6\sqrt{r}$$

by the free Khintchine inequality [49, eq. (9.9.8)]. The conclusion follows readily, for example, by choosing the parameters $\varepsilon = \frac{1}{6}$ and $\sqrt{r} \geq 24$. □

The proof is readily adapted to achieve the same conclusion in the setting of Theorem 1.5; then one may use [49, eq. (9.7.1)] instead of the free Khintchine inequality. However, the situation for other ensembles may be even more restrictive. For example, for random permutations as in section 9.2, the set $\mathcal{B}_N$ can be replaced by S_N to show that strong convergence fails already when $D_N = e^{O(N \log N)}$.

Appendix B. Strong approximation of tensor models

In this appendix we adopt the setting and notations of section 9.4. We aim to prove Lemma 9.10. We begin by noting the following polynomial encoding.

Lemma B.1. *For every $P \in \mathrm{M}_D(\mathbb{C}) \otimes \mathbb{C}\langle \hat{\boldsymbol{s}} \rangle$, there is a polynomial Φ so that*

$$\mathbf{E}[\mathrm{tr}\, P(\hat{\boldsymbol{G}}^{N,T})] = \Phi(\tfrac{1}{N^2}, \tfrac{1}{T}) \quad \text{for all } N, T \in \mathbb{N}.$$

Proof. Let $Z_{v,t}$ be independent random matrices whose distribution does not depend on t, and let $Z_v^T = T^{-1/2}\sum_{t=1}^T Z_{v,t}$. In the following, we fix indices $v_1,\ldots,v_q$. Then $c(\pi(t_1,\ldots,t_q)) := \mathbf{E}[\mathrm{tr}\, Z_{v_1,t_1}\cdots Z_{v_q,t_q}]$ only depends on $(t_1,\ldots,t_q)$ through the partition $\pi(t_1,\ldots,t_q) \in \mathrm{P}([q])$ whose elements are $\{i\in[q] : t_i = t\}$. Thus

$$\mathbf{E}[\mathrm{tr}\, Z_{v_1}^T\cdots Z_{v_q}^T] = \frac{1}{T^{q/2}}\sum_{\pi\in\mathrm{P}([q])} T(T-1)\cdots(T-|\pi|+1)\,c(\pi).$$

If in addition $Z_{v,t}$ has the same distribution as $-Z_{v,t}$, then $c(\pi)=0$ unless q is even and $|\pi|\le\frac{q}{2}$. It is then evident that $\mathbf{E}[\mathrm{tr}\, Z_{v_1}^T\cdots Z_{v_q}^T]$ is a polynomial of $\frac{1}{T}$.

The special case $Z_v^T = \hat G_v^{N,T}$ satisfies the above conditions, and moreover each $c(\pi)$ is a polynomial of $\frac{1}{N^2}$ as noted in the proof of Theorem 9.7. The conclusion follows in the case $P(\hat{\boldsymbol{s}}) = \hat s_{v_1}\cdots\hat s_{v_q}$, and extends to general P by linearity. □

Lemma B.1 readily implies the following weak convergence statement.

Corollary B.2. *For every $P\in \mathrm{M}_D(\mathbb{C})\otimes\mathbb{C}\langle\boldsymbol{s}\rangle$, we have*

$$(\mathrm{tr}\otimes\tau)(P(\hat{\boldsymbol{s}}^T)) = (1+o(1))\,(\mathrm{tr}\otimes\tau)(P(\hat{\boldsymbol{s}})) \quad \text{as} \quad T\to\infty.$$

Proof. By Lemma 9.9, the central limit theorem, and weak convergence of $\hat{\boldsymbol{G}}^N$ [14]

$$\lim_{N\to\infty}\lim_{T\to\infty}\mathbf{E}[\mathrm{tr}\,P(\hat{\boldsymbol{G}}^{N,T})] = \lim_{N\to\infty}\mathbf{E}[\mathrm{tr}\,P(\hat{\boldsymbol{G}}^N)] = (\mathrm{tr}\otimes\tau)(P(\hat{\boldsymbol{s}})).$$

On the other hand, Lemma 2.11 yields

$$\lim_{T\to\infty}\lim_{N\to\infty}\mathbf{E}[\mathrm{tr}\,P(\hat{\boldsymbol{G}}^{N,T})] = \lim_{T\to\infty}(\mathrm{tr}\otimes\tau)(P(\hat{\boldsymbol{s}}^T)).$$

To conclude the proof, it remains to note that Lemma B.1 enables us to exchange the order of the limits as $N\to\infty$ and $T\to\infty$. □

To upgrade the weak convergence of Corollary B.2 to norm convergence, we will use that the models $\hat{\boldsymbol{s}}^T$ satisfy a Haagerup inequality uniformly in T.

Lemma B.3. *There exist constants $C,a>0$ so that $\|P(\hat{\boldsymbol{s}}^T)\|\le Cq^a\|P(\hat{\boldsymbol{s}}^T)\|_{L^2(\tau)}$ for every $T,q\in\mathbb{N}$ and noncommutative polynomial $P\in\mathbb{C}\langle\hat{\boldsymbol{s}}\rangle$ of degree q.*

Proof. Define in $\bigotimes_{\ell\in[L]} C^*_{\mathrm{red}}(\mathrm{F}_R) \simeq \bigotimes_{\ell\in K_v} C^*_{\mathrm{red}}(\mathrm{F}_R)\otimes\bigotimes_{\ell\in[L]\setminus K_v} C^*_{\mathrm{red}}(\mathrm{F}_R)$

$$\hat s_v^{T,S} := \frac{1}{\sqrt{T}}\sum_{t=1}^T\big(s_{v,t}^S\otimes\cdots\otimes s_{v,t}^S\big)\otimes\mathbf{1}$$

where $s_{v,t}^S := \frac{1}{\sqrt{2S}}\sum_{s=1}^S(\lambda(g_{v,t,s})+\lambda(g_{v,t,s})^*)$, and let $\hat{\boldsymbol{s}}^{T,S} = (\hat s_v^{T,S})_{v\in V}$. Here $R = VTS$, and $\{g_{v,t,s} : v\in[V], t\in[T], s\in[S]\}$ are the free generators of F_R.

By the Haagerup inequality for the free group F_R [26, Lemma 1.4] and stability of Haagerup inequalities under direct products [33, Lemma 2.1.2] (see also [16]), there exist $C,a>0$ such that for all $p,T,S\in\mathbb{N}$ and $P\in\mathbb{C}\langle\hat{\boldsymbol{s}}\rangle$ of degree q

$$\|P(\hat{\boldsymbol{s}}^{T,S})\|_{L^{2p}(\tau)}\le Cq^a\|P(\hat{\boldsymbol{s}}^{T,S})\|_{L^2(\tau)}.$$

As $\hat{\boldsymbol{s}}^{T,S}$ converges weakly to $\hat{\boldsymbol{s}}^T$ as $S\to\infty$ by the free central limit theorem [44, Theorem 8.17], the conclusion follows by taking $S\to\infty$ and then $p\to\infty$. □

We can now complete the proof of Lemma 9.10.

Proof of Lemma 9.10. We first note that by Corollary B.2,

$$\|P(\hat{\boldsymbol{s}}^T)\| \ge \|P(\hat{\boldsymbol{s}}^T)\|_{L^{2p}(\mathrm{tr}\otimes\tau)} = (1+o(1))\|P(\hat{\boldsymbol{s}})\|_{L^{2p}(\mathrm{tr}\otimes\tau)} \quad \text{as } T\to\infty$$

for every $p \in \mathbb{N}$. Taking $p \to \infty$ yields the lower bound.

Next, we apply Lemma B.3 as in the proof of [3, Theorem 4.1] to estimate

$$\|P(\hat{\boldsymbol{s}}^T)\| \le D^{3/4p}(Cpq_0)^{a/2p}\|P(\hat{\boldsymbol{s}}^T)\|_{L^{4p}(\mathrm{tr}\otimes\tau)}$$

for every $p \in \mathbb{N}$, where q_0 is the degree of P. The upper bound follows by first taking $T \to \infty$ on the right-hand side using Corollary B.2, and then $p \to \infty$. □

Appendix C. Supersymmetric duality for stable characters of U(N)

Lemma 9.12 relies on an unpublished result of Magee which states that every stable representation of U(N) exhibits a form of supersymmetric duality. We are most grateful to Michael Magee for allowing us to include the proof here.

Lemma C.1 (Magee). *Fix $L, M \in \mathbb{Z}_+$ with $L + M \ge 1$, let $\lambda \vdash L$ and $\mu \vdash M$, and let $w(U_1^N, \ldots, U_r^N)$ be any word in i.i.d. Haar-distributed elements of* U(N). *Then there exists a rational function Ψ such that*

$$\mathbf{E}[\mathrm{Tr}\,\pi_{\lambda,\mu}(w(U_1^N,\ldots,U_r^N))] = \Psi(\tfrac{1}{N}),$$
$$\mathbf{E}[\mathrm{Tr}\,\pi_{\bar\lambda,\bar\mu}(w(U_1^N,\ldots,U_r^N))] = (-1)^{L+M}\Psi(-\tfrac{1}{N})$$

for all $N \ge L + M$, where $\bar\lambda, \bar\mu$ denote the conjugate partitions of λ, μ.

Note that the duality statement in Lemma 9.12 follows directly from Lemma C.1, as both $\pi_{\lambda,\mu}$ and $\pi_{\bar\lambda,\bar\mu}$ appear in $\pi_{L,M}$ with multiplicity one.

Proof of Lemma C.1. For every permutation σ with cycle type $(\sigma_1, \ldots, \sigma_r)$, let

$$p_\sigma(U) := \mathrm{Tr}[U^{\sigma_1}]\,\mathrm{Tr}[U^{\sigma_2}]\cdots\mathrm{Tr}[U^{\sigma_r}].$$

By combining [34, eq. (0.3)] and [25, p. 78, eq. (12)], we can express the character $\mathrm{Tr}\,\pi_{\lambda,\mu}(U)$ for any $U \in \mathrm{U}(N)$ by the explicit formula

$$\mathrm{Tr}\,\pi_{\lambda,\mu}(U) = \sum_{\alpha,\beta,\gamma} \frac{(-1)^{|\alpha|}c^\lambda_{\alpha,\beta}c^\mu_{\bar\alpha,\gamma}}{|\beta|!|\gamma|!} \sum_{\sigma\in\mathrm{S}_{|\beta|}}\sum_{\sigma'\in\mathrm{S}_{|\gamma|}} \chi_\beta(\sigma)\,\chi_\gamma(\sigma')\,p_\sigma(U)\,p_{\sigma'}(U^*),$$

where $c^\lambda_{\alpha,\beta}$ are Littlewood-Richardson coefficients, χ_β is the character of $\mathrm{S}_{|\beta|}$ associated to β, and the sum is over all integer partitions α, β, γ; here and below, we write $|\alpha| := k$ for $\alpha \vdash k$. The sum is finite as $c^\lambda_{\alpha,\beta} = 0$ unless $|\alpha| + |\beta| = |\lambda|$.

Denote $W^N := w(U_1^N, \ldots, U_r^N)$. Then [38, Proposition 1.1 and Remark 1.9] imply that for every $\sigma \in \mathrm{S}_k$ and $\sigma' \in \mathrm{S}_l$, there is a rational function $\Psi_{\sigma,\sigma'}$ so that

$$\mathbf{E}[p_\sigma(W^N)\,p_{\sigma'}(W^{N*})] = \Psi_{\sigma,\sigma'}(\tfrac{1}{N}) = (-1)^{c(\sigma)+c(\sigma')}\Psi_{\sigma,\sigma'}(-\tfrac{1}{N})$$

for all $N \ge k + l$, where $c(\sigma)$ denotes the number of cycles of σ.

Now recall that for every $\sigma \in \mathrm{S}_k$ and $\alpha \vdash k$, we have

$$\chi_{\bar\alpha}(\sigma) = \mathrm{sgn}(\sigma)\chi_\alpha(\sigma) = (-1)^{k-c(\sigma)}\chi_\alpha(\sigma).$$

The proof is concluded by noting that

$$(-1)^{L+M}\sum_{\alpha,\beta,\gamma}\frac{(-1)^{|\alpha|}c^{\lambda}_{\alpha,\beta}c^{\mu}_{\bar\alpha,\gamma}}{|\beta|!|\gamma|!}\sum_{\sigma\in\mathrm{S}_{|\beta|}}\sum_{\sigma'\in\mathrm{S}_{|\gamma|}}\chi_\beta(\sigma)\,\chi_\gamma(\sigma')\,\Psi_{\sigma,\sigma'}(-\tfrac{1}{N}) =$$
$$\sum_{\alpha,\beta,\gamma}\frac{(-1)^{|\alpha|}c^{\lambda}_{\alpha,\beta}c^{\mu}_{\bar\alpha,\gamma}}{|\beta|!|\gamma|!}\sum_{\sigma\in\mathrm{S}_{|\beta|}}\sum_{\sigma'\in\mathrm{S}_{|\gamma|}}\chi_{\bar\beta}(\sigma)\,\chi_{\bar\gamma}(\sigma')\,\Psi_{\sigma,\sigma'}(\tfrac{1}{N}) = \mathbf{E}[\mathrm{Tr}\,\pi_{\bar\lambda,\bar\mu}(W^N)],$$

where we used $L=|\alpha|+|\beta|$, $M=|\alpha|+|\gamma|$, and that $c^{\lambda}_{\alpha,\beta}=c^{\bar\lambda}_{\bar\alpha,\bar\beta}$ (cf. [25, p. 62]). □

Acknowledgments. We thank Pierre Deligne, Michael Magee, Félix Parraud, and Mikael de la Salle for helpful discussions. We are grateful to Michael Magee for allowing us to include the argument of Appendix C in this paper, and to Mikael de la Salle for suggesting Remark 1.6. This work was completed while RvH was a member of the Institute for Advanced Study in Princeton, NJ, which is gratefully acknowledged for providing a fantastic mathematical environment.

CFC was supported in part by a Simons-CIQC postdoctoral fellowship through NSF grant QLCI-2016245. JGV was supported in part by Joel A. Tropp through NSF grant FRG-1952777, ONR grant N-00014-24-1-2223, and the Carver Mead New Adventures Fund, and in part by Caltech IST and a Baer–Weiss CMI Fellowship. RvH was supported in part by NSF grant DMS-2347954.

References

[1] G. Aubrun, S. J. Szarek, and D. Ye. Entanglement thresholds for random induced states. *Comm. Pure Appl. Math.*, 67(1):129–171, 2014.

[2] A. S. Bandeira, M. T. Boedihardjo, and R. van Handel. Matrix concentration inequalities and free probability. *Invent. Math.*, 234(1):419–487, 2023.

[3] A. S. Bandeira, G. Cipolloni, D. Schröder, and R. van Handel. Matrix concentration inequalities and free probability II. Two-sided bounds and applications, 2024. Preprint arxiv:2406.11453.

[4] S. Belinschi and M. Capitaine. Strong convergence of tensor products of independent GUE matrices, 2022. Preprint arXiv:2205.07695.

[5] P. Biane and R. Speicher. Stochastic calculus with respect to free Brownian motion and analysis on Wigner space. *Probab. Theory Related Fields*, 112(3):373–409, 1998.

[6] C. Bordenave and B. Collins. Eigenvalues of random lifts and polynomials of random permutation matrices. *Ann. of Math. (2)*, 190(3):811–875, 2019.

[7] C. Bordenave and B. Collins. Norm of matrix-valued polynomials in random unitaries and permutations, 2024. Preprint arxiv:2304.05714v2.

[8] C. Bordenave and B. Collins. Strong asymptotic freeness for independent uniform variables on compact groups associated to nontrivial representations. *Invent. Math.*, 237(1):221–273, 2024.

[9] C. Borell. The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, 30(2):207–216, 1975.

[10] P. Borwein and T. Erdélyi. *Polynomials and polynomial inequalities*. Springer Science & Business Media, 2012.

[11] T. Brailovskaya and R. van Handel. Universality and sharp matrix concentration inequalities. *Geom. Funct. Anal.*, 34(6):1734–1838, 2024.

[12] W. Bryc and V. Pierce. Duality of real and quaternionic random matrices. *Electr. J. Probab.*, 14:452–476, 2009.

[13] E. Cassidy. Random permutations acting on k-tuples have near-optimal spectral gap for $k=\mathrm{poly}(n)$, 2024. Preprint.

[14] I. Charlesworth and B. Collins. Matrix models for ε-free independence. *Arch. Math. (Basel)*, 116(5):585–600, 2021.
[15] C.-F. Chen, J. Garza-Vargas, J. A. Tropp, and R. van Handel. A new approach to strong convergence, 2024. Preprint arXiv:2405.16026.
[16] L. Ciobanu, D. F. Holt, and S. Rees. Rapid decay is preserved by graph products. *J. Topol. Anal.*, 5(2):225–237, 2013.
[17] B. Collins, A. Guionnet, and F. Parraud. On the operator norm of non-commutative polynomials in deterministic matrices and iid GUE matrices. *Camb. J. Math.*, 10(1):195–260, 2022.
[18] B. Collins and C. Male. The strong asymptotic freeness of Haar and deterministic matrices. *Ann. Sci. Éc. Norm. Supér. (4)*, 47(1):147–163, 2014.
[19] B. Collins and S. Matsumoto. On some properties of orthogonal Weingarten functions. *J. Math. Phys.*, 50(11), 2009.
[20] B. Collins, S. Matsumoto, and J. Novak. The Weingarten calculus. *Notices Amer. Math. Soc.*, 69(5):734–745, 2022.
[21] B. Collins and W. Yuan. Strong convergence for tensor GUE random matrices, 2024. Preprint arXiv:2407.09065.
[22] D. Coppersmith and T. J. Rivlin. The growth of polynomials bounded at equally spaced points. *SIAM J. Math. Anal.*, 23(4):970–983, 1992.
[23] P. Deligne. La catégorie des représentations du groupe symétrique S_t, lorsque t n'est pas un entier naturel. In *Algebraic groups and homogeneous spaces*, volume 19 of *Tata Inst. Fund. Res. Stud. Math.*, pages 209–273. Tata Inst. Fund. Res., Mumbai, 2007.
[24] P. Erdős. On extremal properties of the derivatives of polynomials. *Ann. of Math. (2)*, 41:310–313, 1940.
[25] W. Fulton. *Young tableaux*, volume 35 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 1997.
[26] U. Haagerup. An example of a nonnuclear C^*-algebra, which has the metric approximation property. *Invent. Math.*, 50(3):279–293, 1978/79.
[27] U. Haagerup. Quasitraces on exact C^*-algebras are traces. *C. R. Math. Acad. Sci. Soc. R. Can.*, 36(2-3):67–92, 2014.
[28] 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.
[29] U. Haagerup and S. Thorbjørnsen. Asymptotic expansions for the Gaussian unitary ensemble. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.*, 15(1):1250003, 41, 2012.
[30] B. Hayes. A random matrix approach to the Peterson-Thom conjecture. *Indiana Univ. Math. J.*, 71(3):1243–1297, 2022.
[31] L. Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis*. Springer-Verlag, Berlin, 2003.
[32] J. Huang, T. McKenzie, and H.-T. Yau. Optimal eigenvalue rigidity of random regular graphs, 2024. Preprint arxiv:2405.12161.
[33] P. Jolissaint. Rapidly decreasing functions in reduced C^*-algebras of groups. *Trans. Amer. Math. Soc.*, 317(1):167–196, 1990.
[34] K. Koike. On the decomposition of tensor products of the representations of the classical groups: by means of the universal characters. *Adv. Math.*, 74(1):57–86, 1989.
[35] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.
[36] M. Magee. Random unitary representations of surface groups II: The large n limit. *Geom. Topol*, 2024. To appear.
[37] M. Magee and M. de la Salle. Strong asymptotic freeness of Haar unitaries in quasi-exponential dimensional representations, 2024. Preprint arXiv:2409.03626.
[38] M. Magee and D. Puder. Matrix group integrals, surfaces, and mapping class groups I: $U(n)$. *Invent. Math.*, 218:341–411, 2019.
[39] M. Magee and D. Puder. Matrix group integrals, surfaces, and mapping class groups II: $O(n)$ and $Sp(n)$. *Math. Ann.*, 388(2):1437–1494, 2024.

[40] M. Magee and J. Thomas. Strongly convergent unitary representations of right-angled Artin groups, 2023. Preprint arxiv:2308.00863.
[41] E. S. Meckes. *The random matrix theory of the classical compact groups*, volume 218 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2019.
[42] J. A. Mingo and R. Speicher. *Free probability and random matrices*. Springer, 2017.
[43] S. C. Morampudi and C. R. Laumann. Many-body systems with random spatially local interactions. *Phys. Rev. B*, 100:245152, 2019.
[44] 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.
[45] F. Parraud. On the operator norm of non-commutative polynomials in deterministic matrices and iid Haar unitary matrices. *Probab. Theory Related Fields*, 182(3-4):751–806, 2022.
[46] F. Parraud. Asymptotic expansion of smooth functions in deterministic and iid Haar unitary matrices, and application to tensor products of matrices, 2023. Preprint arXiv:2302.02943.
[47] F. Parraud. Asymptotic expansion of smooth functions in polynomials in deterministic matrices and iid GUE matrices. *Comm. Math. Phys.*, 399(1):249–294, 2023.
[48] F. Parraud. The spectrum of a tensor of random and deterministic matrices, 2024. Preprint arXiv:2410.04481.
[49] G. Pisier. *Introduction to operator space theory*, volume 294 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2003.
[50] G. Pisier. Random matrices and subexponential operator spaces. *Israel J. Math.*, 203:223–273, 2014.
[51] E. A. Rakhmanov. Bounds for polynomials with a unit discrete norm. *Ann. of Math. (2)*, 165(1):55–88, 2007.
[52] H. Schultz. Non-commutative polynomials of independent Gaussian random matrices. The real and symplectic cases. *Probab. Theory Related Fields*, 131:261–309, 2005.
[53] R. Speicher and J. Wysoczański. Mixtures of classical and free independence. *Arch. Math. (Basel)*, 107(4):445–453, 2016.
[54] T. Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.
[55] D. Voiculescu. Limit laws for random matrices and free products. *Invent. Math.*, 104(1):201–220, 1991.

EECS, University of California, Berkeley, CA 94720, USA
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Email address: achifchen@gmail.com

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA
Email address: jgarzav@caltech.edu

PACM, Princeton University, Princeton, NJ 08544, USA
Email address: rvan@math.princeton.edu