On the operator norm of non-commutative polynomials in deterministic matrices and iid GUE matrices

Benoît Collins[1], Alice Guionnet[2], and Félix Parraud[1,2]

[1]Department of Mathematics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan.
[2]Université de Lyon, CNRS, ENSL, 46 allée d'Italie, 69007 Lyon.

Abstract

Let $X^N = (X_1^N, \dots, X_d^N)$ be a d-tuple of $N \times N$ independent GUE random matrices and Z^{NM} be any family of deterministic matrices in $\mathbb{M}_N(\mathbb{C}) \otimes \mathbb{M}_M(\mathbb{C})$. Let P be a self-adjoint non-commutative polynomial. A seminal work of Voiculescu shows that the empirical measure of the eigenvalues of $P(X^N)$ converges towards a deterministic measure defined thanks to free probability theory. Let now f be a smooth function, the main technical result of this paper is a precise bound of the difference between the expectation of

$$\frac{1}{MN} \operatorname{Tr}_{\mathbb{M}_N(\mathbb{C})} \otimes \operatorname{Tr}_{\mathbb{M}_M(\mathbb{C})} \left(f(P(X^N \otimes I_M, Z^{NM})) \right) ,$$

and its limit when N goes to infinity. If f is six times differentiable, we show that it is bounded by $M^2 \|f\|_{\mathcal{C}^6} N^{-2}$. As a corollary we obtain a new proof and slightly improve a result of Haagerup and Thorbjørnsen, later developed by Male, which gives sufficient conditions for the operator norm of a polynomial evaluated in (X^N, Z^{NM}, Z^{NM*}) to converge almost surely towards its free limit.

1 Introduction

Given several deterministic matrices whose spectra are known, the spectra of a non-commutative polynomial evaluated in these matrices is not well defined since it depends as well on the eigenvectors of these matrices. If one takes these vectors at random, it is possible to get some surprisingly good results, in particular when the dimension of these matrices goes to infinity. Indeed, the limit can then be computed thanks to free probability. This theory was introduced by Voiculescu in the early nineties as a non-commutative probability theory equipped with a notion of freeness analogous to independence in classical probability theory. Voiculescu showed that this theory was closely related with Random Matrix Theory in a seminal paper [30]. He considered independent matrices taken from the Gaussian Unitary Ensemble (GUE), which are random matrix is an $N \times N$ self-adjoint random matrix whose distribution is proportional to the measure $\exp\left(-N/2 \operatorname{Tr}_N(A^2)\right) dA$, where dA denotes the Lebesgue measure on the set of $N \times N$ Hermitian matrices. We refer to Definition 2.8 for a more precise statement. Voiculescu proved that given $X_1^N, \dots, X_d^N$ independent GUE matrices, the renormalized trace of a polynomial P evaluated in these matrices converges towards a deterministic limit $\alpha(P)$. Specifically, the following holds true almost surely:

$$\lim_{N \to \infty} \frac{1}{N} \operatorname{Tr}_N \left(P(X_1^N, \dots, X_d^N) \right) = \alpha(P) . \tag{1}$$

Voiculescu computed the limit $\alpha(P)$ with the help of free probability. If A_N is a self-adjoint matrix of size N, then one can define the empirical measure of its (real) eigenvalues by

$$\mu_{A_N} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i} ,$$

where δ_λ is the Dirac mass in λ and $\lambda_1, \ldots, \lambda_N$ are the eingenvalue of A_N. In particular, if P is a self-adjoint polynomial, that is such that for any self adjoint matrices $A_1, \ldots, A_d$, $P(A_1, \ldots, A_d)$ is a self-adjoint matrix, then one can define the random measure $\mu_{P(X_1^N, \ldots, X_d^N)}$. In this case, Voiculescu's result (1) implies that there exists a measure μ_P with compact support such that almost surely $\mu_{P(X_1^N, \ldots, X_d^N)}$ converges weakly towards μ_P : it is given by $\mu_P(x^k) = \alpha(P^k)$ for all integer numbers k.

However, the convergence of the empirical measure of the eigenvalues of a matrix does not say anything about the local properties of its spectrum, in particular about the convergence of the norm of this matrix, or the local fluctuations of its spectrum. When dealing with a single matrix, incredibly precise results are known. For exemple it is well-known that the largest eigenvalue of a GUE random matrix converges almost surely towards 2. More precisely, if X_N is a GUE random matrix of size N, then almost surely

$$\lim_{N \to \infty} \|X_N\| = 2 \ .$$

The proof, for the more general case of a Wigner matrix with entries with finite moments, was given in [13]. This result was later obtained under the optimal assumption that their fourth moment is finite [3]. Concerning the GUE, much more precise results were obtained by Tracy and Widom in the early nineties in [29]. The main result of their paper is the existence of a continuous decreasing function F_2 from $\mathbb{R}$ to $[0, 1]$ such that if $\lambda_1(X^N)$ denotes the largest eigenvalue of X^N,

$$\lim_{N \to \infty} P\big(N^{2/3}(\lambda_1(X^N) - 2) \geq s\big) = F_2(s) \ .$$

This was recently generalized to Wigner matrices [27, 11, 28, 18] up to optimal hypotheses. One can as well study the localization of the eigenvalues in the bulk as well as their fluctuations [10, 11].

On the other hand, there are much less results available when one deals with a polynomial in several random matrices. In fact, up to today, the only local fluctuations results concern perturbative polynomials [12] or local laws [9] under some assumptions which are shown to hold for homogeneous polynomials of degree two. However, a beautiful breakthrough was made in 2005 by Haagerup and Thorbjørnsen [17]: they proved the almost sure convergence of the norm of a polynomial evaluated in independent GUE matrices. For P a self-adjoint polynomial, they proved that almost surely, for N large enough,

$$\sigma\left(P(X_1^N, \ldots, X_d^N)\right) \subset \operatorname{Supp} \mu_P + (-\varepsilon, \varepsilon) \ , \tag{2}$$

where $\sigma(H)$ is the spectrum of H and $\operatorname{Supp} \mu_P$ the support of the measure μ_P. This is equivalent to saying that for any polynomial P, $\left\|P(X_1^N, \ldots, X_d^N)\right\|$ converges almost surely towards $\sup\{|x| \ | x \in \operatorname{Supp} \mu_P\}$ (see proposition 2.2). The result (2) was a major progress in free probability. It was was refined in multiple ways. In [25], Schultz used the method of [17] to prove the same result with Gaussian orthogonal or symplectic matrices instead of Gaussian unitary matrices. In [6], Capitaine and Donati-Martin proved it for Wigner matrices under some technical hypothesis on the law of the entries. This result itself was then extended by Anderson in [1] to remove most of the technical assumptions. In [19], Male made a conceptual improvement to the result of Haagerup and Thorbjørnsen, by allowing to work both with GUE and deterministic matrices. Finally, Belinschi and Capitaine proved in [7] that one could even work with Wigner and determinisic matrices, while keeping the same assumptions on the Wigner matrices as Anderson. It is also worth noting that Collins and Male proved in [8] the same result with unitary Haar matrices instead of GUE matrices by using Male's former paper.

With the exception of [8], all of these results are essentially based on the method introduced by Haagerup and Thorbjørnsen. Their first tool is called the linearization trick: it allows to relate the spectrum of a polynomial of degree d with coefficients in $\mathbb{C}$ by a polynomial of degree 1 with coefficients in $\mathbb{M}_{k(d)}(\mathbb{C})$. The second idea to understand the spectrum of the spectral measure of this larger matrix is to study its Stieltjes transform close to the real axis by using the Dyson-Schwinger equations. An issue of this method is that it does not give easily good quantitative estimates. One aim of this paper is to remedy to this problem. We develop a new method that allows us to give a new proof of the main theorem of Male in [19], and thus a new proof of the result of Haagerup and Thorbjørnsen. Our approach requires neither the linearization trick, nor the study of the Stieljes transform and attacks the problem directly. In this sense the proof is more direct and less algebraic. We will apply it to a generalization of GUE matrices by tackling the case of GUE random matrices tensorized with deterministic matrices.

A usual strategy to study outliers, that are the eigenvalues going away from the spectrum, is to study the *non-renormalized* trace of smooth non-polynomial functions evaluated in independent GUE matrices i.e. if P is self-adjoint:

$$\mathrm{Tr}_N\left(f(P(X_1^N,\dots,X_d^N))\right) \ .$$

This strategy was also used by Haagerup, Thorbjørnsen and Male. Indeed it is easy to see that if f is a function which takes value 0 on $(-\infty, C-\varepsilon]$, 1 on $[C,\infty)$ and in $[0,1]$ elsewhere, then

$$\mathbb{P}\Big(\lambda_1(P(X_1^N,\dots,X_d^N)) \geq C\Big) \leq \ \mathbb{P}\Big(\ \mathrm{Tr}_N\left(f(P(X_1^N,\dots,X_d^N))\right) \geq 1\Big)$$

Hence, if we can prove that $\mathrm{Tr}_N\left(f(P(X_1^N,\dots,X_d^N))\right)$ converges towards 0 in probability, this would already yield expected results. The case where f is a polynomial function has already been studied a long time ago, starting with the pioneering works [5, 16], and later formalized by the concept of second order freeness [20]. However here we have to deal with a function f which is at best C^∞. This makes things considerably more difficult and forces us to adopt a completely different approach. The main result is the following Theorem. For the notations, we refer to Section 2 – for now, let us specify that $\frac{1}{N}\mathrm{Tr}_N$ denotes the usual renormalized trace on $N\times N$ matrices whereas τ denotes its free limit.

Theorem 1.1. *Let the following objects be given,*

- $X^N = (X_1^N,\dots,X_d^N)$ *independent GUE matrices in* $\mathbb{M}_N(\mathbb{C})$,
- $x = (x_1,\dots,x_d)$ *a system of free semicircular variables,*
- $Z^{NM} = (Z_1^{NM},\dots,Z_q^{NM})$ *deterministic matrices in* $\mathbb{M}_N(\mathbb{C})\otimes\mathbb{M}_M(\mathbb{C})$,
- $P\in\mathbb{C}\langle X_1,\dots,X_{d+2q}\rangle_{sa}$ *a self-adjoint polynomial,*
- $f\in\mathcal{C}^6(\mathbb{R})$.

Then there exists a polynomial L_P *which only depends on* P *such that for any* N, M,

$$\left|\mathbb{E}\left[\frac{1}{MN}\mathrm{Tr}_{MN}\left(f\left(P\left(X^N\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\right)\right] - \tau_N\otimes\tau_M\Big(f\left(P\left(x\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\Big)\right|$$
$$\leq \frac{M^2}{N^2}\,\|f\|_{\mathcal{C}^6}\,L_P\left(\left\|Z^{NM}\right\|\right) \ ,$$

where $\|f\|_{\mathcal{C}^6}$ *is the sum of the supremum on* $\mathbb{R}$ *of the first six derivatives. Besides if* $Z^{NM} = (I_N\otimes Y_1^M,\dots,I_N\otimes Y_q^M)$ *and that these matrices commute, then we have the same inequality without the* M^2.

This theorem is a consequence of the slightly sharper, but less explicit, Theorem 3.1. It is essentially the same statement, but instead of having the norm C^6 of f, we have the fourth moment of the Fourier transform of f. The above Theorem calls for a few remarks.

- We assumed that the matrices Z^{NM} were deterministic, but thanks to Fubini's Theorem we can assume that they are random matrices as long as they are independent from X^N. In this situation though, $L_P\left(\left\|Z^{NM}\right\|\right)$ in the right side of the inequality is a random variable (and thus we need some additional assumptions if we want its expectation to be finite for instance).
- In Theorems 1.1 and 3.1 we have $X^N\otimes I_M$ and $x\otimes I_M$, however it is very easy to replace them by $X^N\otimes Y^M$ and $x\otimes Y^M$ for some matrices $Y_i^M\in M_M(\mathbb{C})$. Indeed we just need to apply Theorem 1.1 or 3.1 with $Z^{NM} = I_N\otimes Y^M$. Besides, in this situation, $L_P\left(\left\|Z^{NM}\right\|\right) = L_P\left(\left\|Y^M\right\|\right)$ does not depend on N. What this means is that if we have a matrix whose coefficients are polynomial in X^N, and that we replace X^N by x, we only change the spectra of this matrix by M^2N^{-2} in average.
- Unfortunately we cannot get rid of the M^2 in all generality. The specific case where we can is when $Z^{NM} = (I_N\otimes Y_1^M,\dots,I_N\otimes Y_q^M)$, where the Y_i^M commute: this indicates that the M^2 term is really a non-commutative feature.

A detailed overview of the proof is given in Subsection 3.1. The main idea of the proof is to use a free version of Stein's method by interpolating GUE matrices with a free semicircular system with the help of a free Ornstein-Uhlenbeck process. For a reference, see [4]. When using this process, the Schwinger-Dyson equations, which can be seen as an integration by part formula, appear in the computation. We we refer to Proposition 2.10 for more information which will play a major role in this paper. Theorem 1.1 is the crux of the paper and allows us to deduce many corollaries. Firstly we rederive a new proof of the following theorem. The first statement is basically Theorem 1.6 from [19]. The second one is an improvement of Theorem 7.8 from [23] on the size of the tensor from $N^{1/4}$ to $N^{1/3}$. This theorem is about strong convergence of random matrices, that is the convergence of the norm of polynomials in these matrices, see definition 2.1.

Theorem 1.2. *Let the following objects be given:*

- $X^N = (X_1^N, \dots, X_d^N)$ *independent GUE matrices of size* N,
- $x = (x_1, \dots, x_d)$ *a system of free semicircular variable,*
- $Y^M = (Y_1^M, \dots, Y_p^M)$ *random matrices of size* M*, which almost surely, as* M *goes to infinity, converge strongly in distribution towards a* p*-tuple* y *of non-commutative random variables in a* $\mathcal{C}^*$*-probability space* $\mathcal{B}$ *with a faithful trace* $\tau_{\mathcal{B}}$,
- $Z^N = (Z_1^N, \dots, Z_q^N)$ *random matrices of size* N*, which almost surely, as* N *goes to infinity, converges strongly in distribution towards a* q*-tuple* z *of non-commutative random variables in a* $\mathcal{C}^*$*-probability space with a faithful trace.*

Then, the following holds true:

- *If* X^N *and* Z^N *are independent, almost surely,* (X^N, Z^N) *converges strongly in distribution towards* $\mathcal{F} = (x, z)$*, where* $\mathcal{F}$ *belongs to a* $\mathcal{C}^*$*- probability space* $(\mathcal{A}, *, \tau_{\mathcal{A}}, \|.\|)$ *in which* x *and* z *are free.*
- *If* X^N *and* Y^{M_N} *are independent and* $M_N = o(N^{1/3})$*, almost surely,* $(X^N \otimes I_{M_N}, I_N \otimes Y^{M_N})$ *converges strongly in distribution towards* $\mathcal{F} = (x \otimes 1, 1 \otimes y)$*. The family* $\mathcal{F}$ *thus belongs to* $\mathcal{A} \otimes_{\min} \mathcal{B}$ *(see definition 4.1). Besides if the matrices* Y^{M_N} *commute, then we can weaken the assumption on* M_N *by only assuming that* $M_N = o(N)$*.*

As we mentioned earlier, understanding the Stieljes transform of a matrix gives a lot of information about its spectrum. This was actually a very important point in the proof of Haagerup and Thorbjørnsen's Theorem. Our proof does not use this tool, however our final result, Theorem 3.1, allows us to deduce the following estimate with sharper constant than what has previously been done. Being given a self- adjoint $NM \times NM$ matrix, we denote by G_A its Stieltjes transform:

$$G_A(z) = \frac{1}{NM} \operatorname{Tr}_{NM} \left(\frac{1}{z - A} \right) .$$

This definition extends to the tensor product of free semi-circular variables by replacing Tr_{NM} by $\tau_N \otimes \tau_M$.

Corollary 1.3. *Given*

- $X^N = (X_1^N, \dots, X_d^N)$ *independent GUE matrices of size* N,
- $x = (x_1, \dots, x_d)$ *a system of free semicircular variable,*
- $Y^M = (Y_1^M, \dots, Y_p^M, Y_1^{M^*}, \dots, Y_p^{M^*})$ *deterministic matrices of size* M *a fixed integer and their adjoints,*
- $P \in \mathbb{C}\langle X_1, \dots, X_d, Y_1, \dots, Y_{2p} \rangle_{sa}$ *a self-adjoint polynomial,*

there exists a polynomial L_P *such that for every* Y^M*,* $z \in \mathbb{C} \backslash \mathbb{R}$*,* $N \in \mathbb{N}$*,*

$$\left| \mathbb{E}\left[G_{P(X^N \otimes I_M, I_N \otimes Y^M)}(z) \right] - G_{P(x \otimes I_M, I_N \otimes Y^M)}(z) \right| \leq L_P\left(\left\| Y^M \right\| \right) \frac{M^2}{N^2} \left(\frac{1}{|\Im(z)|^5} + \frac{1}{|\Im(z)|^2} \right) .$$

One of the limitation of Theorem 1.1 is that we need to pick f regular enough. Actually by approximating f, we can afford to take f less regular at the cost of a slower speed of convergence. In other words, we trade some degree of regularity on f for a smaller exponent in N. The best that we can achieve is to take f Lipschitz. Thus it makes sense to introduce the Lipschitz-bounded metric. This metric is compatible with the topology of the convergence in law of measure. Let $\mathcal{F}_{LU}$ be the set of Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$, uniformly bounded by 1 and with Lipschitz constant at most 1, then

$$d_{LU}(\mu,\nu) = \sup_{f\in\mathcal{F}_{LU}} \left| \int_{\mathbb{R}} f d\mu - \int_{\mathbb{R}} f d\nu \right| .$$

For more information about this metric we refer to Annex C.2 of [2]. In this paper, we get the following result:

Corollary 1.4. *Under the same notations as in Corollary 1.3, there exists a polynomial L_P such that for every matrices Y^M and $M, N \in \mathbb{N}$,*

$$d_{LU}\left(\mathbb{E}[\mu_{P(X^N\otimes I_M, I_N\otimes Y_M)}], \mu_{P(x\otimes I_M, I_N\otimes Y_M)}\right) \leq L_P\left(\left\|Y^M\right\|\right)\frac{M^2}{N^{1/3}} .$$

One of the advantage of Theorem 1.1 over the original proof of Haagerup and Thorbjørnsen is that if we take f which depends on N, we get sharper estimates in N. For exemple if we assume that g is a $\mathcal{C}^\infty$ function with bounded support, as we will see later in this paper we like to work with $f : x \mapsto g(N^\alpha x)$ for some constant α. Then its n-th derivative will be of order $N^{n\alpha}$. In the original work of Haagerup, Thorbjørnsen (see [17], Theorem 6.2) the eighth derivative appears for the easiest case where our polynomial P is of degree 1, and the order is even higher in the general case. But in Theorem 1.1 the sixth derivative appears in the general case. Actually if we look at the sharper Theorem 3.1, the fourth moment of the Fourier transform appears, which is roughly equivalent to the fourth derivative for our computations. This allows us to compute an estimate of the difference between $\mathbb{E}\left[\left\|P(X^N\otimes I_M, I_N\otimes Y^M)\right\|\right]$ and its limit. To do that, we use Proposition 4.8 from [26, Theorem 1.1] which implies that if we denote by $\mu_{P(x\otimes I_M, 1\otimes Y^M)}$ the spectral measure of $P(x\otimes I_M, 1\otimes Y^M)$, then there exists $\beta\in\mathbb{R}^+$ such that

$$\limsup_{\varepsilon\to 0} \quad \varepsilon^{-\beta}\mu_{P(x\otimes I_M,1\otimes Y^M)}\left(\left[\left\|P(x\otimes I_M,1\otimes Y^M)\right\| - \varepsilon, \left\|P(x\otimes I_M,1\otimes Y^M)\right\|\right]\right) > 0 . \tag{3}$$

With the help of standard measure concentration estimates, we then get the following Theorem:

Theorem 1.5. *We consider*

- $X^N = (X_1^N, \ldots, X_d^N)$ *independent GUE matrices of size N,*
- $x = (x_1, \ldots, x_d)$ *a system of free semicircular variable,*
- $Y^M = (Y_1^M, \ldots, Y_p^M)$ *deterministic matrices of size M a fixed integer and their adjoints.*

Almost surely, for any polynomial $P \in \mathbb{C}\langle X_1, \ldots, X_d, Y_1, \ldots, Y_p\rangle$, there exists constants K and C such that for any $\delta > 0$,

$$\mathbb{P}\left(N^{1/4}\left(\left\|P(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|P(x\otimes I_M, 1\otimes Y^M)\right\|\right) \geq \delta + C\right) \leq e^{-K\delta^2\sqrt{N}} + de^{-N} , \tag{4}$$

$$\mathbb{P}\left(N^{1/(3+\beta)}\left(\left\|P(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|P(x\otimes I_M, 1\otimes Y^M)\right\|\right) \leq -\delta - C\right) \leq e^{-K\delta^2 N^{\frac{1+\beta}{3+\beta}}} + de^{-N} . \tag{5}$$

This theorem is interesting because of its similarity with Tracy and Widom's result about the tail of the law of the largest eingenvalue of a GUE matrix. We have smaller exponent in N, and thus we can only show the convergence towards 0 with exponential speed, however we are not restricted to a single GUE matrix, we can chose any polynomial evaluated in GUE matrices. Besides by applying Borel-Cantelli's Lemma, we immediately get:

Theorem 1.6. *We consider*

- $X^N = (X_1^N, \dots, X_d^N)$ *independent GUE matrices of size* N,
- $x = (x_1, \dots, x_d)$ *a system of free semicircular variable,*
- $Y^M = (Y_1^M, \dots, Y_p^M)$ *deterministic matrices of size* M *a fixed integer and their adjoints.*

Then almost surely, for any polynomial $P \in \mathbb{C}\langle X_1, \dots, X_d, Y_1, \dots, Y_p\rangle$, *there exists a constant* $c(P) > 0$ *such that for any* $c(P) > \alpha > 0$,

$$\lim_{N\to\infty} N^{\alpha} \Big| \big\|P(X^N \otimes I_M, I_N \otimes Y^M)\big\| - \big\|P(x \otimes I_M, 1 \otimes Y^M)\big\| \Big| = 0 \ .$$

Moreover, if β *satisfies* (3), *then almost surely for any* $\alpha < (3+\beta)^{-1}$ *and* $\varepsilon < 1/4$, *for* N *large enough,*

$$-N^{-\alpha} \leq \big\|P(X^N \otimes I_M, I_N \otimes Y^M)\big\| - \big\|P(x \otimes I_M, 1 \otimes Y^M)\big\| \leq N^{-\varepsilon} \ .$$

In order to conclude this introduction, we would like to say that while it is not always easy to compute the constant β in all generality, it is possible for some polynomials. In particular, if our polynomial is evaluated in a single GUE matrix, then the computation is heavily simplified by the fact that we know the distribution of a single semicircular variable. Finally, the constant $(3+\beta)^{-1}$ is clearly a worst case scenario and can be easily improved if β is explicit.

This paper is organised as follows. In Section 2, we recall the definitions and properties of free probability, non-commutative calculus and Random Matrix Theory needed for this paper. Section 3 contains the proof of Theorem 1.1. And finally in Section 4 we give the proof of the remaining Theorem and Corollaries.

2 Framework and standard properties

2.1 Usual definitions in free probability

In order to be self-contained, we begin by reminding the following definitions from free probability.

Definition 2.1.

- *A* $\mathcal{C}^*$*-probability space* $(\mathcal{A}, *, \tau, \|.\|)$ *is a unital* $\mathcal{C}^*$*-algebra* $(\mathcal{A}, *, \|.\|)$ *endowed with a state* τ, *i.e. a linear map* $\tau : \mathcal{A} \to \mathbb{C}$ *satisfying* $\tau(1_{\mathcal{A}}) = 1$ *and* $\tau(a^*a) \geq 0$ *for all* $a \in \mathcal{A}$. *In this paper we always assume that* τ *is a trace, i.e. that it satisfies* $\tau(ab) = \tau(ba)$ *for any* $a, b \in \mathcal{A}$. *An element of* $\mathcal{A}$ *is called a (non commutative) random variable. We will always work with a faithful trace, namely, for* $a \in \mathcal{A}$, $\tau(a^*a) = 0$ *if and only if* $a = 0$. *In this case the norm is determined by* τ *thanks to the formula:*
$$\|a\| = \lim_{k\to\infty} \big(\tau\big((a^*a)^{2k}\big)\big)^{1/2k}.$$

- *Let* $\mathcal{A}_1, \dots, \mathcal{A}_n$ *be* $*$*-subalgebras of* $\mathcal{A}$, *having the same unit as* $\mathcal{A}$. *They are said to be free if for all* k, *for all* $a_i \in \mathcal{A}_{j_i}$ *such that* $j_1 \neq j_2$, $j_2 \neq j_3$, $\dots$, $j_{k-1} \neq j_k$:
$$\tau\Big((a_1 - \tau(a_1))(a_2 - \tau(a_2)) \dots (a_k - \tau(a_k))\Big) = 0.$$
Families of non-commutative random variables are said to be free if the $*$*-subalgebras they generate are free.*

- *Let* $A = (a_1, \dots, a_k)$ *be a* k*-tuple of random variables. The joint distribution of the family* A *is the linear form* $\mu_A : P \mapsto \tau\big[P(A, A^*)\big]$ *on the set of polynomials in* $2k$ *non commutative indeterminates. By convergence in distribution, for a sequence of families of variables* $(A_N)_{N\geq 1} = (a_1^N, \dots, a_k^N)_{N\geq 1}$ *in* $\mathcal{C}^*$*-algebras* $\big(\mathcal{A}_N, ^*, \tau_N, \|.\|\big)$, *we mean the pointwise convergence of the map*
$$\mu_{A_N} : P \mapsto \tau_N\big[P(A_N, A_N^*)\big],$$
and by strong convergence in distribution, we mean convergence in distribution, and pointwise convergence of the map
$$P \mapsto \big\|P(A_N, A_N^*)\big\|.$$

- *A family of non commutative random variables $x = (x_1, \ldots, x_p)$ is called a free semicircular system when the non commutative random variables are free, selfadjoint ($x_i = x_i^*$, $i = 1 \ldots p$), and for all k in $\mathbb{N}$ and $i = 1, \ldots, p$, one has*

$$\tau(x_i^k) = \int t^k d\sigma(t),$$

with $d\sigma(t) = \frac{1}{2\pi}\sqrt{4-t^2}\ \mathbf{1}_{|t|\leq 2}\ dt$ the semicircle distribution.

The strong convergence of non-commutative random variables is actually equivalent to the convergence of the spectrum of their polynomials for the Hausdorff distance. More precisely we have the following proposition whose proof can be found in [8, Proposition 2.1] :

Proposition 2.2. *Let $\mathbf{x}_N = (x_1^N, \ldots, x_p^N)$ and $\mathbf{x} = (x_1, \ldots, x_p)$ be p-tuples of variables in $\mathcal{C}^*$-probability spaces, $(\mathcal{A}_N, .^*, \tau_N, \|\cdot\|)$ and $(\mathcal{A}, .^*, \tau, \|\cdot\|)$, with faithful states. Then, the following assertions are equivalent.*

- *$\mathbf{x}_N$ converges strongly in distribution to $\mathbf{x}$.*
- *For any self-adjoint variable $h_N = P(\mathbf{x}_N)$, where P is a fixed polynomial, μ_{h_N} converges in weak-$*$ topology to μ_h where $h = P(\mathbf{x})$. Weak-$*$ topology means relatively to continuous functions on $\mathbb{C}$. Moreover, the spectrum of h_N converges in Hausdorff distance to the spectrum of h, that is, for any $\varepsilon > 0$, there exists N_0 such that for any $N \geq N_0$,*

$$\sigma(h_N) \ \subset \ \sigma(h) \ + (-\varepsilon, \varepsilon). \tag{6}$$

In particular, the strong convergence in distribution of a single self-adjoint variable is equivalent to its convergence in distribution together with the Hausdorff convergence of its spectrum.

It is important to note that thanks to [22, Theorem 7.9], that we recall below, one can consider free version of any random variable.

Theorem 2.3. *Let $(\mathcal{A}_i, \phi_i)_{i\in I}$ be a family of $\mathcal{C}^*$-probability spaces such that the functionals $\phi_i : \mathcal{A}_i \to \mathbb{C}$, $i \in I$, are faithful traces. Then there exist a $\mathcal{C}^*$-probability space $(\mathcal{A}, \phi)$ with ϕ a faithful trace, and a family of norm- preserving unital $*$-homomorphism $W_i : \mathcal{A}_i \to \mathcal{A}$, $i \in I$, such that:*

- *$\phi \circ W_i = \phi_i$, $\forall i \in I$.*
- *The unital $\mathcal{C}^*$-subalgebras form a free family in $(\mathcal{A}, \phi)$.*

Let us finally fix a few notations concerning the spaces and traces that we use in this paper.

Definition 2.4.
- *$(\mathcal{A}_N, \tau_N)$ is the free sum of $\mathbb{M}_N(\mathbb{C})$ with a system of d free semicircular variable, this is the $\mathcal{C}^*$- probability space built in Theorem 2.3. Note that when restricted to $\mathbb{M}_N(\mathbb{C})$, τ_N is just the regular renormalized trace on matrices. The restriction of τ_N to the $\mathcal{C}^*$-algebra generated by the free semicircular system x is denoted as τ.*
- *Tr_N is the non-renormalized trace on $\mathbb{M}_N(\mathbb{C})$.*
- *$\mathbb{M}_N(\mathbb{C})_{sa}$ is the set of self adjoint matrix of $\mathbb{M}_N(\mathbb{C})$. We denote $E_{r,s}$ the matrix with coefficients equal to 0 except in (r, s) where it is equal to one.*
- *We regularly identify $\mathbb{M}_N(\mathbb{C}) \otimes \mathbb{M}_k(\mathbb{C})$ with $\mathbb{M}_{kN}(\mathbb{C})$ through the isomorphism $E_{i,j} \otimes E_{r,s} \mapsto E_{i+rN,j+sN}$, similarly we identify $\mathrm{Tr}_N \otimes \mathrm{Tr}_k$ with Tr_{kN}.*
- *If $A^N = (A_1^N, \ldots, A_d^N)$ and $B^M = (B_1^M, \ldots, B_d^M)$ are two vectors of random matrices, then we denote $A^N \otimes B^M = (A_1^N \otimes B_1^M, \ldots, A_d^N \otimes B_d^M)$. We typically use the notation $X^N \otimes I_M$ for the vector $(X_1^N \otimes I_M, \ldots, X_1^N \otimes I_M)$.*

2.2 Non-commutative polynomials and derivatives

We set $\mathcal{A}_{d,q} = \mathbb{C}\langle X_1, \ldots, X_d, Y_1, \ldots, Y_q, Y_1^*, \ldots, Y_q^* \rangle$ the set of non-commutative polynomial in $d+2q$ indeterminates. We endow this vector space with the norm

$$\|P\|_A = \sum_{M \text{monomial}} |c_M(P)| A^{\deg M} \ , \tag{7}$$

where $c_M(P)$ is the coefficient of P for the monomial M and $\deg M$ the total degree of M (that is the sum of its degree in each letter $X_1, \ldots, X_d, Y_1, \ldots, Y_q, Y_1^*, \ldots, Y_q^*$). Let us define several maps which we use frequently in the sequel First, for $A, B, C \in \mathcal{A}_{d,q}$, let

$$A \otimes B \# C = ACB \ ,$$

$$A \otimes B \tilde{\#} C = BCA \ ,$$

$$m(A \otimes B) = BA \ .$$

Definition 2.5. *If $1 \leq i \leq d$, one defines the non-commutative derivative $\partial_i : \mathcal{A}_{d,q} \longrightarrow \mathcal{A}_{d,q} \otimes \mathcal{A}_{d,q}$ by its value on a monomial $M \in \mathcal{A}_{d,q}$ given by*

$$\partial_i M = \sum_{M = A X_i B} A \otimes B \ ,$$

and then extend it by linearity to all polynomials. Similarly one defines the cyclic derivative $D_i : \mathcal{A}_{d,q} \longrightarrow \mathcal{A}_{d,q}$ for $P \in \mathcal{A}_{d,q}$ by

$$D_i P = m \circ \partial_i P \ .$$

The map ∂_i is called the non-commutative derivative. It is related to Schwinger-Dyson equation on semicircular variable thanks to the following property 2.6. One can find a proof of the first part in [2], Lemma 5.4.7. As for the second part it is a direct consequence of the first one which can easily be verified by taking P monomial and then concluding by linearity.

Proposition 2.6. *Let $x = (x_1, \ldots, x_p)$ be a free semicircular system, $y = (y_1, \ldots, y_q)$ be non-commutative random variables free from x, if the family (x, y) belongs to the $\mathcal{C}^*$-probability space $(\mathcal{A}, *, \tau, \|.\|)$, then for any $P \in \mathcal{A}_{d,q}$,*

$$\tau(P(x, y, y^*)\ x_i) = \tau \otimes \tau(\partial_i P(x, y, y^*)) \ .$$

Moreover, one can deduce that if Z^{NM} are matrices in $\mathbb{M}_N(\mathbb{C}) \otimes \mathbb{M}_M(\mathbb{C})$ that we view as a subspace of $\mathcal{A}_N \otimes \mathbb{M}_M(\mathbb{C})$, then for any $P \in \mathcal{A}_{d,q}$,

$$\tau_N \otimes \tau_M \Big(P(x \otimes I_M, Z^{NM}, Z^{NM^*})\ x_i \otimes I_M \Big) = \tau_M \Big((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \big(\partial_i P(x \otimes I_M, Z^{NM}, Z^{NM^*}) \big) \Big) \ .$$

We define an involution $*$ on $\mathcal{A}_{d,q}$ such that

$$(X_i)^* = X_i \ , \quad (Y_i)^* = Y_i^* \ , \quad (Y_i^*)^* = Y_i$$

and then we extend it to $\mathcal{A}_{d,q}$ by the formula $(\alpha PQ)^* = \overline{\alpha} Q^* P^*$. $P \in \mathcal{A}_{d,q}$ is said to be self- adjoint if $P^* = P$. Self-adjoint polynomials have the property that if $x_1, \ldots, x_d, z_1, \ldots, z_q$ are elements of a $\mathcal{C}^*$-algebra such as $x_1, \ldots, x_d$ are self-adjoint, then so is $P(x_1, \ldots, x_d, z_1, \ldots, z_q, z_1^*, \ldots, z_q^*)$. Now that we have defined the notion of self-adjoint polynomial we remark for later use that

Proposition 2.7. *Let the following objects be given,*

- *$x = (x_1, \ldots, x_p)$ a free semicircular system ,*
- *$X^N = (X_1^N, \ldots, X_d^N)$ self-ajoint matrices of size N,*
- *$X_t^N = e^{-t/2} X^N + (1 - e^{-t})^{1/2} x$ elements of $\mathcal{A}_N$,*

- Z^{NM} *matrices in* $\mathbb{M}_N(\mathbb{C}) \otimes \mathbb{M}_M(\mathbb{C})$,
- $f \in \mathcal{C}^0(\mathbb{R})$,
- P *a self-adjoint polynomial.*

Then the following map is measurable:

$$(X^N, Z^{NM}) \mapsto \tau_N \otimes \tau_M \left(f \left(P(X_t^N \otimes I_M, Z^{NM}, Z^{NM*}) \right) \right).$$

Proof. This is obvious if f is a polynomial and the general case is obtained by approximation. □

Actually we could easily prove that this map is continuous, however we do not need it. The only reason we need this property is to justify that if X^N is a vector of d independent GUE matrices, then the random variable $\tau_N \otimes \tau_M \left(f \left(P(X_t^N \otimes I_M, Z^{NM}, Z^{NM*}) \right) \right)$ is well-defined and measurable.

2.3 GUE random matrices

We conclude this section by reminding the definition of Gaussian random matrices and stating a few useful properties about them.

Definition 2.8. *A GUE random matrix X^N of size N is a self adjoint matrix whose coefficients are random variables with the following laws:*

- *For $1 \leq i \leq N$, the random variables $\sqrt{N} X_{i,i}^N$ are independent centered Gaussian random variables of variance 1.*
- *For $1 \leq i < j \leq N$, the random variables $\sqrt{2N}\ \Re X_{i,j}^N$ and $\sqrt{2N}\ \Im X_{i,j}^N$ are independent centered Gaussian random variables of variance 1, independent of $\left(X_{i,i}^N\right)_i$.*

We now present two of the most useful tools when it comes to computation with Gaussian variable, the Poincaré inequality and Gaussian integration by part. Firstly, the Poincaré inequality:

Proposition 2.9. *Let $(x_1, \ldots, x_n)$ be i.i.d. centered Gaussian random variable with variance 1, let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^1$, then*

$$\mathrm{Var}\left(f(x_1, \ldots, x_n)\right) \leq \mathbb{E}\left[\, \|\nabla f(x_1, \ldots, x_n)\|_2^2 \,\right].$$

For more details about the Poincaré inequality, we refer to Definition 4.4.2 in [2]. As for Gaussian integration by part, it comes from the following formula, if Z is a centered Gaussian variable with variance 1 and f a $\mathcal{C}^1$ function, then

$$\mathbb{E}[Zf(Z)] = \mathbb{E}[\partial_Z f(Z)]\,. \tag{8}$$

A direct consequence of this, is that if x and y are centered Gaussian variable with variance 1, and $Z = \frac{x+\mathbf{i}y}{\sqrt{2}}$, then

$$\mathbb{E}[Zf(x,y)] = \mathbb{E}[\partial_Z f(x,y)] \quad \text{and} \quad \mathbb{E}[\overline{Z} f(x,y)] = \mathbb{E}[\partial_{\overline{Z}} f(x,y)]\,, \tag{9}$$

where $\partial_Z = \frac{1}{2}(\partial_x + \mathbf{i}\partial_y)$ and $\partial_{\overline{Z}} = \frac{1}{2}(\partial_x - \mathbf{i}\partial_y)$. When working with GUE matrices, an important consequence of this are the so-called Schwinger-Dyson equation, which we summarize in the following proposition. For more information about these equations and their applications, we refer to [2], Lemma 5.4.7.

Proposition 2.10. *Let X^N be GUE matrices of size N, $Q \in \mathcal{A}_{d,q}$, then for any i,*

$$\mathbb{E}\left[\frac{1}{N}\,\mathrm{Tr}_N(X_i^N\ Q(X^N))\right] = \mathbb{E}\left[\left(\frac{1}{N}\,\mathrm{Tr}_N\right)^{\otimes 2}(\partial_i Q(X^N))\right].$$

Proof. One can write $X_i^N = \frac{1}{\sqrt{N}}(x_{r,s}^i)_{1\leq r,s\leq N}$ and thus

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N}\operatorname{Tr}_N(X_i^N\ Q(X^N))\right] &= \frac{1}{N^{3/2}}\sum_{r,s}\mathbb{E}\left[x_{r,s}^i\ \operatorname{Tr}_N(E_{r,s}\ Q(X^N))\right]\\ &= \frac{1}{N^{3/2}}\sum_{r,s}\mathbb{E}\left[\operatorname{Tr}_N(E_{r,s}\ \partial_{x_{r,s}^i}Q(X^N))\right]\\ &= \frac{1}{N^2}\sum_{r,s}\mathbb{E}\left[\operatorname{Tr}_N(E_{r,s}\ \partial_i Q(X^N)\#E_{s,r})\right]\\ &= \mathbb{E}\left[\left(\frac{1}{N}\operatorname{Tr}_N\right)^{\otimes 2}(\partial_i Q(X^N))\right].\end{aligned}$$

□

Now to finish this section we state a property that we use several times in this paper:

Proposition 2.11. *There exist constants C, D and α such that for any $N\in\mathbb{N}$, if X^N is a GUE random matrix of size N, then for any $u\geq 0$,*

$$\mathbb{P}\left(\left\|X^N\right\|\geq u+D\right)\leq e^{-\alpha u N}.$$

Consequently, for any $k\leq \alpha N/2$,

$$\mathbb{E}\left[\left\|X^N\right\|^k\right]\leq C^k.$$

Proof. The first part is a direct consequence of Lemma 2.2 from [15] in the specific case of the GUE. As for the second part, if $k\leq \alpha N/2$, then we have,

$$\begin{aligned}\mathbb{E}\left[\left\|X^N\right\|^k\right] &= k\int_0^\infty \mathbb{P}\left(\left\|X^N\right\|\geq u\right)u^{k-1}du\\ &\leq kD^k + k\int_D^\infty e^{-N\alpha(u-D)}u^{k-1}du\\ &\leq kD^k + ke^{DN\alpha}\int_D^\infty e^{(k-N\alpha)u}du\\ &\leq kD^k + \frac{2k}{\alpha N}e^{kD}\leq C^k\end{aligned}$$

for some C independent of N and k. In the third line we used that $\ln|u|\leq u$ for all positive real numbers, □

3 Proof of Theorem 1.1

3.1 Overview of the proof

Given two families of non-commutative random variables, $(X^N\otimes I_M, Z^{NM})$ and $(x\otimes I_M, Z^{NM})$, and we want to study the difference between their distributions. As mentioned in the introduction, the main idea of the proof is to interpolate these two families with the help of a free Ornstein-Uhlenbeck process $X^{t,N}=(X_i^{t,N})_i$ started in deterministic matrices $(X_i^{N,0})_i$ of size N. However, as we shall explain in this subsection, we are only interested into the law of the marginals at time t of this process, hence we do not need to define it globally. We refer to [4] for more information about it. Some properties of this process are well understood. For example, like in the classical case, we know its distribution at time t. In the classical case, if $(S_t)_t$ was an Ornstein-Uhlenbeck process, then it is well-known that for any function f and $t\geq 0$,

$$\mathbb{E}[f(S_t)] = \mathbb{E}[f(e^{-t/2}S_0 + (1-e^{-t})^{1/2}X)]$$

where X is a centered Gaussian random variable of variance 1 independent of S_0. Likewise, if μ is the trace on the $\mathcal{C}^*$-algebra which contains $(X_t^N)_{t\geq 0}$, we have for any function f such that this is well-defined and $t \geq 0$,

$$\mu(f(X_t^N)) = \tau_N\left(f(e^{-t/2}X_0^N + (1-e^{-t})^{1/2}x)\right) \tag{10}$$

where x is a system of free semicircular variables, free from $\mathbb{M}_N(\mathbb{C})$. Thus a free Ornstein-Uhlenbeck process started at time t has the same distribution in the sense of Definition 2.1 as the family

$$e^{-t/2}X_0^N + (1-e^{-t})^{1/2}x \ .$$

Consequently, from now on, we write $X_t^N = e^{-t/2}X_0^N + (1-e^{-t})^{1/2}x$. Since our aim in this subsection is not to give a rigorous proof but to outline the strategy used in subsection 3.2, we also assume that we have no matrix Z^{NM} and that $M = 1$. Now under the assumption that this is well-defined, if $Q \in \mathcal{A}_{d,0} = \mathbb{C}\langle X_1, \ldots, X_d\rangle$,

$$\mathbb{E}\left[\frac{1}{N}\operatorname{Tr}_N\left(Q\left(X^N\right)\right)\right] - \tau\Big(Q\left(x\right)\Big) = -\int_0^\infty \mathbb{E}\left[\frac{d}{dt}\Big(\tau_N\left(Q(X_t^N)\right)\Big)\right]\ dt \ .$$

On the other hand, using the free Markov property of the free Brownian motion, we have for $Q \in \mathcal{A}_{d,0}$

$$\frac{d}{dt}\tau_N(Q(X_t^N)) = -\frac{1}{2}\sum_i \left\{\tau_N\left((X_t^N)_i(D_iQ)(X_t^N)\right) - \tau_N\otimes\tau_N\left((\partial_i D_i Q)(X_t^N)\right)\right\}.$$

One can already recognize the Schwinger-Dyson equation. Indeed thanks to Proposition 2.10, one can see that

$$\mathbb{E}\left[\frac{d}{dt}\tau_N(Q(X_t^N))\right]\Big|_{t=0} = -\frac{1}{2}\sum_i \mathbb{E}\left[\tau_N\left(X_i^N\ (D_iQ)(X^N)\right) - \tau_N\otimes\tau_N\left((\partial_i D_i Q)(X^N)\right)\right] = 0 \ .$$

And then, thanks to Proposition 2.6,

$$\mathbb{E}\left[\frac{d}{dt}\tau_N(Q(X_t^N))\right]\Big|_{t=\infty} = -\frac{1}{2}\sum_i \left\{\tau\left(x_i\ (D_iQ)(x)\right) - \tau\otimes\tau\left((\partial_i D_i Q)(x)\right)\right\} = 0 \ .$$

However what happens at time t is much harder to estimate and is the core of the proof. The main idea to deal with this issue is to view the family (X^N, x) as the asymptotic limit when k goes to infinity of the family $(X^N \otimes I_k, R^{kN})$ where R^{kN} are independent GUE matrices of size kN and independent of X^N.

Another issue is that to prove Theorem 1.1, we would like to set $Q = f(P)$ but since f is not polynomial this means that we need to extend the definition of operators such as ∂_i. In order to do so we assume that there exist μ a measure on $\mathbb{R}$ such that,

$$\forall x \in \mathbb{R}, \quad f(x) = \int_{\mathbb{R}} e^{\mathbf{i}xy}\ d\mu(y) \ .$$

While we have to assume that the support of μ is indeed on the real line, μ can be a complex measure. However we will usually work with measure such that $|\mu|(\mathbb{R})$ is finite. Indeed under this assumption we can use Fubini's Theorem, and we get

$$\mathbb{E}\left[\frac{1}{M}\operatorname{Tr}_N\left(f\left(P(X^N)\right)\right)\right] - \tau\Big(f\left(P(x)\right)\Big) = \int_{\mathbb{R}}\left\{\mathbb{E}\left[\frac{1}{N}\operatorname{Tr}_N\left(e^{\mathbf{i}yP(X^N)}\right)\right] - \tau\Big(e^{\mathbf{i}yP(x)}\Big)\right\}\ d\mu(y) \ .$$

We can then set $Q = e^{\mathbf{i}yP}$. And even though this is not a polynomial function, since it is a power series, most of the properties associated to polynomials remain true with some assumption on the convergence. The main difficulty with this method is that we need to find a bound which does not depend on too high moments of y. Indeed terms of the form

$$\int_{\mathbb{R}} |y|^l\ d|\mu|(y)$$

appear in our estimates. Thanks to Fourier integration we can relate the exponent l to the regularity of the function f, thus we want to find a bound with l as small as possible. It turns out that with our proof $l = 4$.

3.2 Proof of Theorem 1.1

In this section we focus on proving Theorem 1.1 from which we deduce all of the important corollaries. It will be a consequence of the following Theorem :

Theorem 3.1. *Let the following objects be given,*

- $X^N = (X_1^N, \dots, X_d^N)$ *independent GUE matrices of size* N,
- $x = (x_1, \dots, x_d)$ *a system of free semicircular variables,*
- $Z^{NM} = (Z_1^{NM}, \dots, Z_q^{NM})$ *deterministic matrices,*
- $P \in \mathcal{A}_{d,q}$ *a polynomial that we assume to be self-adjoint,*
- $f : \mathbb{R} \mapsto \mathbb{R}$ *such that there exists a measure on the real line* μ *with* $\int (1+y^4)\ d|\mu|(y) < +\infty$ *and for any* $x \in \mathbb{R}$,

$$f(x) = \int_{\mathbb{R}} e^{\mathbf{i}xy}\ d\mu(y)\ .$$

Then, there exists a polynomial L_P *which only depends on* P *such that for any* N, M,

$$\left| \mathbb{E}\left[\frac{1}{MN} \mathrm{Tr}_{MN}\left(f\left(P\left(X^N \otimes I_M, Z^{NM}, Z^{NM*} \right) \right) \right) \right] - \tau_N \otimes \tau_M \Big(f\left(P\left(x \otimes I_M, Z^{NM}, Z^{NM*} \right) \right) \Big) \right|$$
$$\leq \frac{M^2}{N^2} L_P\left(\left\| Z^{NM} \right\| \right) \int_{\mathbb{R}} (|y| + y^4)\ d|\mu|(y)\ .$$

The proof is a direct corollary of Lemmas 3.3 and 3.4 below. The first one shows that the crux of the proof lies in understanding the following quantity:

Definition 3.2. *Let the following objects be given,*

- $\alpha, \beta \in [0,1]$,
- $A, B, C, D \in \mathcal{A}_{d,q}$ *monomials,*
- $X_t^N = e^{-t/2} X^N + (1 - e^{-t})^{1/2} x$
- $Z_t^N = (X_t^N \otimes I_M, Z^{NM}, Z^{NM*})$,
- $S_t = (A e^{\mathbf{i}\beta y P} B)(Z_t^N)$,
- $V_t = (C e^{\mathbf{i}\alpha y P} D)(Z_t^N)$.

Then we define:

$$\mathcal{S}_{N,t}^{\alpha,\beta}\Big(A, B, C, D\Big) = \mathbb{E}\left[\frac{1}{N} \sum_{1 \leq s,r \leq N} \tau_N \otimes \tau_M \Big(E_{s,r} \otimes I_M \times S_t \times E_{r,s} \otimes I_M \times V_t \Big) \right]$$
$$- \mathbb{E}\left[\tau_M \Big((\tau_N \otimes I_M)(S_t)\ (\tau_N \otimes I_M)(V_t) \Big) \right].$$

We can now state the next lemma which explains why this object appears:

Lemma 3.3. *Let* f *be a function such that there exists a measure* μ *such that for any* $x \in \mathbb{R}$,

$$f(x) = \int_{\mathbb{R}} e^{ixy} d\mu(y)$$

We also assume that $\int_{\mathbb{R}} (1+y^4) d|\mu|(y) < \infty$. *Then one can write*

$$\mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}\left(f\left(P\left(X^N\otimes I_M, Z^{NM}, Z^{NM^*}\right)\right)\right)\right] - \tau_N\otimes\tau_M\Big(f\left(P\left(x\otimes I_M, Z^{NM}, Z^{NM^*}\right)\right)\Big)$$

as a finite linear combination of terms of the following kinds :

$$\int_0^\infty e^{-t}\ \int y^2\ \int_0^1 \mathcal{S}_{N,t}^{\alpha,1-\alpha}(A,B,C,D)d\alpha\ d\mu(y)\ dt\ , \tag{11}$$

and

$$\int_0^\infty e^{-t}\ \int y\ \mathcal{S}_{N,t}^{1,0}(A,B,C,D)d\mu(y)\ dt \tag{12}$$

where the monomials $A,B,C,D\in\mathcal{A}_{d,q}$ and the coefficients of the linear combination are uniquely determined by P.

Proof. First, we define the natural interpolation between the trace of matrices at size N and the trace of semicircular variables,

$$s(t,y) = \mathbb{E}\left[\tau_N\otimes\tau_M\left(e^{\mathbf{i}yP(Z_t^N)}\right)\right]\ .$$

By definition of f we have

$$\begin{aligned}\int_{\mathbb{R}} s(0,y)\ d\mu(y) &= \mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}_{MN}\left(f(P(X^N\otimes I_M, Z^{NM}, Z^{NM^*}))\right)\right],\\ \int_{\mathbb{R}} s(\infty,y)\ d\mu(y) &= \tau_N\otimes\tau_M\Big(f(P(x\otimes I_M, Z^{NM}, Z^{NM^*}))\Big).\end{aligned}$$

Thus under the assumption that this is well-defined, we have

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}_{MN}\left(f\left(P\left(X^N\otimes I_M, Z^{NM}, Z^{NM^*}\right)\right)\right)\right] - \tau_N\otimes\tau_M\Big(f\left(P\left(x\otimes I_M, Z^{NM}, Z^{NM^*}\right)\right)\Big)\\ &\quad= -\int_0^\infty\int_{\mathbb{R}}\partial_t s(t,y)\ d\mu(y)\ dt\ .\end{aligned} \tag{13}$$

We compute

$$\partial_t s(t,y) = \mathbf{i}y\frac{e^{-t}}{2}\ \mathbb{E}\left[\tau_N\otimes\tau_M\left(e^{\mathbf{i}yP(Z_t^N)}\ \sum_i \partial_i P(Z_t^N)\#\left(\left(\frac{x_i}{(1-e^{-t})^{1/2}} - e^{t/2}X_i^N\right)\otimes I_M\right)\right)\right].$$

Since we assumed that μ is such that $\int(1+y^4)d\mu(y)<+\infty$ and that since X_i^N and x_i have all moments uniformly bounded by Lemma 2.11, we can find a constant C independent from y and t such that

$$|\partial_t s(t,y)| \le C\ ye^{-t/2}\ ,$$

we can deduce that (13) is well-defined. Besides, writing $P=\sum c_Q(P)Q$ with monomials $Q\in\mathcal{A}_{d,q}$, we get

$$\partial_t s(t,y) = \mathbf{i}y\frac{e^{-t}}{2}\sum c_Q(P)\sum_{Q=BX_iA}\mathbb{E}\left[\tau_N\otimes\tau_M\left(e^{\mathbf{i}yP(Z_t^N)}\ B(Z_t^N)\left(\frac{x_i}{(1-e^{-t})^{1/2}} - e^{t/2}X_i^N\right)\otimes I_M A(Z_t^N)\right)\right]. \tag{14}$$

Hence, $\partial_t s$ is a finite linear combination of terms of the form

$$ye^{-t}S_t(A,B) = ye^{-t}S_t^1(A,B) - ye^{-t}S_t^2(A,B) \tag{15}$$

with

$$S_t^1(A,B) = S_t(A,B,(1-e^{-t})^{-1/2}x_i) \text{ and } S_t^2(A,B) = S_t(A,B,e^{t/2}X_i^N)$$

where

$$S_t(A,B,G) = \mathbb{E}\left[\tau_N \otimes \tau_M \left(A(Z_t^N)\ e^{\mathbf{i}yP(Z_t^N)}B(Z_t^N)G \otimes I_M\right)\right]. \tag{16}$$

We first study $S_t^2(A,B)$. We denote by $Q = Ae^{\mathbf{i}yP}B$. We want to use Gaussian integration by part: if we set $\sqrt{N}X_i^N = (x_{s,r}^i)_{1\le s,r\le N}$, then with $\partial_{x_{s,r}^i}$ as in equations (8) and (9), thanks to Duhamel formula

$$\begin{aligned}\sqrt{N}e^{t/2}\ \partial_{x_{s,r}^i}Q(Z_t^N) &= \partial_i A(Z_t^N)\#(E_{r,s}\otimes I_M)\ e^{\mathbf{i}yP(Z_t^N)}B(Z_t^N)\\ &+ \mathbf{i}y\int_0^1 A(Z_t^N)e^{\mathbf{i}(1-\alpha)yP(Z_t^N)}\ \partial_i P(Z_t^N)\#(E_{r,s}\otimes I_M)\ e^{\mathbf{i}\alpha yP(Z_t^N)}B(Z_t^N)\ d\alpha \\ &+ A(Z_t^N)e^{\mathbf{i}yP(Z_t^N)}\ \partial_i B(Z_t^N)\#(E_{r,s}\otimes I_M)\ .\end{aligned} \tag{17}$$

Consequently, expanding in $S_t^2(A,B)$ the product by X_i^N in terms of its entries, we have

$$\begin{aligned}S_t^2(A,B) =& e^{t/2}\mathbb{E}\left[\tau_N\otimes\tau_M\left((Ae^{\mathbf{i}yP}B)(Z_t^N)\ X_i^N\otimes I_M\right)\right]\\ =& N^{-1/2}e^{t/2}\sum_{1\le s,r\le N}\mathbb{E}\left[x_{s,r}^i\ \tau_N\otimes\tau_M\Big(E_{s,r}\otimes I_M\ (Ae^{\mathbf{i}yP}B)(Z_t^N)\Big)\right]\\ =& \frac{1}{N}\sum_{1\le s,r\le N}\mathbb{E}\left[\tau_N\otimes\tau_M\Big(E_{s,r}\otimes I_M\ e^{t/2}\partial_{x_{s,r}^i}Q(Z_t^N)\Big)\right]\\ =& \mathbb{E}\left[\frac{1}{N}\sum_{1\le s,r\le N}\tau_N\otimes\tau_M\Big(E_{s,r}\otimes I_M\ \partial_i A\#(E_{r,s}\otimes I_M)\ e^{\mathbf{i}yP}B\Big)\right]\\ &+\mathbf{i}y\int_0^1\mathbb{E}\left[\frac{1}{N}\sum_{1\le s,r\le N}\tau_N\otimes\tau_M\Big(E_{s,r}\otimes I_M\ Ae^{\mathbf{i}(1-\alpha)yP}\ \partial_i P\#(E_{r,s}\otimes I_M)\ e^{\mathbf{i}\alpha yP}B\Big)\right]d\alpha\\ &+\mathbb{E}\left[\frac{1}{N}\sum_{1\le s,r\le N}\tau_N\otimes\tau_M\Big(E_{s,r}\otimes I_M\ Ae^{\mathbf{i}yP}\ \partial_i B\#(E_{r,s}\otimes I_M)\Big)\right]\end{aligned} \tag{18}$$

where A, B, P are evaluated at Z_t^N. To deal with $S_t^1(A,B)$, since a priori we defined free integration by parts only for polynomials, we expand the exponential as a power series,

$$\begin{aligned}&\tau_N\otimes\tau_M\left(A(Z_t^N)\ e^{\mathbf{i}yP(Z_t^N)}\ B(Z_t^N)\frac{x_i\otimes I_M}{(1-e^{-t})^{1/2}}\right)\\ &=\sum_{k\ge 0}\frac{1}{k!}\ \tau_N\otimes\tau_M\left(A(Z_t^N)\ (\mathbf{i}yP(Z_t^N))^k\ B(Z_t^N)\frac{x_i\otimes I_M}{(1-e^{-t})^{1/2}}\right).\end{aligned}$$

We define $(\tau_N\otimes I_M)\bigotimes(\tau_N\otimes I_M) : (\mathcal{A}_N\otimes\mathbb{M}_M(\mathbb{C}))^{\otimes 2}\to M_M(\mathbb{C})$ the linear application which is defined on simple tensor by $(\tau_N\otimes I_M)\bigotimes(\tau_N\otimes I_M)(A\otimes B) = (\tau_N\otimes I_M)(A)\times(\tau_N\otimes I_M)(B)$. Hence, thanks to Proposition 2.6, with the convention that $A\times(B\otimes C)\times D = (AB)\otimes(CD)$, we have

$$
\begin{aligned}
&\tau_N \otimes \tau_M \left(A(Z_t^N)\ (\mathbf{i}yP(Z_t^N))^k\ B(Z_t^N) \frac{x_i \otimes I_M}{(1-e^{-t})^{1/2}} \right) \\
&= \tau_M \Big((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Big(\partial_i A(Z_t^N)\ (\mathbf{i}yP(Z_t^N))^k\ B(Z_t^N) \Big) \Big) \\
&\quad + \mathbf{i}y \tau_M \left((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Big(A(Z_t^N)\ (\mathbf{i}y)^{k-1} \sum_{1 \le l \le k} P(Z_t^N)^{l-1} \partial_i P(Z_t^N)\ P(Z_t^N)^{k-l}\ B(Z_t^N) \Big) \right) \\
&\quad + \tau_M \Big((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Big(A(Z_t^N)\ (\mathbf{i}yP(Z_t^N))^k\ \partial_i B(Z_t^N) \Big) \Big) .
\end{aligned}
$$

Now we can use the fact that

$$
\frac{1}{k!} = \int_0^1 \frac{\alpha^{l-1}(1-\alpha)^{k-l}}{(l-1)!(k-l)!} d\alpha \ ,
$$

to deduce that

$$
\begin{aligned}
&\tau_M \left((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Big(A(Z_t^N)\ \sum_{k \ge 1} \frac{(\mathbf{i}y)^{k-1}}{k!} \sum_{l=1}^{k} P(Z_t^N)^{l-1} \partial_i P(Z_t^N)\ P(Z_t^N)^{k-l} B(Z_t^N) \Big) \right) \\
&= \int_0^1 \sum_{k \ge 1} \sum_{l=1}^{k} \tau_M \Bigg((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Bigg(A(Z_t^N)\ \frac{(\mathbf{i}y\alpha P(Z_t^N))^{l-1}}{(l-1)!}\ \partial_i P(Z_t^N) \\
&\qquad\qquad \frac{(\mathbf{i}y(1-\alpha)P(Z_t^N))^{k-l}}{(k-l)!} B(Z_t^N) \Bigg) \Bigg) d\alpha \\
&= \int_0^1 \tau_M \Big((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Big(A(Z_t^N)\ e^{\mathbf{i}(1-\alpha)yP(Z_t^N)}\ \partial_i P(Z_t^N)\ e^{\mathbf{i}\alpha yP(Z_t^N)}\ B(Z_t^N) \Big) \Big) d\alpha \ .
\end{aligned}
$$

And thus, after summation, we obtain

$$
\begin{aligned}
S_t^1(A,B) = &\ \tau_M \Big((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Big(\partial_i A\ e^{\mathbf{i}yP} B \Big) \Big) \\
&+ \mathbf{i}y \int_0^1 \tau_M \Big((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Big(A e^{\mathbf{i}(1-\alpha)yP}\ \partial_i P\ e^{\mathbf{i}\alpha yP} B \Big) \Big) d\alpha \\
&+ \tau_M \Big((\tau_N \otimes I_M) \bigotimes (\tau_N \otimes I_M) \Big(A\ e^{\mathbf{i}yP}\ \partial_i B \Big) \Big) \ .
\end{aligned}
$$

Therefore, after making the difference (15) to compute $S_t(A,B)$, we conclude that the difference we wish to estimate in (13) is a linear combination of terms, whose coefficients only depend on P, of the form (11) and (12).

□

We need to study the quantity $\mathcal{S}_{N,t}^{\alpha,\beta}(A,B,C,D)$. Let us first explain why one can expect it to be small. Let $(g_i)_{1 \le i \le N}$ be the canonical basis of $\mathbb{C}^N$ so that $E_{r,s} = g_r g_s^*$. We observe that $\mathcal{S}_{N,0}^{\alpha,\beta}(A,B,C,D)$ vanishes since

$$\frac{1}{N}\sum_{1\le s,r\le N}\tau_N\otimes\tau_M\Big(E_{s,r}\otimes I_M\ S_0\ E_{r,s}\otimes I_M\ V_0\Big)$$
$$=\frac{1}{N^2}\sum_{1\le s,r\le N}\tau_M\Big(g_r^*\otimes I_M\ S_0\ g_r\otimes I_M\ g_s^*\otimes I_M\ V_0\ g_s\otimes I_M\ \Big)$$
$$=\tau_M\Big((\tau_N\otimes I_M)(S_0)\ (\tau_N\otimes I_M)(V_0)\Big)\ .$$

Let us now estimate $\mathcal{S}^{\alpha,\beta}_{N,\infty}(A,B,C,D)$. We first notice that if $X,Y\in\mathcal{A}_N$ are free from $\mathbb{M}_N(\mathbb{C})$ then (with constants being identified with constants times identity):

- If $r\neq s$: $\tau_N\Big(E_{r,s}(X-\tau_N(X))E_{s,r}(Y-\tau(Y))\Big)=0$.
- If $r=s$: $\tau_N\Big(\big(E_{r,r}-\frac{1}{N}\big)\,(X-\tau_N(X))\,\big(E_{r,r}-\frac{1}{N}\big)\,(Y-\tau_N(Y))\Big)=0.$

Consequently, since $\tau_N(E_{r,s}E_{s,r})=1/N$ for all r,s, we get:

- If $r\neq s$: $\tau_N\Big(E_{r,s}XE_{s,r}Y\Big)=\frac{1}{N}\tau_N(X)\tau_N(Y)$.
- If $r=s$: $\tau_N\Big(E_{r,s}XE_{s,r}Y\Big)=\frac{1}{N}\tau_N(X)\tau_N(Y)+\frac{1}{N^2}\Big(\tau_N(XY)-\tau_N(X)\tau_N(Y)\Big)$.

Hence
$$\frac{1}{N}\sum_{1\le s,r\le N}\tau_N\Big(E_{r,s}XE_{s,r}Y\Big)=\tau_N(X)\tau_N(Y)+\frac{1}{N^2}\Big(\tau_N(XY)-\tau_N(X)\tau_N(Y)\Big).$$

This implies that $N^2\mathcal{S}^{\alpha,\beta}_{N,\infty}(A,B,C,D)$ is bounded by a constant independent of N or y since

$$\frac{1}{N}\sum_{1\le s,r\le N}\tau_N\otimes\tau_M\Big(E_{s,r}\otimes I_M\ S_\infty\ E_{r,s}\otimes I_M\ V_\infty\Big)$$
$$=\tau_M\Big((\tau_N\otimes I_M)(S_\infty)\ (\tau_N\otimes I_M)(V_\infty)\Big)+\frac{1}{N^2}\Big(\tau_N\otimes\tau_M(S_\infty V_\infty)-\tau_M(\tau_N\otimes I_M(S_\infty)\tau_N\otimes I_M(V_\infty))\Big)\ .$$

With this in mind, we now study what happens at time t. More precisely we show:

Lemma 3.4. *There is a polynomial L which only depends on A,B,C,D and P such that for any $\alpha,\beta\in[0,1]$, $N\in\mathbb{N}$, $t\in\mathbb{R}^+$ and $y\in\mathbb{R}$,*

$$\Big|\mathcal{S}^{\alpha,\beta}_{N,t}\Big(A,B,C,D\Big)\Big|\le\frac{(1+y^2)M^2}{N^2}L(\|Z^{NM}\|)\ .$$

This lemma is a direct consequence of Lemmas 3.6 and 3.7. We first show that the family $(X^N\otimes I_M, x\otimes I_M, Z^{NM})$ is actually the asymptotic distribution (in the sense of Definition 2.1) as k goes to infinity of the family $(X^N\otimes I_{kM}, R^{kN}\otimes I_M, Z^{NM}\otimes I_k)$ where R^{kN} are independent GUE random matrices of size kN. The advantage of this representation is that it allows us to use classical analysis, and to treat the GUE variables and the semi-circle variables in a more symmetric way. A direct proof using semi-circular variables should however be possible.

Proposition 3.5. *If R^{kN} are independent GUE random matrices of size kN, independent of X^N, we set*

$$U_t^k=\Big(\Big(e^{-t/2}X^N\otimes I_k+(1-e^{-t})^{1/2}R^{kN}\Big)\otimes I_M\ ,Z^{NM}\otimes I_k,Z^{NM*}\otimes I_k\Big)\ .$$

Then if $q=Ae^{\mathbf{i}\beta yP}B$, we have that $\mathbb{P}_{X^N}$-almost surely for any t,

$$(\tau_N \otimes I_M)\big(q(Z_t^N)\big) = \lim_{k\to\infty} \mathbb{E}_R\left[(\tau_{kN} \otimes I_M)\big(q(U_t^k)\big)\right] \ ,$$

where $\mathbb{E}_R$ is the expectation with respect to R^{kN}. Here M, N are kept fixed.

Proof. This proposition is mostly a corollary of Theorem 5.4.5 of [2]. Indeed this theorem states that if R^{kN} are GUE matrices and D^{kN} are deterministic matrices such that

$$\sup_{l\in\mathbb{N}} \max_i \sup_{k\in\mathbb{N}} \left(\frac{1}{N}\operatorname{Tr}(|D_i^{kN}|^l)\right)^{1/l} < \infty \ ,$$

and if D^{kN} converges in distribution towards a family of non-commutative random variables d, then the family (R^{kN}, D^{kN}) in the non-commutative probability space $(\mathbb{M}_{kN}(\mathbb{C}), *, \mathbb{E}[\frac{1}{kN}\operatorname{Tr}])$ converges in distribution towards the family (x, d) where x is a system of free semicircular variables free from d. In our situation we can write for every i,

$$Z_i^{NM} = \sum_{1\le r,s\le N} E_{r,s} \otimes A_{r,s,i}^M \ .$$

Thus, if $E^N = (E_{r,s})_{1\le r,s\le N}$, we fix $D^{k,N} = (X^N \otimes I_k, E^N \otimes I_k)$, and we can apply Theorem 5.4.5 from [2] to get that for any non-commutative polynomial P,

$$\lim_{k\to\infty} \mathbb{E}_R\left[\tau_{kN}(P(R^{kN}, D^{k,N}))\right] = \tau_N\left(P(x, D^{k,1})\right).$$

Consequently, for any non-commutative polynomial P, we also have

$$\lim_{k\to\infty} \mathbb{E}_R\left[\tau_{kN} \otimes I_M(P(R^{kN}, D^{k,N}, A^M, A^{M*}))\right] = \tau_N \otimes I_M\left(P(x, X^N, E^N, A^M, A^{M*})\right).$$

Hence, for any $P \in \mathcal{A}_{d,q}$,

$$\lim_{k\to\infty} \mathbb{E}_R\left[\tau_{kN} \otimes I_M(P(U_t^k))\right] = \tau_N \otimes I_M\left(P(Z_t^N)\right). \tag{19}$$

Thanks to Property 2.11, we know that there exist $\alpha > 0$ and $D < \infty$ such that for all $u \ge D$, for N large enough,

$$\mathbb{P}\left(\left\|R_1^{kN}\right\| \ge u\right) \le e^{-\alpha\ u\ kN} \ . \tag{20}$$

Since if $c_M(P)$ is the coefficient of P associated with the monomial M, one has

$$\left\|P(U_t^k)\right\| \le \sum_{M \text{ monomials}} |c_M(P)| \left\|M(U_t^k)\right\| \ ,$$

there exist constants L and C which do depend on $\left\|Z_j^{NM}\right\|$ and $\left\|X_i^N\right\|$ such that for N large enough

$$\mathbb{P}\left(\left\|P(U_t^k)\right\| \ge C\right) \le e^{-LkN} \ . \tag{21}$$

Knowing this, let $f_\varepsilon \in \mathbb{C}[X]$ be a polynomial which is ε-close from $x \mapsto e^{\mathbf{i}\beta y x}$ on the interval $[-1-C, C+1]$. Since one can always assume that $C > \left\|P(Z_t^N)\right\|$, we have, with $q = Ae^{\mathbf{i}\beta y P}B$:

$$\|(\tau_N \otimes I_M)\big(q(Z_t^N)\big) - (\tau_N \otimes I_M)\big((Af_\varepsilon(P)B)(Z_t^N)\big)\| \le D\varepsilon \ ,$$

where D is some constant which can depend on the dimensions N, M but not on k.
Thus

$$\begin{aligned}\|(\tau_N \otimes I_M)\big(q(Z_t^N)\big) - \mathbb{E}_R\left[(\tau_{kN} \otimes I_M)\big(q(U_t^k)\big)\right]\| \le D\varepsilon + D\mathbb{E}_R\left[\left\|(q - Af_\varepsilon(P)B)(U_t^k)\right\| \mathbf{1}_{\|P(U_t^k)\|\ge C+1}\right] \\ + \|(\tau_N \otimes I_M)\big((Af_\varepsilon(P)B)(Z_t^N)\big) - \mathbb{E}_R\left[(\tau_{kN} \otimes I_M)\big((Af_\varepsilon(P)B)(U_t^k)\big)\right]\|\end{aligned}$$

The last term goes to zero as k goes to infinity by (19). Besides

$$\mathbb{E}_R\left[\left\|(q-Af_\varepsilon(P)B)(U_t^k)\right\|\mathbf{1}_{\|P(U_t^k)\|\geq C+1}\right]$$
$$\leq \mathbb{E}_R\left[\left(\left\|A(U_t^k)\right\|\left\|B(U_t^k)\right\|+\left\|(Af_\varepsilon(P)B)(U_t^k)\right\|\right)^2\right]^{1/2}\mathbb{P}(\left\|P(U_t^k)\right\|\geq C+1)^{1/2}\ .$$

The first term is bounded independently of k thanks to (20) and the second converges exponentially fast towards 0 thanks to (21). Consequently

$$\limsup_{k\to\infty}\|(\tau_N\otimes I_M)\big(q(Z_t^N)\big)-\mathbb{E}_R\left[(\tau_{kN}\otimes I_M)\big(q(U_t^k)\big)\right]\|\leq D\varepsilon\ .$$

Hence the conclusion follows since the left hand side does not depend on ε.

□

Recall that by definition

$$\mathcal{S}_{N,t}^{\alpha,\beta}\Big(A,B,C,D\Big):=\mathbb{E}[\Lambda_{N,t}^{\alpha,\beta}\Big(A,B,C,D\Big)] \tag{22}$$

with, following the notations of Definition 3.2 :

$$\Lambda_{N,t}^{\alpha,\beta}\Big(A,B,C,D\Big)=\frac{1}{N}\sum_{1\leq s,r\leq N}\tau_N\otimes\tau_M\Big(E_{s,r}\otimes I_M\times S_t\times E_{r,s}\otimes I_M\times V_t\Big)$$
$$-\tau_M\Big((\tau_N\otimes I_M)(S_t)\ (\tau_N\otimes I_M)(V_t)\Big).$$

By Proposition 3.5, we deduce that

$$\Lambda_{N,t}^{\alpha,\beta}\Big(A,B,C,D\Big)=\lim_{k\to\infty}\Lambda_{k,N,t}^{\alpha,\beta}\Big(A,B,C,D\Big) \tag{23}$$

where $\Lambda_{k,N,t}^{\alpha,\beta}\Big(A,B,C,D\Big)$ equals

$$\mathbb{E}_R\Big[\frac{1}{N}\sum_{1\leq s,r\leq N}\tau_{kN}\otimes\tau_M\Big(E_{s,r}\otimes I_k\otimes I_M\ (Ae^{\mathbf{i}\beta yP}B)(U_t^k)\ E_{r,s}\otimes I_k\otimes I_M\ (Ce^{\mathbf{i}\alpha yP}D)(U_t^k)\Big)\Big]$$
$$-\tau_M\big(\mathbb{E}_R\left[\tau_{kN}\otimes I_M(Ae^{\mathbf{i}\beta yP}B)(U_t^k)\right]\mathbb{E}_R\left[\tau_{kN}\otimes I_M(Ce^{\mathbf{i}\alpha yP}D)(U_t^k))\right]\big) \tag{24}$$

We can now prove the following intermediary lemma in view of deriving Lemma 3.4.

Lemma 3.6. *Define U_t^k as in Proposition 3.5, and let*

- $P_{1,2}=I_N\otimes E_{1,2}\otimes I_M$,
- $Q=(Ae^{\mathbf{i}\beta yP}B)(U_t^k)$,
- $T=(Ce^{\mathbf{i}\alpha yP}D)(U_t^k)$.

Then there is a constant C and a polynomial L which only depend on A,B,C,D and P such that for any $\alpha,\beta\in[0,1]$, $M,N\in\mathbb{N}$, $t\in\mathbb{R}^+$ and $y\in\mathbb{R}$,

$$|\Lambda_{k,N,t}^{\alpha,\beta}\Big(A,B,C,D\Big)|\leq\frac{(1+y^2)M^2}{N^2}L\left(\left\|Z^{NM}\right\|,\left\|X^N\right\|\right) \tag{25}$$
$$+k^3\left|\tau_M\left(\mathbb{E}_R\left[(\tau_{kN}\otimes I_M)(QP_{1,2})\right]\mathbb{E}_R\left[(\tau_{kN}\otimes I_M)(TP_{1,2})\right]\right)\right|.$$

Proof. We denote in short $\Lambda_{k,N,t}^{\alpha,\beta}\Big(A,B,C,D\Big)=\Lambda_{k,N,M}=\mathbb{E}_R[\Lambda_{k,N,M}^1]-\Lambda_{k,N,M}^2$ with

$$\Lambda_{k,N,M}^1 = \frac{1}{N}\sum_{1\leq s,r\leq N}\tau_{kN}\otimes I_M\left(E_{s,r}\otimes I_k\otimes I_M\ Q\ E_{r,s}\otimes I_k\otimes I_M\ T\right)$$
$$\Lambda_{k,N,M}^2 = \tau_M\big(\mathbb{E}_R\left[\tau_{kN}\otimes I_MQ\right]\mathbb{E}_R\left[\tau_{kN}\otimes I_M(T)\right]\big) \tag{26}$$

Let $(g_i)_{i\in[1,N]}$ and $(f_i)_{i\in[1,k]}$ be the canonical basis of $\mathbb{C}^N$ and $\mathbb{C}^k$, $E_{i,j}$ is the matrix whose only non-zero coefficient is (i,j) and this coefficient has value 1, the size of the matrix $E_{i,j}$ will depend on the context. We use the fact that $E_{r,s} = g_r g_s^*$ and $I_k = \sum_l E_{l,l}$ with $E_{l,l} = f_l^* f_l$ to deduce that

$$\begin{aligned}
\Lambda^1_{k,N,M} &= \frac{1}{N} \sum_{1\le s,r\le N} \sum_{1\le l,l'\le k} \tau_{kN}\otimes I_M \left(E_{s,r}\otimes E_{l,l}\otimes I_M \; Q \; E_{r,s}\otimes E_{l',l'}\otimes I_M \; T\right) \\
&= \frac{1}{N^2 k} \sum_{1\le l,l'\le k} \sum_{1\le r\le N} g_r^*\otimes f_l^*\otimes I_M \; Q \; g_r\otimes f_{l'}\otimes I_M \sum_{1\le s\le N} g_s^*\otimes f_{l'}^*\otimes I_M \; T \; g_s\otimes f_l\otimes I_M \\
&= \frac{1}{k} \sum_{1\le l,l'\le k} (\tau_N\otimes I_M)\left(I_N\otimes f_l^*\otimes I_M \; Q \; I_N\otimes f_{l'}\otimes I_M\right) \; (\tau_N\otimes I_M)\left(I_N\otimes f_{l'}^*\otimes I_M \; T \; I_N\otimes f_l\otimes I_M\right) \\
&= k \sum_{1\le l,l'\le k} (\tau_{kN}\otimes I_M)\big(Q \; I_N\otimes E_{l',l}\otimes I_M\big) \; (\tau_{kN}\otimes I_M)\big(T \; I_N\otimes E_{l,l'}\otimes I_M\big). \qquad (27)
\end{aligned}$$

The last line of the above equation prompts us to set $P_{l',l} = I_N\otimes E_{l',l}\otimes I_M$. If $(e_i)_{i\in[1,M]}$ is the canonical basis of $\mathbb{C}^M$, we set

$$F^q_{l,l',u,v}(R^{kN}) = e_u^* \; (\tau_{kN}\otimes I_M)\Big(q\big((e^{-t/2}X^N\otimes I_k + (\frac{1-e^{-t}}{Nk})^{1/2}R^{kN})\otimes I_M, Z^{NM}, Z^{NM*}\big) \; P_{l',l}\Big) \; e_v$$

with $q = Q = Ae^{\mathbf{i}\beta y_P}B$ or $q = T = Ce^{\mathbf{i}\alpha y_P}D$. We thus have with (27)

$$\begin{aligned}
\mathbb{E}_R\left[\Lambda^1_{k,N,M}\right] &= k \sum_{1\le l,l'\le k} \tau_M\left(\mathbb{E}_R\left[(\tau_{kN}\otimes I_M)\big(Q \; P_{l',l}\big) \; (\tau_{kN}\otimes I_M)\big(T \; P_{l,l'}\big)\right]\right) \qquad (28)\\
&= \frac{k}{M} \sum_{\substack{1\le l,l'\le k\\ 1\le u,v\le M}} \mathrm{Cov}_R\left(F^Q_{l,l',u,v}(R^{kN}), F^T_{l',l,u,v}(R^{kN})\right) \\
&\quad + k \sum_{1\le l,l'\le k} \tau_M\left(\mathbb{E}_R\left[(\tau_{kN}\otimes I_M)\big(Q \; P_{l',l}\big)\right] \; \mathbb{E}_R\left[(\tau_{kN}\otimes I_M)\big(T \; P_{l,l'}\big)\right]\right) \; .
\end{aligned}$$

However, the law of U_t^k is invariant under conjugation by $I_N\otimes U\otimes I_M$, where $U\in M_k(\mathbb{C})$ is a permutation matrix. Therefore, if $l = l'$, $\mathbb{E}_R[\tau_{kN}(Q \; P_{l',l})] = \mathbb{E}_R[\tau_{kN}(Q \; P_{1,1})]$, and if $l \neq l'$, $\mathbb{E}_R[\tau_{kN}(Q \; P_{l',l})] = \mathbb{E}_R[\tau_{kN}(Q \; P_{1,2})]$. We get the same equation when replacing Q by T. Consequently, we get

$$\begin{aligned}
&k \sum_{1\le l,l'\le k} \mathbb{E}_R\Big[(\tau_{kN}\otimes I_M)\big(Q \; P_{l',l}\big)\Big]\mathbb{E}_R\Big[(\tau_{kN}\otimes I_M)\big(T \; P_{l,l'}\big)\Big] \\
&\quad = k^2 \; \mathbb{E}_R[(\tau_{kN}\otimes I_M)(QP_{1,1})] \; \mathbb{E}_R[(\tau_{kN}\otimes I_M)(TP_{1,1})] \\
&\quad + (k-1)k^2 \; \mathbb{E}_R[(\tau_{kN}\otimes I_M)(QP_{1,2})] \; \mathbb{E}_R[(\tau_{kN}\otimes I_M)(TP_{1,2})] \; .
\end{aligned}$$

where the first term in the right hand side equals $\Lambda^2_{k,N,M} = \mathbb{E}_R[(\tau_{kN}\otimes I_M)(Q)] \; \mathbb{E}_R[(\tau_{kN}\otimes I_M)(T)]$ because $I_M = \sum_l P_{l,l}$. Thus equation (28) yields

$$\begin{aligned}
|\Lambda_{k,N,M}| &\le \frac{k}{M} \sum_{\substack{1\le l,l'\le k\\ 1\le u,v\le M}} \left|\mathrm{Cov}_R\left(F^Q_{l,l',u,v}(R^{kN}), F^T_{l',l,u,v}(R^{kN})\right)\right| \qquad (29)\\
&\quad + \left|k^3\tau_M\Big(\mathbb{E}_R[(\tau_{kN}\otimes I_M)(QP_{1,2})] \; \mathbb{E}_R[(\tau_{kN}\otimes I_M)(TP_{1,2})]\Big)\right| \; .
\end{aligned}$$

Hence, we only need to bound the first term to complete the proof of the lemma. Thanks to Cauchy-Schwartz's inequality, it is enough to bound the covariance of $F^q_{l,l',u,v}(R^{kN})$, for $q = Q$ and T. To study these variances, we shall use the Poincaré inequality, see Proposition 2.9. If we set $x^i_{r,s}$ and $y^i_{r,s}$ the real and imaginary part of $\sqrt{2kN}(R_i^{kN})_{r,s}$ for $r < s$ and $x^i_{r,r} = \sqrt{kN}(R_i^{kN})_{r,r}$, then these are real

centered Gaussian random variables of variance 1 and one can view $F^q_{l,l',u,v}$ as a function on $(x^i_{r,s})_{r\leq s,i}$ and $(y^i_{r,s})_{r<s,i}$. By a computation similar to (17), we find

$$\frac{kN}{1-e^{-t}}\left\|\nabla F^q_{l,l',u,v}\right\|_2^2=\sum_i\sum_{1\leq r,s\leq kN}e_u^*\,(\tau_{kN}\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ P_{l',l}\Big)e_v e_v^*\,(\tau_{kN}\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ P_{l',l}\Big)^* e_u.$$

It is worth noting that here the matrices $E_{r,s}$ have size kN in this formula. Thanks to Poincaré inequality (see Proposition 2.9), we deduce

$$\begin{aligned}
&\frac{k}{M}\sum_{1\leq u,v\leq M}\mathrm{Var}_R(F^q_{l,l',u,v}(R_{kN}))\leq\frac{k}{M}\sum_{1\leq u,v\leq M}\mathbb{E}\left[\left\|\nabla F^q_{l,l',u,v}\right\|_2^2\right]\\
&\leq\frac{1}{N}\sum_i\sum_{1\leq r,s\leq kN}\mathbb{E}_R\Bigg[\frac{1}{M}\sum_{1\leq u,v\leq M}e_u^*\,(\tau_{kN}\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ P_{l',l}\Big)e_v e_v^*\\
&\qquad\qquad\times(\tau_{kN}\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ P_{l',l}\Big)^* e_u\Bigg]\\
&\leq\frac{1}{N}\sum_i\sum_{1\leq r,s\leq kN}\mathbb{E}_R\left[\tau_M\left((\tau_{kN}\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ P_{l',l}\Big)(\tau_{kN}\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ P_{l',l}\Big)^*\right)\right]\ .
\end{aligned}\tag{30}$$

Moreover we have, if e_l is an orthornormal basis of $\mathbb{C}^k$,

$$\begin{aligned}
&\sum_{1\leq l,l'\leq k}\tau_M\left((\tau_{kN}\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ P_{l',l}\Big)(\tau_{kN}\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ P_{l',l}\Big)^*\right)\\
&=\frac{1}{k^2}\sum_{1\leq l,l'\leq k}\tau_M\Bigg(e_l^*\otimes I_M\ (\tau_N\otimes I_k\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ \Big)\ e_{l'}e_{l'}^*\otimes I_M\\
&\qquad\qquad(\tau_N\otimes I_k\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\Big)^* e_l\otimes I_M\ \Bigg)\\
&=\frac{1}{k}\tau_k\otimes\tau_M\left((\tau_N\otimes I_k\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\ \Big)\ (\tau_N\otimes I_k\otimes I_M)\Big(\partial_i q\#E_{r,s}\otimes I_M\Big)^*\right)\ .
\end{aligned}\tag{31}$$

Hence by combining equations (30) and (31) we have proved that

$$\begin{aligned}
&\frac{k}{M}\sum_{\substack{1\leq l,l'\leq k\\1\leq u,v\leq M}}\mathrm{Var}_R\left(F^q_{l,l',v,u}(R^{kN})\right)\\
&\leq\frac{1}{kN}\sum_i\sum_{1\leq r,s\leq kN}\mathbb{E}_R\left[\tau_k\otimes\tau_M\left((\tau_N\otimes I_{kM})\Big(\partial_i q\#E_{r,s}\otimes I_M\ \Big)\ (\tau_N\otimes I_{kM})\Big(\partial_i q\#E_{r,s}\otimes I_M\Big)^*\right)\right]
\end{aligned}\tag{32}$$

Moreover, let us remind that, with the convention $A\times(B\otimes C)\times D=(AB)\otimes(CD)$, we have (for $q=Q=Ae^{\mathbf{i}\beta yP}B$ but with obvious changes for $q=T$)

$$\partial_i q=\partial_i A\ e^{\mathbf{i}\beta yP}\ B+\mathbf{i}\beta yA\int_0^1 e^{\mathbf{i}(1-u)\beta yP}\ \partial_i P\ e^{\mathbf{i}u\beta yP}\ Bdu+A\ e^{\mathbf{i}\beta yP}\ \partial_i B\ .$$

Consequently, (32) is a finite linear combination of terms of the three following types $Q^i_N=\mathbb{E}_R[q^i_N]$, $1\leq i\leq 3$, with

$$\begin{aligned}
q^1_N=\frac{1}{kN}\sum_{1\leq r,s\leq kN}\tau_k\otimes\tau_M\Big((\tau_N\otimes I_{kM})\Big(A_1E_{r,s}\otimes I_M\ A_2e^{\mathbf{i}\beta yP}\ A_3\Big)\\
(\tau_N\otimes I_{kM})\Big(B_3E_{s,r}\otimes I_M\ B_2e^{-\mathbf{i}\beta yP}\ B_1\Big)\Big)\ ,
\end{aligned}$$

$$q_N^2 = \frac{\beta y}{kN}\int_0^1 \sum_{1\le r,s\le kN} \tau_k\otimes\tau_M\Big((\tau_N\otimes I_{kM})\Big(A_1 e^{\mathbf{i}(1-u)\beta y P}\ A_2E_{r,s}\otimes I_M\ A_3 e^{\mathbf{i}u\beta y P}\ A_4\Big)$$
$$(\tau_N\otimes I_{kM})\Big(B_3\ E_{s,r}\otimes I_M\ B_2 e^{-\mathbf{i}\beta y P}\ B_1\Big)\Big)du\ ,$$

$$q_N^3 = \frac{(\beta y)^2}{kN}\int_0^1\int_0^1 \sum_{1\le r,s\le kN} \tau_k\otimes\tau_M\Big((\tau_N\otimes I_{kM})\Big(A_1 e^{\mathbf{i}(1-u)\beta y P}\ A_2E_{r,s}\otimes I_M\ A_3 e^{\mathbf{i}u\beta y P}\ A_4\Big) \tag{33}$$
$$(\tau_N\otimes I_{kM})\Big(B_4 e^{-\mathbf{i}v\beta y P}\ B_3E_{s,r}\otimes I_M\ B_2 e^{-\mathbf{i}(1-v)\beta y P}\ B_1\Big)\Big)du\ dv,$$

where the A_i and B_i are monomials in U_t^k. Besides the coefficients of this linear combination only depend on A, B and P.
We first show how to estimate q_N^3. Let us recall that we set $(e_i)_{1\le i\le N}$, $(f_i)_{1\le i\le k}$ and $(g_i)_{1\le i\le M}$ as the canonical basis of $\mathbb{C}^M$, $\mathbb{C}^k$ and $\mathbb{C}^N$. Then, for any matrices $A, B, C, D \in \mathbb{M}_N(\mathbb{C})\otimes\mathbb{M}_k(\mathbb{C})\otimes\mathbb{M}_M(\mathbb{C})$, we have

$$\sum_{1\le r,s\le kN} \mathrm{Tr}_{kM}\left(\mathrm{Tr}_N\otimes I_{kM}\Big(A\ E_{r,s}\otimes I_M\ B\Big)\times \mathrm{Tr}_N\otimes I_{kM}\Big(C\ E_{s,r}\otimes I_M\ D\Big)\right) \tag{34}$$
$$= \sum_{1\le a,b,r_1,s_1\le N}\ \sum_{1\le c,d,r_2,s_2\le k}\ \sum_{1\le e,f,g,h\le M} g_a^*\otimes f_c^*\otimes e_e^*\ A\ g_{r_1}\otimes f_{r_2}\otimes e_f\times g_{s_1}^*\otimes f_{s_2}^*\otimes e_f^*\ B\ g_a\otimes f_d\otimes e_g$$
$$\times\, g_b^*\otimes f_d^*\otimes e_g^*\ C\ g_{s_1}\otimes f_{s_2}\otimes e_h\times g_{r_1}^*\otimes f_{r_2}^*\otimes e_h^*\ D\ g_b\otimes f_c\otimes e_e$$

$$= \sum_{\substack{1\le a\le N\\ 1\le c,d\le k\\ 1\le e,f,g,h\le M}} g_a^*\otimes f_c^*\otimes e_e^*\ A\ I_N\otimes I_k\otimes(e_fe_h^*)\ D\ I_N\otimes(f_cf_d^*)\otimes(e_ee_g^*)\ C\ I_N\otimes I_k\otimes(e_he_f^*)\ B\ g_a\otimes f_d\otimes e_g$$
$$= \sum_{1\le u,v\le M} \mathrm{Tr}_N\Big(I_N\otimes\mathrm{Tr}_{kM}(A\ I_{kN}\otimes e_ue_v^*\ D)\ I_N\otimes\mathrm{Tr}_{kM}(C\ I_{kN}\otimes e_ve_u^*\ B)\Big)\ .$$

Let K_M be a GUE matrix of size M, independent of everything else. Performing a Gaussian integration by part, we get

$$\frac{1}{M}\sum_{1\le u,v\le M} \mathrm{Tr}_N\Big(I_N\otimes\mathrm{Tr}_{kM}(A\ I_{kN}\otimes e_ue_v^*\ D)\ I_N\otimes\mathrm{Tr}_{kM}(C\ I_{kN}\otimes e_ve_u^*\ B)\Big) \tag{35}$$
$$= \mathbb{E}_K\left[\mathrm{Tr}_N\left(I_N\otimes\mathrm{Tr}_{kM}\Big(A\ I_{kN}\otimes K_M\ D\Big)\ I_N\otimes\mathrm{Tr}_{kM}\Big(C\ I_{kN}\otimes K_M\ B\Big)\right)\right].$$

Consequently by combining equations (34) and (35), we have

$$q_N^3 = \left(\frac{\beta yM}{N}\right)^2\int_0^1\int_0^1 \mathbb{E}_K\Big[\tau_N\Big((I_N\otimes\tau_{kM})\Big(A_1e^{\mathbf{i}(1-u)\beta y P}A_2\ I_{kN}\otimes K_M\ B_2e^{-\mathbf{i}(1-v)\beta y P}B_1\Big)$$
$$\times\,(I_N\otimes\tau_{kM})\Big(B_4e^{-\mathbf{i}v\beta y P}B_3\ I_{kN}\otimes K_M\ A_3e^{\mathbf{i}u\beta y P}A_4\Big)\Big)\Big]du\ dv\ .$$

Since P is self-adjoint, we know that for any real r, $\left\|e^{\mathbf{i}\ rP(U_t^k)}\right\| = 1$. Besides $\|I_N\otimes\tau_{kM}(A)\|\le\|A\|$, thus we can bound q_N^3 in (33) by

$$|q_N^3| \leq \left(\frac{yM}{N}\right)^2 \|A_1\| \|A_2\| \|A_3\| \|A_4\| \|B_1\| \|B_2\| \|B_3\| \|B_4\| \ \mathbb{E}_K\left[\|K_M\|^2\right]. \tag{36}$$

Finally, by [13], $\mathbb{E}_K\left[\|K_M\|^2\right]$ is bounded by 3. One can bound similarly q_N^1 and q_N^2, the only difference on the final result is that we would have 1 or y instead of y^2. Finally after taking the expectation with respect to R^{kN} in equation (36) and using Proposition 2.11, we deduce that there exists S which only depends on A, B and P, hence is independent of N, M, y, t, α or β, such that the covariance in (32) is bounded by

$$\frac{k}{M} \sum_{\substack{1\leq l,l'\leq k \\ 1\leq u,v\leq M}} \mathrm{Var}_R\left(F^q_{l,l',v,u}(R^{kN})\right) \leq \frac{(1+y^2)M^2}{N^2} S\left(\|X^N\|, \|Z^{NM}\|\right).$$

Thus, we deduce that there exists a polynomial H which only depends on A, B, C, D and P such that the first term in the right hand side of (29) is bounded by

$$\frac{k}{M} \sum_{\substack{1\leq l,l'\leq k \\ 1\leq u,v\leq M}} \left|\mathrm{Cov}_R\left(F^Q_{l,l',u,v}(R^{kN}), F^T_{l',l,u,v}(R^{kN})\right)\right| \leq \frac{(1+y^2)M^2}{N^2} H\left(\|X^N\|, \|Z^{NM}\|\right). \tag{37}$$

This completes the proof of the Lemma in the general case. For the specific case where $Z^{NM} = (I_N \otimes Y_1^M, \ldots, I_N \otimes Y_q^M)$ and that these matrices commute, we can get better estimate in equation (36) thanks to a refinement of equation (35). Indeed if A, B, C, D are monomials in U_t^k, then we can write $A = A_1 \otimes A_2$ in $\mathbb{M}_{kN}(\mathbb{C}) \otimes \mathbb{M}_M(\mathbb{C})$ and likewise for B, C, D such that A_2, B_2, C_2, D_2 commute. Thus,

$$\begin{aligned}
&\frac{1}{M} \sum_{1\leq u,v\leq M} \mathrm{Tr}_N\left(I_N \otimes \mathrm{Tr}_{kM}(A\ I_{kN} \otimes e_u e_v^*\ D)\ I_N \otimes \mathrm{Tr}_{kM}(C\ I_{kN} \otimes e_v e_u^*\ B)\right) \\
&= \frac{1}{M} \mathrm{Tr}_N\left(I_N \otimes \mathrm{Tr}_k(A_1 D_1)\ I_N \otimes \mathrm{Tr}_k(C_1 B_1)\right) \sum_{1\leq u,v\leq M} \mathrm{Tr}_M(A_2\ e_u e_v^*\ D_2)\, \mathrm{Tr}_M(C_2\ e_u e_v^*\ B_2) \\
&= \frac{1}{M} \mathrm{Tr}_N\left(I_N \otimes \mathrm{Tr}_k(A_1 D_1)\ I_N \otimes \mathrm{Tr}_k(C_1 B_1)\right) \mathrm{Tr}_M(D_2 A_2 B_2 C_2) \\
&= \frac{1}{M} \mathrm{Tr}_N\left(I_N \otimes \mathrm{Tr}_k(A_1 D_1)\ I_N \otimes \mathrm{Tr}_k(C_1 B_1)\right) \mathrm{Tr}_M(A_2 D_2 C_2 B_2) \\
&= \frac{1}{M} \mathrm{Tr}_{NM}\left(I_{NM} \otimes \mathrm{Tr}_k(AD)\ I_{NM} \otimes \mathrm{Tr}_k(CB)\right).
\end{aligned}$$

By linearity and density this equality is true if we assume that A, B, C, D are power series in U_t^k. Thus combining this equality with equation (34), we get that in this case

$$|q_N^3| \leq \left(\frac{y}{N}\right)^2 \|A_1\| \|A_2\| \|A_3\| \|A_4\| \|B_1\| \|B_2\| \|B_3\| \|B_4\| \ .$$

The same argument as in the general case applies and the proof follows.

□

In order to prove Lemma 3.4, we show in the following lemma that the term appearing in the second line of equation (25) vanishes.

Lemma 3.7. *Let $U_t^k, P_{1,2}, Q$ and T be defined as in Lemma 3.6, then $\mathbb{P}_{X^N}$-almost surely,*

$$\lim_{k\to\infty} k^3 \tau_M\left(\mathbb{E}_R\left[(\tau_{kN} \otimes I_M)(QP_{1,2})\right] \mathbb{E}_R\left[(\tau_{kN} \otimes I_M)(TP_{1,2})\right]\right) = 0 \ .$$

Proof. It is enough to show that given $y \in \mathbb{R}$ and monomial A and B, we have

$$\lim_{k\to\infty} k^{3/2} \mathbb{E}_R\left[(\tau_{kN} \otimes I_M)((A\ e^{\mathbf{i}yP}\ B)(U_t^k)\ P_{1,2})\right] = 0 \ .$$

For this purpose, let us define for monomials A, B and $y \geq 0$

$$f_{A,B}(y) = \mathbb{E}_R\left[(\mathrm{Tr}_{kN} \otimes I_M)((A\ e^{\mathbf{i}yP}\ B)(U_t^k)\ P_{1,2})\right]\ .$$

We want to show that $f_{A,B}$ goes to zero faster than $k^{-1/2}$. We first show that we can reduce the problem to the case $y = 0$. To this end, we also define

$$d_n(y) = \sup_{\deg(A)+\deg(B)\leq n} \|f_{A,B}(y)\|\ .$$

We know thanks to Proposition 2.11 that there exist constants α and C such that for any i and $n \leq \alpha kN/2$,

$$\mathbb{E}\left[\left\|R_i^{kN}\right\|^n\right] \leq C^n\ .$$

Consequently, P_{X^N}-almost surely, there exist constants γ and D (which do depend on, N, $\left\|X^N\right\|$ and $\left\|Z^{NM}\right\|$) such that for any $n \leq \gamma k$,

$$d_n(y) \leq D^n\ . \tag{38}$$

It is important to point out that this constant D can be very large when N is, it does not matter since, in the end, we will show that this quantity will go towards 0 when k goes to infinity and the other parameters such as N, M or y are fixed. Next, we define

$$g_{k,a}(y) = \sum_{0\leq n\leq \gamma k} d_n(y)a^n\ .$$

But we have

$$\frac{df_{A,B}(y)}{dy} = \mathbf{i}\mathbb{E}_R\left[(\mathrm{Tr}_{kN} \otimes I_M)((A\ Pe^{\mathbf{i}yP}\ B)(U_t^k)\ P_{1,2})\right]$$

so that if we set $c_L(P)$ to be the coefficient associated to the monomial L in P, $P = \sum c_L(P)L$,

$$\left|\frac{df_{A,B}(y)}{dy}\right| \leq \sum_{L \text{ monomials}} |c_L(P)|\ d_{\deg(A)+\deg(B)+\deg(L)}(y)\ .$$

Thus, for any $y \geq 0$, any monomials A, B with $\deg(A) + \deg(B) = n$,

$$f_{A,B}(y) \leq f_{A,B}(0) + \sum_{L \text{ monomials}} |c_L(P)|\ \int_0^y d_{n+\deg(L)}(u)\ du\ .$$

Therefore, we have for $y \geq 0$ and any $n \geq 0$,

$$a^n d_n(y) \leq a^n d_n(0) + \sum_{L \text{ monomials}} |c_L(P)|a^{-\deg(L)}\ \int_0^y d_{n+\deg(L)}(u)a^{n+\deg(L)}du\ .$$

And with $\|.\|_{a^{-1}}$ defined as in (7), thanks to (38), we find a finite constant c_a independent of k such that

$$g_{k,a}(y) \leq g_{k,a}(0) + c_a(aD)^{\gamma k} + \|P\|_{a^{-1}} \int_0^y g_{k,a}(u)du\ ,$$

where we used (38). As a consequence of Gronwall's inequality, we deduce that for $y \geq 0$,

$$g_{k,a}(y) \leq \left(g_{k,a}(0) + c_a(aD)^{\gamma k}\right) e^{y\|P\|_{a^{-1}}}\ . \tag{39}$$

Hence, it is enough to find an estimate on $g_{k,a}(0)$. First for any j, one can write $Z_j^{NM} = \sum_{1\leq u,v\leq N} E_{u,v} \otimes I_k \otimes A_{u,v}^j$ for some matrices $A_{u,v}^j$, then we define

$$U_{N,k} = \left(R^{kN}, X^N \otimes I_k, (E_{u,v} \otimes I_k)_{u,v}\right)\ , \quad c_n = \sup_{\deg(L)\leq n,\ L \text{ monomial}} \left|\mathbb{E}_R\left[\mathrm{Tr}_{kN}(L(U_{N,k})\ P_{1,2})\right]\right|\ .$$

Note that since we are taking the trace of $L(U_{N,k})P_{1,2}$ with $P_{1,2} = I_N \otimes f_1 f_2^* \otimes I_M$, we have $c_0 = c_1 = 0$. We consider K the supremum over u, v, j of $\left\|A_{u,v}^j\right\|$, we also assume without loss of generality that $K \geq 1$. Thus, since

$$Z_j^{NM} = \sum_{1 \leq u,v \leq N} E_{u,v} \otimes I_k \otimes A_{u,v}^j \ , \quad X_t^N = e^{-t/2} X^N \otimes I_k + (1 - e^{-t})^{1/2} R^{kN} \ ,$$

if L is a monomial in $U_t^k = (X_t^N \otimes I_M, Z^{NM} \otimes I_k, Z^{NM^*} \otimes I_k)$ of degree n, then we can view $L(U_t^k)$ as a sum of at most $2^n N^{2n}$ monomials in $e^{-t/2} X^N \otimes I_k$, $(1 - e^{-t})^{1/2} R^{kN}$, $E_{u,v} \otimes I_k \otimes A_{u,v}^j$, $E_{v,u} \otimes I_k \otimes A_{u,v}^{j\ *}$. Consequently, since $\sup_{u,v,j} \left\|A_{u,v}^j\right\| \leq K$, we have

$$\left\| \mathbb{E}_R \left[\mathrm{Tr}_{kN} \otimes I_M (L(U_t^k) P_{1,2}) \right] \right\| \leq 2^n N^{2n} K^n c_n \ .$$

Thus, if we set

$$f_p(a) = \sum_{0 \leq n \leq p} c_n a^n \ ,$$

we have

$$g_{k,a}(0) \leq f_{\gamma k}(2N^2 K a) \ . \tag{40}$$

Now we need to study the behaviour of $f_k(a)$ when k goes to infinity for a small enough. In order to do so, let us consider a monomial L in $U_{N,k}$. Since $X^N \otimes I_k$ and $E_{u,v} \otimes I_k$ commute with $P_{1,2}$, one can assume that $L = R_i^{kN} S$ for some i (unless L is a monomial in $X^N \otimes I_k$ and $E_{u,v} \otimes I_k$ in which case $\mathrm{Tr}_{kN}(LP_{1,2}) = 0$), thus thanks to Schwinger-Dyson equation (see Proposition 2.10),

$$\mathbb{E}_R \left[\mathrm{Tr}_{kN}(LP_{1,2}) \right] = \frac{1}{Nk} \mathbb{E}_R \left[\mathrm{Tr}_{kN} \otimes \mathrm{Tr}_{kN}(\partial_i (SP_{1,2})) \right] = \frac{1}{Nk} \sum_{S = UR_iV} \mathbb{E}[\mathrm{Tr}_{Nk}(U) \, \mathrm{Tr}_{Nk}(VP_{1,2})]. \tag{41}$$

To use this Schwinger-Dyson equation as an inductive bound we shall use Poincaré inequality to bound the covariance in the above right hand side.We hence compute for any monomial V,

$$\begin{aligned} \|\nabla \, \mathrm{Tr}_{kN}(VP_{1,2})\|_2^2 &= \frac{1}{Nk} \sum_i \sum_{r,s} \mathrm{Tr}_{kN}(\partial_s V \# E_{r,s} P_{1,2}) \, \mathrm{Tr}_{kN}(\partial_s V \# E_{s,r} P_{1,2})^* \\ &= \sum_i \sum_{V = AR_iB, V = CR_iD} \frac{1}{Nk} \, \mathrm{Tr}_{kN}(BP_{1,2} AC^* P_{1,2}^* D^*) \end{aligned} \tag{42}$$

Thus with $\Theta = \max\left\{C, \left\|X^N\right\|, 1\right\}$, since $P_{1,2}$ is of rank N, we get

$$\mathrm{Var}_R(\mathrm{Tr}_{kN}(VP_{1,2})) \leq \frac{1}{k} (\deg V)^2 \Theta^{2 \deg V}.$$

Likewise, for any monomial U, we find

$$\mathrm{Var}_R(\mathrm{Tr}_{kN}(U)) \leq (\deg U)^2 \Theta^{2 \deg U}.$$

Therefore, if n is the degree of L, we deduce from (42), (41) and Poincaré inequality that

$$\begin{aligned} |\mathbb{E}_R \left[\mathrm{Tr}_{kN}(LP_{1,2}) \right]| &\leq \frac{1}{k^{3/2} N} \sum_{i=0}^{n-2} i(n-2-i) \Theta^n + \sum_{S = UR_iV} \left| \frac{1}{Nk} \mathbb{E}_R[\mathrm{Tr}_{kN}(U)] \mathbb{E}_R[\mathrm{Tr}_{kN}(VP_{1,2})] \right| \\ &\leq \frac{n^3 \Theta^n}{k^{3/2} N} + \sum_{S = UR_iV} |\mathbb{E}_R[\mathrm{Tr}_{kN}(VP_{1,2})]| \, \Theta^{\deg U} \ . \end{aligned}$$

By replacing D by $\max\{D, \Theta\}$, we can always assume that $\Theta < D$. We also bound N^{-1} by 1, thus for $n \geq 2$,

$$c_n \leq \frac{n^3 D^n}{k^{3/2}} + \sum_{i=0}^{n-2} c_i D^{n-2-i} \ .$$

We use this estimate to bound $f_g(a)$ with g such that $g^3 D^g \leq \sqrt{k}$. Since $c_0 = c_1 = 0$ and for any $n \leq g$, $n^3 D^n k^{-3/2} \leq k^{-1}$, we have for $aD < 1$

$$f_g(a) = \sum_{n=2}^{g} c_n a^n \leq \frac{1}{k} \times \frac{a^2 - a^{g+1}}{1-a} + a^2 \sum_{m=0}^{g-2} \sum_{n=0}^{m} c_i D^{n-i} a^n \leq \frac{1}{k} \times \frac{a^2}{1-a} + a^2 \frac{f_g(a)}{1-Da} \ .$$

Thus, for a small enough,

$$f_g(a) \leq \frac{(1-Da)a^2}{(1-a)(1-Da-a^2)} \times \frac{1}{k} \ .$$

Besides, we want g such that $g^3 D^g \leq \sqrt{k}$, hence we can take g the integer part of $\frac{\ln k}{2(\ln D+3)}$. Since by definition we have $c_n \leq \Theta^n$, this also means that $c_n \leq D^n$, thus

$$\sum_{g<n\leq \gamma k} c_n a^n \leq \sum_{n>g} (Da)^n \leq \frac{(Da)^{g+1}}{1-Da} \leq k^{\frac{\ln(Da)}{2(\ln D+3)}} \times \frac{1}{1-Da} \ .$$

Thus, if we fix a small enough, $f_{\gamma k}(a) = O(1/k)$. Hence, we deduce from (40) that for a small enough (depending on N, K but not k) there exists a finite constant C independent of$k$ such that

$$g_{k,a}(0) \leq f_k(2N^2 K a) \leq \frac{C}{k}.$$

Therefore, by plugging this inequality in (39), we obtain for a small enough and $y \geq 0$, $g_{k,a}(y) = O(1/k)$. By replacing P by $-P$, we have for a small enough and any $y \in \mathbb{R}$, $g_{k,a}(y) = O(1/k)$. This completes the proof by teh definitions of $g_{k,a}$ and d_n.

□

We can now prove Theorem 1.1.

Proof of Theorem 1.1. It is based on Theorem 3.1. To use it, we would like to take the Fourier transform of f and use Fourier inversion formula. However we did not assume that f is integrable. Thus the first step of the proof is to show that up to a term of order e^{-N}, we can assume that f has compact support. Thanks to Proposition 2.11, there exist constants D and α such that for any N and i, for any $u \geq 0$,

$$\mathbb{P}\left(\left\|X_i^N\right\| \geq u + D\right) \leq e^{-\alpha u N} \ .$$

Thus, there exist constants C and K, independent of M, N, P and f, such that

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}\left(f\left(P\left(X^N \otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\right)\mathbf{1}_{\left\{\exists i, \left\|X_i^N\right\| > D+1\right\}}\right]\right| \\
&\leq \mathbb{E}\left[\left\|f\left(P\left(X^N \otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\right\| \mathbf{1}_{\left\{\exists i, \left\|X_i^N\right\| > D+1\right\}}\right] \\
&\leq \|f\|_\infty \mathbb{P}\left(\exists i, \left\|X_i^N\right\| > D+1\right) \\
&\leq C\|f\|_\infty e^{-KN} \ .
\end{aligned}$$

There exists a polynomial H which only depends on P such that

$$\left\|P\left(X^N \otimes I_M, Z^{NM}, Z^{NM*}\right)\right\| \mathbf{1}_{\left\{\forall i, \left\|X_i^N\right\| \leq D+1\right\}} \leq H\left(\left\|Z^{NM}\right\|\right) \ .$$

We can also assume that $\left\|P(x\otimes I_M, Z^{NM}, Z^{NM*})\right\| \leq H\left(\left\|Z^{NM}\right\|\right)$. We take g a $\mathcal{C}^\infty$-function which takes value 1 on $[-H\left(\left\|Z^{NM}\right\|\right), H\left(\left\|Z^{NM}\right\|\right)]$, 0 on $[-H\left(\left\|Z^{NM}\right\|\right)-1, H\left(\left\|Z^{NM}\right\|\right)+1]^c$ and belongs to $[0,1]$ elsewhere. From the bound above, we deduce

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}\left(f\left(P\left(X^N\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\right)\right] - \tau\Big(f\left(P\left(x\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\Big)\right| \\
&\leq \left|\mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}\left(f\left(P\left(X^N\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\right)\mathbf{1}_{\left\{\forall i, \|X_i^N\|\leq D+1\right\}}\right]\right. \\
&\qquad \left. - \tau\Big(f\left(P\left(x\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\Big)\right| + C\left\|f\right\|_\infty e^{-KN} \qquad (43)\\
&\leq \left|\mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}\left((fg)\left(P\left(X^N\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\right)\right]\right. \\
&\qquad \left. - \tau\Big((fg)\left(P\left(x\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\Big)\right| + 2C\left\|f\right\|_\infty e^{-KN}.
\end{aligned}$$

Since fg has compact support and can be differentiated six times, we can take its Fourier transform and then invert it so that with the convention $\hat{h}(y) = \frac{1}{2\pi}\int_{\mathbb{R}} h(x)e^{-\mathbf{i}xy}dx$, we have

$$\forall x\in\mathbb{R}, \quad (fg)(x) = \int_{\mathbb{R}} e^{\mathbf{i}xy}\widehat{fg}(y)\ dy\ .$$

Besides, since if h has compact support bounded by K then $\left\|\hat{h}\right\|_\infty \leq 2K\left\|h\right\|_\infty$, we have

$$\begin{aligned}
\int_{\mathbb{R}}(|y|+y^4)\ \left|\widehat{fg}(y)\right|\ dy &\leq \int_{\mathbb{R}}\frac{|y|+|y|^3+y^4+y^6}{1+y^2}\ \left|\widehat{fg}(y)\right|\ dy \\
&\leq \int_{\mathbb{R}}\frac{\left|\widehat{(fg)^{(1)}}(y)\right| + \left|\widehat{(fg)^{(3)}}(y)\right| + \left|\widehat{(fg)^{(4)}}(y)\right| + \left|\widehat{(fg)^{(6)}}(y)\right|}{1+y^2}\ dy \\
&\leq 2\left(H\left(\left\|Z^{NM}\right\|\right)+1\right)\left\|fg\right\|_{\mathcal{C}^6}\int_{\mathbb{R}}\frac{1}{1+y^2}\ dy \\
&\leq C\left(H\left(\left\|Z^{NM}\right\|\right)+1\right)\left\|f\right\|_{\mathcal{C}^6}\ ,
\end{aligned}$$

for some absolute constant C. Hence fg satisfies the hypothesis of Theorem 3.1 with $\mu(dy) = \widehat{fg}(y)dy$. Therefore, combining with equation (43), we conclude that

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}\left(f\left(P\left(X^N\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\right)\right] - \tau\Big(f\left(P\left(x\otimes I_M, Z^{NM}, Z^{NM*}\right)\right)\Big)\right| \\
&\leq \left\|f\right\|_\infty e^{-KN} + \frac{M^2}{N^2}L_P\left(\left\|Z^{NM}\right\|\right)\int_{\mathbb{R}}(|y|+y^4)\ \left|\widehat{fg}(y)\right|\ dy \\
&\leq \frac{M^2}{N^2}\left(CL_P\left(\left\|Z^{NM}\right\|\right)\left(H\left(\left\|Z^{NM}\right\|\right)+1\right) + e^{-KN}\right)\left\|f\right\|_{\mathcal{C}^6}\ .
\end{aligned}$$

□

4 Consequences of the main result

In this section, we deduce Corollaries 1.3 and 1.4, as well as Theorems 1.2 and 1.5.

4.1 Proof of Corollary 1.3

We could directly apply Theorem 1.1 to $f_z : x \to (z-x)^{-1}$, however we have $\|f\|_{\mathbb{C}^6} = O\left((\Im z)^7\right)$ when we want an exponent 5. Since $\overline{G_{P(x)}(z)} = G_{P(x)}(\overline{z})$ we can assume that $\Im z < 0$, but then

$$f_z(x) = \int_0^\infty e^{\mathbf{i}xy} \ (\mathbf{i}e^{-\mathbf{i}yz}) \ dy \ .$$

Consequently, with $\mu_z(dy) = \mathbf{i}e^{-\mathbf{i}yz} \ dy$, we have

$$\int_0^\infty (y+y^4) \ d|\mu_z|(y) = \frac{1}{|\Im z|^2} + \frac{24}{|\Im z|^5} \ .$$

Thus, by applying Theorem 3.1 with $Z^{NM} = \left(I_N \otimes Y_1^M, \ldots, I_N \otimes Y_p^M\right)$, P and f_z, we have

$$\left|\mathbb{E}\left[G_{P(X^N \otimes I_M, I_N \otimes Y^M)}(z)\right] - G_{P(x \otimes I_M, I_N \otimes Y^M)}(z)\right| \leq \frac{M^2}{N^2} L_P\left(\left\|Z^{NM}\right\|\right) \int_{\mathbb{R}} (1+y^4) \ d|\mu_z|(y) \ .$$

Now since $\left\|Z^{NM}\right\| = \left(\left\|Y_1^M\right\|, \ldots, \left\|Y_p^M\right\|\right)$ which does not depend on N, we get the desired estimate

$$\left|\mathbb{E}\left[G_{P(X^N)}(z)\right] - G_{P(x)}(z)\right| \leq \frac{M^2}{N^2} L_P\left(\left\|Y_1^M\right\|, \ldots, \left\|Y_p^M\right\|\right) \left(\frac{1}{|\Im z|^2} + \frac{24}{|\Im_z|^5}\right) \ .$$

4.2 Proof of Corollary 1.4

Let $f : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function uniformly bounded by 1 and with Lipschitz constant at most 1. We want to bound from above the quantity

$$\Delta_{N,M}(f) = \left|\mathbb{E}\left[\frac{1}{MN} \operatorname{Tr}_{NM}\left(f\left(P\left(X^N \otimes I_M, I_N \otimes Y_M\right)\right)\right)\right] - \tau \otimes \tau_M\left(f\left(P\left(x \otimes I_M, I_N \otimes Y_M\right)\right)\right)\right| \quad (44)$$

Firstly, one can see that with the same argument as in the proof of Theorem 1.1 (in particular equation (43)), we can assume that the support of f is bounded by a constant $S = H(\left\|Y^M\right\|)$ for some polynomial H independent of everything. However, we cannot apply directly Theorem 1.1 since f is not regular enough. In order to deal with this issue we use the convolution with Gaussian random variable, thus let G be a centered Gaussian random variable, we set

$$f_\varepsilon : x \to \mathbb{E}[f(x + \varepsilon G)] \ .$$

Since f has Lipschitz constant 1, we have for any $x \in \mathbb{R}$,

$$|\mathbb{E}[f(x+\varepsilon G)] - f(x)| \leq \varepsilon \ .$$

Since f_ε is regular enough we could now apply Theorem 1.1, however we a get better result by using Theorem 3.1. Indeed we have

$$\begin{aligned} f_\varepsilon(x) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x+\varepsilon y) e^{-y^2/2} \ dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(y) \frac{e^{-\frac{(x-y)^2}{2\varepsilon^2}}}{\varepsilon} \ dy \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} f(y) \int_{\mathbb{R}} e^{\mathbf{i}(x-y)u} e^{-(u\varepsilon)^2/2} \ du \ dy \ . \end{aligned}$$

Since the support of f is bounded, we can apply Fubini's Theorem:

$$f_\varepsilon(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{\mathbf{i}ux}\int_{\mathbb{R}} f(y)e^{-\mathbf{i}yu}dy\ e^{-(u\varepsilon)^2/2}\ du\ .$$

And so with the convention $\hat{h}(u) = \frac{1}{2\pi}\int_{\mathbb{R}} h(y)e^{-\mathbf{i}uy}dy$, we have

$$f_\varepsilon(x) = \int_{\mathbb{R}} e^{\mathbf{i}ux}\hat{f}(u)e^{-(u\varepsilon)^2/2}du\ .$$

Thus, if we set $\mu_\varepsilon(dy) = \hat{f}(y)e^{-(y\varepsilon)^2/2}dy$, then, since $\|f\|_\infty \leq 1$,

$$\int_{\mathbb{R}}(1+y^4)d|\mu_\varepsilon|(y) \leq 2S\int_{\mathbb{R}}(1+y^4)e^{-y^2/2}\ dy\ \varepsilon^{-5}\ .$$

Consequently, we can apply Theorem 3.1 with f_ε and since $\|f - f_\varepsilon\|_\infty \leq \varepsilon$, there exists a polynomial R_P such that the difference in (44) can be bounded by:

$$\Delta_{N,M}(f) \leq 2\varepsilon + R_P\left(\left\|Y^M\right\|\right)\frac{M^2}{N^2\varepsilon^5}.$$

We finally choose $\varepsilon = N^{-1/3}$ to get the desired bound

$$\Delta_{N,M}(f) \leq 2R_P\left(\left\|Y^M\right\|\right)\frac{M^2}{N^{1/3}}\ .$$

4.3 Proof of Theorem 1.2

Firstly, we need to define properly the operator norm of tensor of $\mathcal{C}^*$-algebras. When writing the proof it appears that we should work with the minimal tensor product.

Definition 4.1. *Let $\mathcal{A}$ and $\mathcal{B}$ be $\mathcal{C}^*$-algebras with faithful representations $(H_\mathcal{A}, \phi_\mathcal{A})$ and $(H_\mathcal{B}, \phi_\mathcal{B})$, then if $\otimes_2$ is the tensor product of Hilbert spaces, $\mathcal{A}\otimes_{\min}\mathcal{B}$ is the completion of the image of $\phi_\mathcal{A}\otimes\phi_\mathcal{B}$ in $B(H_\mathcal{A}\otimes_2 H_\mathcal{B})$ for the operator norm in this space. This definition is independent of the representations that we fixed.*

The following two lemmas are well known facts in operator algebra. The first one is Lemma 4.1.8 from [31]:

Lemma 4.2. *Let $(\mathcal{A}, \tau_\mathcal{A})$ and $(\mathcal{B}, \tau_\mathcal{B})$ be $\mathcal{C}^*$-algebra with faithful traces, then $\tau_\mathcal{A}\otimes\tau_\mathcal{B}$ extends uniquely to a faithful trace $\tau_\mathcal{A}\otimes_{\min}\tau_\mathcal{B}$ on $\mathcal{A}\otimes_{\min}\mathcal{B}$.*

We did not find a reference with an explicit proof for the following Lemma, so we give our own. In order to learn more about this second lemma, especially how to weaken the hypothesis, we refer to [23].

Lemma 4.3. *Let $\mathcal{C}$ be an exact $\mathcal{C}^*$-algebra endowed with a faithful state $\tau_\mathcal{C}$, let $Y^N \in \mathcal{A}_N$ be a sequence of families of noncommutative random variables in a $\mathcal{C}^*$-algebra $\mathcal{A}_N$ which converges strongly towards a family Y in a $\mathcal{C}^*$-algebra $\mathcal{A}$ endowed with a faithful state $\tau_\mathcal{A}$. Let $S \in \mathcal{C}$ be a family of noncommutative random variables, then the family $(S\otimes 1, 1\otimes Y^N)$ converges strongly in distribution towards the family $(S\otimes 1, 1\otimes Y)$.*

Proof. The following sets

$$\mathcal{M} = \left\{(x_N)_{N\in\mathbb{N}}\ \middle|\ x_N \in \mathcal{A}_N, \sup_{N\geq 0}\|x_N\| < \infty\right\}\ ,$$

$$\mathcal{I} = \left\{(x_N)_{N\in\mathbb{N}} \in \mathcal{M}\ \middle|\ \lim_{N\to\infty}\|x_N\| = 0\right\}\ ,$$

are $\mathcal{C}^*$-algebras for the norm $\|x\| = \sup_{N\geq 0}\|x_N\|$. We also define

$$\mathcal{B} = \mathcal{C}^*\left((Y_N)_{N\in\mathbb{N}}\ , \mathcal{I}\right)\ ,$$

the $\mathcal{C}^*$-algebra generated by $\mathcal{I}$ and the family $(Y_N)_{N\in\mathbb{N}}$. Since $\mathcal{I}$ is a closed ideal of $\mathcal{B}$, by Theorem 3.1.4 of [21], $\mathcal{B}/\mathcal{I}$ is a $\mathcal{C}^*$-algebra for the quotient norm. We naturally have the following exact sequence

$$0 \to \mathcal{I} \to \mathcal{B} \to \mathcal{B}/\mathcal{I} \to 0 \ .$$

And by hypothesis, since $\mathcal{C}$ is exact, we have the following exact sequence

$$0 \to \mathcal{C} \otimes_{\min} \mathcal{I} \to \mathcal{C} \otimes_{\min} \mathcal{B} \to \mathcal{C} \otimes_{\min} (\mathcal{B}/\mathcal{I}) \to 0 \ .$$

By definition, this means that $(\mathcal{C} \otimes_{\min} \mathcal{B})/(\mathcal{C} \otimes_{\min} \mathcal{I}) \simeq \mathcal{C} \otimes_{\min} (\mathcal{B}/\mathcal{I})$. If $\pi_{\mathcal{I}}$ is the quotient map from $\mathcal{B}$ to $\mathcal{B}/\mathcal{I}$, the isomorphism between these two spaces is

$$f : x + \mathcal{C} \otimes_{\min} \mathcal{I} \mapsto \mathrm{id}_{\mathcal{C}} \otimes_{\min} \pi_{\mathcal{I}}(x) \ .$$

Hence

$$f(P\big(1 \otimes (Y_N)_{N\in\mathbb{N}}, S \otimes 1\big) + \mathcal{C} \otimes_{\min} \mathcal{I}) = P\big(1 \otimes ((Y_N)_{N\in\mathbb{N}} + \mathcal{I}), S \otimes 1\big) \ . \tag{45}$$

Let (H, φ) be a faithful representation of $\mathcal{C}$, and (H_N, φ_N) a faithful representation of $\mathcal{A}_N$. The direct sum $(\bigoplus_{N\in\mathbb{N}} H_N, \bigoplus_{N\in\mathbb{N}} \varphi_N)$ is a faithful representation of $\mathcal{M}$ and consequently of $\mathcal{B}$ too. More precisely, it is defined by

$$\bigoplus_{N\in\mathbb{N}} H_N = \left\{ (x_N)_{N\in\mathbb{N}} \;\middle|\; x_N \in H_N, \sum_N \|x_N\|_2^2 < \infty \right\} \ .$$

Consequently, by definition of the spatial tensor product, it is the completion of the algebraic tensor $\mathcal{C}\otimes\mathcal{B}$ in the operator space $B\left(H \otimes_2 (\oplus_N H_N)\right)$ endowed with the operator norm. The notation $\otimes_2$ means that we completed the algeraic tensor $H \,\overline{\otimes}\, (\oplus_N H_N)$ to make it a Hilbert space. It is important to see that this space is isomorphic to $\oplus_N (H \otimes_2 H_N)$, indeed it means that if P is a non-commutative polynomial, then

$$\big\|P\big(1 \otimes (Y_N)_{N\in\mathbb{N}}, S \otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{B}} = \sup_{N\geq 0} \big\|P\big(1 \otimes Y_N, S \otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}_{\mathcal{N}}} \ .$$

Consequently by using the definition of the quotient norm, we have

$$\big\|P\big(1 \otimes (Y_N)_{N\in\mathbb{N}}, S \otimes 1\big) + \mathcal{C} \otimes_{\min} \mathcal{I}\big\|_{(\mathcal{C}\otimes_{\min}\mathcal{B})/(\mathcal{C}\otimes_{\min}\mathcal{I})} = \limsup_{N\to\infty} \big\|P\big(1 \otimes Y_N, S \otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}_{\mathcal{N}}} \ . \tag{46}$$

Since f is a $\mathcal{C}^*$-algebra isomorphism, thanks to (45), we have

$$\big\|P\big(1 \otimes (Y_N)_{N\in\mathbb{N}}, S \otimes 1\big) + \mathcal{C} \otimes_{\min} \mathcal{I}\big\|_{(\mathcal{C}\otimes_{\min}\mathcal{B})/(\mathcal{C}\otimes_{\min}\mathcal{I})} = \big\|P\big(1 \otimes ((Y_N)_{N\in\mathbb{N}} + \mathcal{I}), S \otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}(\mathcal{B}/\mathcal{I})} \ .$$

By definition of $\mathcal{I}$, if P is a non-commutative polynomial, we have

$$\|P((Y_N)_{N\in\mathbb{N}} + \mathcal{I})\|_{\mathcal{B}/\mathcal{I}} = \|P(Y)\|_{\mathcal{A}} \ .$$

For our purposes, we can assume that $\mathcal{A} = \mathcal{C}^*(Y)$. Therefore the map

$$P((Y_N)_{N\in\mathbb{N}} + \mathcal{I}) \in \mathbb{C}\langle (Y_N)_{N\in\mathbb{N}} + \mathcal{I}\rangle \mapsto P(Y) \in \mathbb{C}\langle Y\rangle$$

is well-defined and is an isometry. Thus since $\mathbb{C}\langle (Y_N)_{N\in\mathbb{N}} + \mathcal{I}\rangle$ is dense in $\mathcal{B}/\mathcal{I}$ and $\mathbb{C}\langle Y\rangle$ is dense in $\mathcal{A}$, this isometry extends into an isomorphism between $\mathcal{B}/\mathcal{I}$ and $\mathcal{A}$. Consequently

$$\big\|P\big(1 \otimes ((Y_N)_{N\in\mathbb{N}} + \mathcal{I}), S \otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}(\mathcal{B}/\mathcal{I})} = \big\|P\big(1 \otimes Y, S \otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}} \ .$$

Thus, combined with (46), we have

$$\limsup_{N\to\infty} \big\|P\big(1 \otimes Y_N, S \otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}_{\mathcal{N}}} = \big\|P\big(1 \otimes Y, S \otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}} \ . \tag{47}$$

Finally let f be a function which takes value 0 on $(-\infty, \|P(1 \otimes Y, S \otimes 1)\|_{\mathcal{C}\otimes_{\min}\mathcal{A}} - \varepsilon]$ and positive value on $(\|P(1 \otimes Y, S \otimes 1)\|_{\mathcal{C}\otimes_{\min}\mathcal{A}} - \varepsilon, \infty)$. Since the family $(S \otimes 1, 1 \otimes Y^N)$ converges clearly in distribution towards the family $(S \otimes 1, 1 \otimes Y)$, we have

$$\lim_{N\to\infty} \tau_{\mathcal{C}} \otimes_{\min} \tau_{\mathcal{A}_N}\Big(f(P(1\otimes Y_N, S\otimes 1))\Big) = \tau_{\mathcal{C}} \otimes_{\min} \tau_{\mathcal{A}}\Big(f(P(1\otimes Y, S\otimes 1))\Big) \ .$$

Thanks to Lemma 4.2, we know that $\tau_{\mathcal{C}} \otimes_{\min} \tau_{\mathcal{A}}$ is faithful, consequently

$$\tau_{\mathcal{C}} \otimes_{\min} \tau_{\mathcal{A}}\Big(f(P(1\otimes Y, S\otimes 1))\Big) > 0 \ .$$

This means that for N large enough, $\tau_{\mathcal{C}} \otimes_{\min} \tau_{\mathcal{A}_N}\Big(f(P(1\otimes Y_N, S\otimes 1))\Big) > 0$, thus for any $\varepsilon > 0$,

$$\liminf_{N\to\infty} \ \big\|P\big(1\otimes Y_N, S\otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}_{\mathcal{N}}} \geq \big\|P\big(1\otimes Y, S\otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}} - \varepsilon \ .$$

This allows to conclude with (47) that

$$\lim_{N\to\infty} \ \big\|P\big(1\otimes Y_N, S\otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}_{\mathcal{N}}} = \big\|P\big(1\otimes Y, S\otimes 1\big)\big\|_{\mathcal{C}\otimes_{\min}\mathcal{A}} \ .$$

□

In order to prove Theorem 1.2 we use well-known concentration properties of Gaussian random variable coupled with an estimation of the expectation, let us begin by stating the concentration properties (see [2] Lemma 2.3.3).

Proposition 4.4. *Let G be a Lipschitz function on $\mathbb{R}^n$ with Lipschitz constant K for the ℓ^2- norm $\|\gamma\|_2 = (\sum_i \gamma_i^2)^{1/2}$, $\gamma = (\gamma_1, \ldots, \gamma_n)$ independent centered Gaussian random variable of variance 1. Then for all $\delta > 0$,*

$$\mathbb{P}\left(G(\gamma) - \mathbb{E}[G(\gamma)] \geq \delta\right) \leq e^{-\frac{\delta^2}{2K^2}} \ .$$

In our situation, we have p independent GUE matrices $(X^{N,i})_s$ of size N, hence we fix γ the random vector of size dN^2 which consists of the union of $(\sqrt{N}X^{N,i}_{s,s})_{i,s}$, $(\sqrt{2N}\ \Re X^{N,i}_{s,r})_{s<r,i}$ and $(\sqrt{2N}\ \Im X^{N,i}_{s,r})_{s<r,i}$ which are indeed centered Gaussian random variable of variance 1 as stated in Definition 2.8. We would like to apply Proposition 4.4 to

$$G_N(\gamma) = \Big\|P^*P(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\Big\| \ .$$

However G_N is not Lipschitz on $\mathbb{R}^{dN^2}$ because of its polynomial behaviour at infinity. Hence we cannot use directly Proposition 4.4. The following lemma is a well-known tool for this kind of situation, the proof can be found in [14, Lemma 5.9].

Lemma 4.5. *Let (X, d) be a metric space and μ a probability measure on (X, d) which satisfies a concentration inequality, i.e. for all $f : X \to \mathbb{R}$ with Lipschitz constant $|f|_{\mathcal{L}}$, for all $\delta > 0$,*

$$\mu\Big(|f - \mu(f)| \geq \delta\Big) \leq e^{-g\left(\frac{\delta}{|f|_{\mathcal{L}}}\right)}$$

for some increasing function g on $\mathbb{R}^+$. Let B be a subset of X and $|f|^B_{\mathcal{L}}$ be the Lipschitz constant of f as a function from B to $\mathbb{R}$. Let $\delta(f) = \mu(\ \mathbf{1}_{x\in B^c}(|f(x)| + \sup_{u\in B}|f(u)| + |f|^B_{\mathcal{L}} d(x, B))\ )$, then

$$\mu\Big(|f - \mu(f)| \geq \delta + \delta(f)\Big) \leq \mu(B^c) + e^{-g\left(\frac{\delta}{|f|^B_{\mathcal{L}}}\right)} \ .$$

We can now prove the concentration inequality that we will use in the rest of this paper. To simplify notations we will write M instead of M_N. We also set $\text{ÅŞ}Z^{NM} = (Z^N \otimes I_M, I_N \otimes Y^M)$ and $Z = (z\otimes 1, 1\otimes y)$.

Proposition 4.6. *Let $P \in \mathcal{A}_{d,p+q}$, there are some polynomials H_P and K_P which only depends on P such that for any N, M,*

$$\mathbb{P}\Big(\Big| \Big\| P^*P(X^N \otimes I_M, Z^{NM}, Z^{NM*}) \Big\| - \mathbb{E}\left[\Big\| P^*P(X^N \otimes I_M, Z^{NM}, Z^{NM*}) \Big\| \right] \Big| \\ \geq \delta + K_P\left(\left\| Z^{NM} \right\| \right) \, e^{-N} \Big) \leq d \, e^{-2N} + e^{-\frac{\delta^2 N}{H_P(\|Z^{NM}\|)}} \, .$$

Proof. We want to use Lemma 4.5 and Proposition 4.4. The metric space we will work with is $\mathbb{R}^n$ endowed with the Euclidian norm, and we can take the function g to be $g : x \mapsto x^2/2$ by Lemma 4.4. Thus we get that for any $B \subset R^n$, for any $G : R^n \mapsto \mathbb{R}$, if $\gamma = (\gamma_1, \ldots, \gamma_n)$ is a vector of independent centered Gaussian random variables of variance 1, then for all $\delta > 0$,

$$\mathbb{P}\left(G(\gamma) - \mathbb{E}[G(\gamma)] \geq \delta + \delta(G) \right) \leq e^{-\frac{\delta^2}{2(|G|_{\mathcal{L}}^B)^2}} \, . \tag{48}$$

If $0 \in B$ as it will be the case later on, we have $\delta(G) \leq \mathbb{E}[\mathbf{1}_{\gamma \notin B}(|G(\gamma)| + \sup_{u \in B} |G(u)| + |f|_{\mathcal{L}}^B \, \|\gamma\|_2)]$. We set $B_N = \left\{ \forall i, \left\| X_i^N \right\| \leq D \right\}$ where D was chosen thanks to 2.11 such that for any N and i,

$$\mathbb{P}\left(\left\| X_i^N \right\| \geq D \right) \leq e^{-2N} \, . \tag{49}$$

Thus we have $\mathbb{P}(B_N^c) \leq d \, e^{-2N}$. With γ the vector of size dN^2 which consists of the union of $(\sqrt{N} X_{s,s}^{N,i})_{i,s}$, $(\sqrt{2N} \, \Re X_{s,r}^{N,i})_{s<r,i}$ and $(\sqrt{2N} \, \Im X_{s,r}^{N,i})_{s<r,i}$, we set $G_N(\gamma) = \Big\| P^*P(X^N \otimes I_M, Z^{NM}, Z^{NM*}) \Big\|$. One can see that on B_N we can find a polynomial H_P' such that for any N and Z^{NM},

$$|G_N(\gamma) - G_N(\tilde{\gamma})| \leq H_P'\left(\left\| Z^{NM} \right\| \right) \sum_i \left\| X_i^N - \tilde{X}_i^N \right\| \, ,$$

where $\|.\|$ is the operator norm. Besides

$$\sum_i \left\| X_i^N - \tilde{X}_i^N \right\| \leq \sum_i \mathrm{Tr}_N\left((X_i^N - \tilde{X}_i^N)^*(X^N - \tilde{X}_i^N) \right)^{1/2} \leq \frac{2^d}{\sqrt{N}} \, \|\gamma - \tilde{\gamma}\|_2 \, .$$

Thus, on B_N, G_N has Lipschitz constant $2^d H_P'\left(\left\| Z^{NM} \right\| \right) N^{-1/2}$. Consequently with (48), we get that

$$\mathbb{P}\left(G_N(\gamma) - \mathbb{E}[G_N(\gamma)] \geq \delta + \delta(G_N) \right) \leq e^{-\frac{\delta^2 N}{2^{d+1} H_P'(\|Z^{NM}\|)^2}} \, .$$

Therefore, we set $H_P = 2^{d+1} H_P'$, we also have that $\|\gamma\|_2^2 = N \sum_i \mathrm{Tr}_N((X_i^N)^2)$. Consequently we have some polynomial K_P' such that,

$$\delta(G) \leq \mathbb{E}\left[\mathbf{1}_{\left\{ \exists i, \|X_i^N\| > D \right\}} \left(|G_N(\gamma)| + K_P'(\left\| Z^{NM} \right\|) + 2^d H_P'\left(\left\| Z^{NM} \right\| \right) N^{1/2} \sqrt{\sum_i \left\| X_i^N \right\|^2} \right) \right]$$

Hence the conclusion thanks to Proposition 2.11 and our choice of D in equation (49).

□

We can now prove Theorem 1.2. Firstly, we can assume that Z^N and Y^M are deterministic matrices by Fubini's Theorem. The convergence in distribution is a well-known theorem, we refer to [2], Theorem 5.4.5. We set g a $\mathcal{C}^\infty$ function which takes value 0 on $(-\infty, 1/2]$ and value 1 on $[1, \infty)$, and belongs to $[0, 1]$ otherwise. Let us define $f_\varepsilon : t \mapsto g(\varepsilon^{-1}(t - \|PP^*(x \otimes 1, Z, Z^*)\|))$. By Theorem 1.1, there exists a constant C which only depends on P, $\sup_M \left\| Y^M \right\|$ and $\sup_N \left\| Z^N \right\|$ (which is finite thanks to the strong convergence assumption on Z^N) such that,

$$\left| \mathbb{E}\left[\mathrm{Tr}_{MN}\left(f_\varepsilon\left(PP^*\left(X^N \otimes I_M, Z^{NM}, Z^{NM*} \right) \right) \right) \right] - MN\tau_N \otimes \tau_M\Big(f_\varepsilon\left(PP^*\left(x \otimes I_M, Z^{NM}, Z^{NM*} \right) \right) \Big) \right| \\ \leq C\varepsilon^{-6} \frac{M^3}{N} \, .$$

According to Theorem A.1 from [19], $(x, Z^N)_{N\geq 1}$ converges strongly in distribution towards (x, z). Besides thanks to Lemma 4.3 and Corollary 17.10 from [24], we have that $(x\otimes I_M, 1\otimes Y^M)_{M\geq 1}$ converges strongly in distribution towards $(x\otimes 1, 1\otimes y)$. In Theorem 1.2, we are interested in the situation where $Z^{NM} = Z^N \otimes I_M$ or $Z^{NM} = I_N \otimes Y^M$. So, without loss of generality, we restrict ourselves to this kind of Z^{NM}. We know that $(x\otimes I_M, Z^{NM})$ converges strongly towards $(x\otimes 1, Z)$, but since the support of f_ε is disjoint from the spectrum of $PP^*(x\otimes 1, Z, Z^*)$, thanks to Proposition 2.2, for N large enough, $\tau_N \otimes \tau_M\Big(f_\varepsilon(PP^*(x\otimes I_M, Z^{NM}, Z^{NM^*}))\Big) = 0$ and therefore,

$$\mathbb{E}\left[\mathrm{Tr}_{MN}\left(f_\varepsilon\left(PP^*\left(X^N\otimes I_M, Z^{NM}, Z^{NM^*}\right)\right)\right)\right] \leq C\varepsilon^{-6}\frac{M^3}{N}\ . \tag{50}$$

Hence, using Proposition 2.11, we deduce for N large enough,

$$\begin{aligned}
&\mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\right\|\right] - \|PP^*(x\otimes I_M, Z, Z^*)\| \\
&\leq \varepsilon + \int_\varepsilon^\infty \mathbb{P}\left(\left\|PP^*(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\right\| \geq \|PP^*(x\otimes I_M, Z, Z^*)\| + \alpha\right)\ d\alpha \\
&\leq \varepsilon + \int_\varepsilon^K \mathbb{P}\left(\mathrm{Tr}_{NM}\left(f_\alpha(PP^*(X^N\otimes I_M, Z^{NM}, Z^{NM^*}))\right) \geq 1\right)\ d\alpha + Ce^{-N} \\
&\leq \varepsilon + C'\varepsilon^{-5}\frac{M^3}{N}\ .
\end{aligned}$$

Finally we get that,

$$\underset{N\to\infty}{\mathrm{limsup}}\ \mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\right\|\right] \leq \|PP^*(x\otimes I_M, Z, Z^*)\|\ .$$

Besides, we know thanks to Theorem 5.4.5 of [2] that if h is a continuous function taking positive values on $(\|PP^*(x\otimes 1, Z, Z^*)\| - \varepsilon, \infty)$ and taking value 0 elsewhere. Then $\frac{1}{MN}\mathrm{Tr}_{MN}(h(PP^*(X^N\otimes I_M, Z, Z^*)))$ converges almost surely towards $\tau_{\mathcal{A}}\otimes_{\min}\tau_{\mathcal{B}}(h(PP^*(x\otimes 1, Z, Z^*)))$. If this quantity is positive for any h, then for any $\varepsilon > 0$, for N large enough,

$$\left\|PP^*(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\right\| \geq \|PP^*(x\otimes 1, Z, Z^*)\| - \varepsilon\ .$$

Since h is non-negative and the intersection of the support of h with the spectrum of $PP^*(x\otimes 1, Z, Z^*)$ is non-empty, we have that $h(PP^*(x\otimes 1, Z, Z^*)) \geq 0$ and is not 0. Besides, we know that the trace on the space where z is defined is faithful, and so is the trace on the $\mathcal{C}^*$-algebra generated by a single semicircular variable, hence by Theorem 2.3, so is $\tau_{\mathcal{A}}$. Thus, since both $\tau_{\mathcal{A}}$ and $\tau_{\mathcal{B}}$ are faithful, by Lemma 4.2, so is $\tau_{\mathcal{A}}\otimes_{\min}\tau_{\mathcal{B}}$ and $\tau_{\mathcal{A}}\otimes_{\min}\tau_{\mathcal{B}}(h(PP^*(x\otimes 1, Z, Z^*))) > 0$. As a consequence, almost surely,

$$\underset{N\to\infty}{\mathrm{liminf}}\left\|P(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\right\| \geq \|P(x\otimes 1, Z, Z^*)\|\ . \tag{51}$$

Thanks to Fatou's Lemma, we deduce

$$\underset{N\to\infty}{\mathrm{liminf}}\ \mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\right\|\right] \geq \|PP^*(x\otimes I_M, Z, Z^*)\|\ .$$

We conclude that

$$\lim_{N\to\infty}\ \mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\right\|\right] = \|PP^*(x\otimes I_M, Z, Z^*)\|\ . \tag{52}$$

Let us define the following objects,

$$\varepsilon_N = \left|\mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, Z^{NM}, Z^{NM^*})\right\|\right] - \|PP^*(x\otimes I_M, Z, Z^*)\|\right|\ ,$$

$$K = \sup_{N,M\geq 0} K_P\left(\left\|Z^{NM}\right\|\right) + H_P\left(\left\|Z^{NM}\right\|\right)\ .$$

K is finite thanks to the strong convergence of the families Z^N and Y^M. Then thanks to Proposition 4.6, we have that for any $\delta > 0$,

$$\mathbb{P}\Big(\Big| \Big\| P^*P(X^N \otimes I_M, Z^{NM}, Z^{NM^*}) \Big\| - \|PP^*(x \otimes I_M, Z, Z^*)\| \Big| \geq \delta + Ke^{-N} + \varepsilon_N \Big) \leq d\ e^{-2N} + e^{-\frac{\delta^2 N}{K}} \ .$$

Since this is true for any $\delta > 0$, by Borel-Cantelli's Lemma, almost surely,

$$\lim_{N\to\infty} \Big\| PP^*(X^N \otimes I_M, Z^{NM}, Z^{NM^*}) \Big\| = \|PP^*(x \otimes 1, Z, Z^*)\| \ .$$

We finally conclude thanks to the fact that for any y in a $\mathcal{C}^*$-algebra, $\|yy^*\| = \|y\|^2$.

4.4 Proof of Theorem 1.5

We first prove the following estimate that we use multiple times during the proofs.

Lemma 4.7. *Let g be a $\mathcal{C}^\infty$ function which takes value 0 on $(-\infty, 1/2]$ and value 1 on $[1, \infty)$, and in $[0, 1]$ otherwise. We set $f_\varepsilon : t \mapsto g(\varepsilon^{-1}(t - \alpha))$ with $\alpha = \Big\| PP^*(x \otimes I_M, 1 \otimes Y^M) \Big\|$, then there exists a constant C such that for any $\varepsilon > 0$ and N,*

$$\mathbb{E}\left[\frac{1}{MN} \mathrm{Tr}_{NM}\left(f_\varepsilon(PP^*(X^N \otimes I_M, I_N \otimes Y^M)) \right) \right] \leq C \frac{\varepsilon^{-4}}{N^2} \ .$$

Proof. To estimate the above expectation we use the same method as in the proof of Theorem 1.2 with a few refinements to have an optimal estimate with respect to ε. We set $f_\varepsilon^\kappa : t \mapsto g(\varepsilon^{-1}(t-\alpha))g(\varepsilon^{-1}(\kappa - t) + 1)$ with $\alpha = \|PP^*(x \otimes I_M, 1 \otimes Y_M)\|$ and $\kappa > \alpha$. Since g has compact support and is sufficiently smooth we can apply Theorem 3.1. Setting $h : t \mapsto g(t - \varepsilon^{-1}\alpha)g(\varepsilon^{-1}\kappa + 1 - t) = \hat{f_\varepsilon^\kappa}(\varepsilon t)$, we have

$$\begin{aligned}
2\pi \int y^4 |\hat{f_\varepsilon^\kappa}(y)|\ dy &= \int y^4 \left| \int g(\varepsilon^{-1}(t-\alpha))g(\varepsilon^{-1}(\kappa - t) + 1) e^{-\mathbf{i}yt}\ dt \right|\ dy \\
&= \int y^4 \left| \int h(t) e^{-\mathbf{i}y\varepsilon t}\ \varepsilon dt \right|\ dy \\
&= \varepsilon^{-4} \int y^4 \left| \int h(t) e^{-\mathbf{i}yt}\ dt \right|\ dy \\
&\leq \varepsilon^{-4} \int \frac{1}{1+y^2}\ dy \int (|h^{(4)}(t)| + |h^{(6)}(t)|)\ dt \ .
\end{aligned}$$

The derivatives $h^{(4)}$ and $h^{(6)}$ are uniformly bounded independently of t or ε. Since the support of these functions is included in $[\varepsilon^{-1}\alpha, \varepsilon^{-1}\alpha + 1] \cup [\varepsilon^{-1}\kappa, \varepsilon^{-1}\kappa + 1]$, there is a universal constant C such that for any ε and κ,

$$\int y^4 |\hat{f_\varepsilon^\kappa}(y)|\ dy \leq C\varepsilon^{-4} \ .$$

With similar computations we can find a constant C such that for any ε and κ,

$$\int (|y| + y^4) |\hat{f_\varepsilon^\kappa}(y)|\ dy \leq C\varepsilon^{-4} \ . \tag{53}$$

Since the support of f_ε^κ is disjoint from the spectrum of $PP^*(x \otimes I_M, 1 \otimes Y^M)$, for any ε and N one have $\tau \otimes \tau_M\Big(f_\varepsilon^\kappa(PP^*(x \otimes I_M, 1 \otimes Y^M)) \Big) = 0$. Consequently thanks to Theorem 3.1, we have a constant C such that for any N, ε and κ,

$$\mathbb{E}\left[\frac{1}{MN} \mathrm{Tr}_{NM}\left(f_\varepsilon^\kappa(PP^*(X^N \otimes I_M, I_N \otimes Y^M)) \right) \right] \leq C \frac{\varepsilon^{-4}}{N^2} \ .$$

Then by the monotone convergence Theorem, we deduce

$$\mathbb{E}\left[\mathrm{Tr}_{NM}\left(f_\varepsilon(PP^*(X^N\otimes I_M, I_N\otimes Y^M))\right)\right] = \lim_{\kappa\to\infty}\mathbb{E}\left[\mathrm{Tr}_{NM}\left(f_\varepsilon^\kappa(PP^*(X^N\otimes I_M, I_N\otimes Y^M))\right)\right] \ .$$

Hence we have a constant C such that for any N and $\varepsilon > 0$,

$$\mathbb{E}\left[\frac{1}{MN}\mathrm{Tr}_{NM}\left(f_\varepsilon(PP^*(X^N\otimes I_M, I_N\otimes Y^M))\right)\right] \le C\frac{\varepsilon^{-4}}{N^2} \ .$$

□

We finally complete the proof of Theorem 1.5. One can view $X^N = (X_1^N, \ldots, X_d^N)$ as the random vector of size dN^2 which consists of the union of $(\sqrt{N}X_{s,s}^{N,i})_{i,s}$, $(\sqrt{2N}\,\Re X_{s,r}^{N,i})_{s<r,i}$ and $(\sqrt{2N}\,\Im X_{s,r}^{N,i})_{s<r,i}$ which are indeed centered Gaussian random variable of variance 1 as stated in Definition 2.8. Thus we can apply the Poincaré inequality (see Proposition 2.9) to the function

$$\varphi : X^N \mapsto \frac{1}{MN}\mathrm{Tr}_{MN}\left(f_\varepsilon(PP^*(X^N\otimes I_M, I_N\otimes Y^M))\right) \ ,$$

and we get

$$\mathrm{Var}\left(\varphi(X^N)\right) \le \frac{1}{(MN)^2}\mathbb{E}[\|\nabla\varphi(X^N)\|_2^2]$$

Besides, as in the proof of Lemma 3.6, if $Q \in \mathcal{A}_{d,p+q}$,

$$N\left\|\nabla\,\mathrm{Tr}_{MN}\left(Q(X^N\otimes I_M, I_N\otimes Y^M)\right)\right\|_2^2 = \sum_s\sum_{i,j}\mathrm{Tr}_{MN}\left(D_sQ\ E_{i,j}\otimes I_M\right)\mathrm{Tr}_{MN}\left(D_sQ\ E_{i,j}\otimes I_M\right)^* \ .$$

Besides, if f_k is a polynomial with a single variable, then $D_sf_k(PP^*) = \partial_s(PP^*)\widetilde{\#}f_k'(PP^*)$. Thus, taking f_k such that f_k' converges towards f_ε' for the sup norm on the spectrum of $PP^*(X^N\otimes I_M, I_N\otimes Y^M)$, we deduce that

$$\mathrm{Var}\left(\varphi(X^N)\right) \le \frac{1}{M^2N^3}\sum_{s,i,j}\mathbb{E}\left[\mathrm{Tr}_{MN}\left(\partial_s(PP^*)\widetilde{\#}f_\varepsilon'(PP^*)\ E_{i,j}\otimes I_M\right)\mathrm{Tr}_{MN}\left(\partial_s(PP^*)\widetilde{\#}f_\varepsilon'(PP^*)\ E_{i,j}\otimes I_M\right)^*\right] \ .$$

Now with $A = \partial_s(PP^*)\widetilde{\#}f_\varepsilon'(PP^*)$,

$$\begin{aligned}\sum_{i,j}\mathrm{Tr}_{MN}\left(A\ E_{i,j}\otimes I_M\right)\mathrm{Tr}_{MN}\left(A\ E_{i,j}\otimes I_M\right)^* &= \sum_{i,j,k,l} g_j^*\otimes e_k^*Ag_i\otimes e_k\ g_i^*\otimes e_l^*A^*g_j\otimes f_l\\ &= \sum_{j,k,l} g_j^*\ (I_N\otimes e_k^*\ A\ I_N\otimes e_k\ I_N\otimes e_l^*\ A^*\ I_N\otimes e_l)\ g_j\\ &= \mathrm{Tr}_N\left(I_N\otimes\mathrm{Tr}_M(A)\ I_N\otimes\mathrm{Tr}_M(A^*)\right)\\ &= \mathrm{Tr}_N\left(I_N\otimes\mathrm{Tr}_M(A)\ (I_N\otimes\mathrm{Tr}_M(A))^*\right) \ .\end{aligned}$$

So by contractivity of the conditional expectation over $\mathbb{M}_N(\mathbb{C})\otimes I_M$, that is $I_N\otimes\frac{1}{M}\mathrm{Tr}_M$, we have

$$\sum_{i,j}\mathrm{Tr}_{MN}\left(A\ E_{i,j}\otimes I_M\right)\mathrm{Tr}_{MN}\left(A\ E_{i,j}\otimes I_M\right)^* \le \mathrm{Tr}_{MN}(AA^*)\ M \ .$$

As a consequence, we find that

$$\mathrm{Var}\left(\varphi(X^N)\right) \le \frac{1}{N^3M}\sum_s\mathbb{E}\left[\mathrm{Tr}_{MN}\left(\partial_s(PP^*)\widetilde{\#}f_\varepsilon'(PP^*)\ (\partial_s(PP^*)\widetilde{\#}f_\varepsilon'(PP^*))^*\right)\right] \ .$$

Besides, if U, V and W are monomials,

$$|\mathrm{Tr}_{MN}(Uf'_\varepsilon(PP^*)Vf'_\varepsilon(PP^*)W)| \leq \sqrt{\mathrm{Tr}_{MN}(Uf'^2_\varepsilon(PP^*)U^*)\ \mathrm{Tr}_{MN}(Vf'_\varepsilon(PP^*)WW^*f'_\varepsilon(PP^*)V^*)}$$
$$\leq \mathrm{Tr}_{MN}(f'^2_\varepsilon(PP^*))\,\|U\|\,\|V\|\,\|W\| \ .$$

Therefore there is a constant C depending only on P and $\sup_i \left\|Y_i^M\right\|$ such that

$$\mathrm{Var}\left(\varphi(X^N)\right) \leq \frac{C}{N^2}\mathbb{E}\left[\prod_s \left(\left\|X_s^N\right\|^{2\deg P}+1\right)\frac{1}{MN}\mathrm{Tr}_{NM}\left(\left|f'_\varepsilon(PP^*(X^N\otimes I_M, I_N\otimes Y^M))\right|^2\right)\right] \ .$$

Thanks to Proposition 2.11, we can find w and α such that for any s and $u \geq 0$,

$$\mathbb{P}\left(\left\|X_s^N\right\| \geq w+u\right) \leq e^{-\alpha u N} \ .$$

There is a constant C independent of N and ε such that

$$\mathrm{Var}\left(\varphi(X^N)\right) \leq \frac{C}{N^2}\left(\mathbb{E}\left[\frac{1}{MN}\mathrm{Tr}_{NM}\left((f'_\varepsilon)^2(PP^*(X^N\otimes I_M, I_N\otimes Y^M))\right)\right]+\varepsilon^{-2}e^{-N}\right) \ . \tag{54}$$

We can now apply Theorem 3.1 to the right hand side of the above equation, noticing that (53) still holds if we replace f_ε^κ by $(\varepsilon f'_\varepsilon)^2$. As a consequence, we find an inequality similar the one of Lemma 4.7 and thus a constant C such that for any N or ε,

$$\mathrm{Var}\left(\frac{1}{MN}\mathrm{Tr}_{NM}(f_\varepsilon(PP^*(X^N\otimes I_M, I_N\otimes Y^M)))\right) \leq C\left(\frac{\varepsilon^{-6}}{N^4}+\varepsilon^{-2}e^{-N}\right) \ .$$

Therefore, thanks to Lemma 4.7 there exists a constant C such that for any $N \in \mathbb{N}$ and ε such that $\varepsilon^4 > C\frac{M}{N}$,

$$\begin{aligned}
&\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| \geq \left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\|+\varepsilon\right)\\
&\leq \mathbb{P}\left(\frac{1}{MN}\mathrm{Tr}_{NM}\left(f_\varepsilon(PP^*(X^N\otimes I_M, I_N\otimes Y^M))\right) \geq \frac{1}{MN}\right)\\
&\leq \mathbb{P}\left(\left|\frac{1}{MN}\mathrm{Tr}_{NM}\left(f_\varepsilon(PP^*)\right)-\mathbb{E}\left[\frac{1}{MN}\mathrm{Tr}_{NM}\left(f_\varepsilon(PP^*)\right)\right]\right| \geq \frac{1}{MN}-\frac{C}{N^2\varepsilon^4}\right)\\
&\leq C\left(\frac{\varepsilon^{-6}}{N^4}+\varepsilon^{-2}e^{-N}\right)\left(\frac{1}{MN}-\frac{C}{N^2\varepsilon^4}\right)^{-2} \ .
\end{aligned}$$

Let us now set $s = cN^{-1/4}$ with c a constant such that for any N,

$$\frac{1}{MN}-\frac{C}{N^2s^4} \geq \frac{1}{2MN} \ .$$

Therefore, if $x_+ = \max(x, 0)$, we have for some constant C,

$$\begin{aligned}
&\mathbb{E}\left[\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\|-\left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\|\right)_+\right]\\
&= \int_{\mathbb{R}^+}\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| \geq \left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\|+\varepsilon\right)\,d\varepsilon\\
&\leq s+4CM^2N^2\int_s^\infty \frac{\varepsilon^{-6}}{N^4}+\varepsilon^{-2}e^{-N}d\varepsilon \leq s+4CM^2N^2(s^{-5}N^{-4}+s^{-1}e^{-N})\\
&\leq CN^{-1/4} \ .
\end{aligned}$$

On one side, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\|\right] \geq \delta + K_P\left(\left\|Y^M\right\|\right)\ e^{-N}\right)\\
&\geq \mathbb{P}\Big(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\|\\
&\qquad\geq \delta + K_P(\left\|Y^M\right\|)e^{-N} + \mathbb{E}\left[\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\|\right)_+\right]\Big)\\
&\geq \mathbb{P}\left(\left|\left\|P(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|P(x\otimes I_M, 1\otimes Y^M)\right\|\right| \geq \frac{\delta + CN^{-1/4}}{\left\|P(x\otimes I_M, 1\otimes Y^M)\right\|}\right)\ .
\end{aligned}$$

On the other side, thanks to Proposition 4.6, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\|\right] \geq \delta + K_P\left(\left\|Y^M\right\|\right)\ e^{-N}\right)\\
&\leq e^{-\frac{\delta^2}{H_P(\|Y^M\|)}N} + de^{-2N}\ .
\end{aligned}$$

Hence we can find constants K and C such that for any $N\in\mathbb{N}$ and $\delta>0$,

$$\mathbb{P}\left(\left\|P(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|P(x\otimes I_M, 1\otimes Y^M)\right\| \geq \delta + CN^{-1/4}\right) \leq e^{-K\delta^2 N} + de^{-2N}\ .$$

And we get (4) by replacing δ by $N^{-1/4}\delta$.

The other inequality is trickier because we need to study the spectral measure of $PP^*(x\otimes I_M, 1\otimes Y^M)$, which is far from easy. We mainly rely on the Theorem 1.1 from [26]. We summarize the part of this theorem which is interesting for us in the proposition below.

Proposition 4.8. *Let $x=(x_1,\ldots,x_d)$ be a system of free semicircular variable, $p_{i,j}\in\mathbb{C}\langle X_1,\ldots,X_d\rangle$ be such that $S=(p_{i,j}(x))_{i,j}$ is self-adjoint with spectral measure ρ with support K. Then there exists a finite subset $A\subset\mathbb{R}$ such that if I is a connected component of $\mathbb{R}\backslash A$, then either $\rho_{|I}=0$, or $I\subset K$. In the second situation there exists an analytic function g defined for some $\delta>0$ on*

$$W := \{z\in\mathbb{C}|\ |\Im z|<\delta\}\setminus\bigcup_{a\in A}\{a-\mathbf{i}t|\ t\in\mathbb{R}^+\}$$

such that for each $a\in A$, there exist $N\in\mathbb{N}$ and $\epsilon>0$ such that $(z-a)^N g(z)$ admits an expansion on $W\cap\{z\in\mathbb{C}|\ |z-a|<\epsilon\}$ as a convergent powerseries in $r_N(z-a)$ where $r_N(z)$ is the analytic N^{th}-root of z defined with branch $C\setminus\{-\mathbf{i}t|\ t\in\mathbb{R}^+\}$. Then $\Im g_{|I}$ is the probability density function of $\rho_{|I}$.

What this means for us is that at the edge of the spectrum of $PP^*(x\otimes I_M, 1\otimes Y^M)$, either we have an atom or the density of the spectral measure decays like $\frac{1}{|x-a|^r}$ with $r\in\mathbb{Q}$ when approaching a. Consequently we can find $\beta\geq 0$ such that if ρ is the spectral measure of $PP^*(x\otimes I_M, 1\otimes Y^M)$ then for $\varepsilon>0$ small enough,

$$\rho\left(\left[\left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\| - \varepsilon, \infty\right]\right) \geq \varepsilon^\beta\ .$$

Consequently if once again g is a $\mathcal{C}^\infty$ function which takes value 0 on $(-\infty,0]$, 1 on $[1/2,\infty)$, and belongs to $(0,1]$ otherwise. We then take $f_\varepsilon: t\mapsto g(\varepsilon^{-1}(t-\left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\|+\varepsilon))$ for some $\varepsilon\geq 0$. Then

$$\begin{aligned}
&\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| \leq \left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\| - \varepsilon\right)\\
&= \mathbb{P}\left(\frac{1}{MN}\operatorname{Tr}_{NM}\left(f_\varepsilon(PP^*(X^N\otimes I_M, I_N\otimes Y^M))\right) = 0\right)\\
&\leq \mathbb{P}\left(\left|\frac{1}{MN}\operatorname{Tr}_{NM}\left(f_\varepsilon(PP^*)\right) - \mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}_{NM}\left(f_\varepsilon(PP^*)\right)\right]\right| \geq \mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}_{NM}\left(f_\varepsilon(PP^*)\right)\right]\right)\\
&\leq \frac{\operatorname{Var}\left(\frac{1}{MN}\operatorname{Tr}_{NM}\left(f_\varepsilon(PP^*)\right)\right)}{\mathbb{E}\left[\frac{1}{MN}\operatorname{Tr}_{NM}\left(f_\varepsilon(PP^*)\right)\right]^2}\ .
\end{aligned}$$

Thanks to (54), we have

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{MN}\,\mathrm{Tr}_N\left(f_\varepsilon(PP^*)\right)\right) &\le \frac{C}{N^2}\left(\mathbb{E}\left[\frac{1}{MN}\,\mathrm{Tr}_{NM}\left((f_\varepsilon')^2(PP^*)\right)\right]+\varepsilon^{-2}e^{-N}\right)\\ &\le \frac{C}{N^2}\left(\|f_\varepsilon'\|^2+\varepsilon^{-2}\right)\le \frac{C'}{N^2}\varepsilon^{-2}\,.\end{aligned}$$

On the contrary, with the same kind of computations which let us get Lemma 4.7, we can find constants C and K such that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{MN}\,\mathrm{Tr}_{NM}\left(f_\varepsilon(PP^*)\right)\right] &\ge \tau\otimes\tau_M(f_\varepsilon(PP^*))-C\frac{\varepsilon^{-4}}{N^2}\\ &\ge \rho\left(\left[\left\|PP^*(x\otimes I_M,1\otimes Y^M)\right\|-\varepsilon/2,\infty\right]\right)-C\frac{\varepsilon^{-4}}{N^2}\ge K\min(1,\varepsilon)^\beta-C\frac{\varepsilon^{-4}}{N^2}\,.\end{aligned}$$

Therefore we find finite constants C and K such that

$$\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|\le\left\|PP^*(x\otimes I_M,1\otimes Y^M)\right\|-\varepsilon\right)\le\frac{K}{N^2\varepsilon^2}\left(\min(1,\varepsilon)^\beta-C\frac{\varepsilon^{-4}}{N^2}\right)^{-2}\,.$$

Now we fix $r=cN^{-1/(3+\beta)}$, with c constant such that for any N,

$$\min(1,r)^\beta-\frac{C}{N^2r^4}\ge\frac{\min(1,r)^\beta}{2}\,.$$

Then, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\left\|PP^*(x\otimes I_M,1\otimes Y^M)\right\|-\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|\right)_+\right]\\ &=\int_{\mathbb{R}^+}\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|\le\left\|PP^*(x\otimes I_M,1\otimes Y^M)\right\|-\varepsilon\right)d\varepsilon\\ &\le r+4KN^{-2}\int_r^\infty\varepsilon^{-2}\min(1,\epsilon)^{-2\beta}d\varepsilon\le r+4KN^{-2}(r^{-1-2\beta}+1)\\ &\le CN^{-1/(3+\beta)}\,.\end{aligned}$$

We deduce the following bound

$$\begin{aligned}&\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|-\mathbb{E}\left[\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|\right]\le-\delta-K_P\left(\left\|Y^M\right\|\right)\ e^{-N}\right)\\ &\ge\mathbb{P}\Big(\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|-\left\|PP^*(x\otimes I_M,1\otimes Y^M)\right\|\\ &\qquad\le-\delta-K_P(\left\|Y^M\right\|)e^{-N}-\mathbb{E}\left[\left(\left\|PP^*(x\otimes I_M,1\otimes Y^M)\right\|-\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|\right)_+\right]\Big)\\ &\ge\mathbb{P}\Big(\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|-\left\|PP^*(x\otimes I_M,1\otimes Y^M)\right\|\le-\delta-CN^{-1/(3+\beta)}\Big)\,.\end{aligned}$$

Since on the event $\left\{\forall i,\left\|X_i^N\right\|\le D\right\}$ with D as in (49), we have

$$\begin{aligned}&\left\|PP^*(X^N\otimes I_M,I_N\otimes Y^M)\right\|-\left\|PP^*(x\otimes I_M,1\otimes Y^M)\right\|\\ &\le\left(\left\|P(X^N\otimes I_M,I_N\otimes Y^M)\right\|-\left\|P(x\otimes I_M,1\otimes Y^M)\right\|\right)\left(J_P(\left\|Y^M\right\|)+\left\|P(x\otimes I_M,1\otimes Y^M)\right\|\right)\ ,\end{aligned}$$

we deduce that

$$\begin{aligned}
&\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\|\right] \le -\delta - K_P\left(\left\|Y^M\right\|\right)\ e^{-N}\right)\\
&\ge \mathbb{P}\Big(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|PP^*(x\otimes I_M, 1\otimes Y^M)\right\| \le -\delta - CN^{-1/(3+\beta)} \text{ and } \forall i, \left\|X_i^N\right\| \le D\Big)\\
&\ge \mathbb{P}\left(\left\|P(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|P(x\otimes I_M, 1\otimes Y^M)\right\| \le \frac{-\delta - CN^{-1/(3+\beta)}}{J_P(\|Y^M\|) + \|P(x\otimes I_M, 1\otimes Y^M)\|}\right)\\
&\quad - \mathbb{P}(\exists i, \left\|X_i^N\right\| \ge D)\\
&\ge \mathbb{P}\left(\left\|P(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|P(x\otimes I_M, 1\otimes Y^M)\right\| \le \frac{-\delta - CN^{-1/(3+\beta)}}{J_P(\|Y^M\|) + \|P(x\otimes I_M, 1\otimes Y^M)\|}\right)\\
&\quad - de^{-2N}\ .
\end{aligned}$$

On the other side thanks to Proposition 4.6, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \mathbb{E}\left[\left\|PP^*(X^N\otimes I_M, I_N\otimes Y^M)\right\|\right] \le -\delta - K_P\left(\left\|Y^M\right\|\right)\ e^{-N}\right)\\
&\le d\ e^{-2N} + e^{-\frac{\delta^2 N}{H_P(\|Y^M\|)}}\ .
\end{aligned}$$

Hence we can find constants K and C such that for any $\delta > 0$,

$$\mathbb{P}\left(\left\|P(X^N\otimes I_M, I_N\otimes Y^M)\right\| - \left\|P(x\otimes I_M, 1\otimes Y^M)\right\| \le -\delta - CN^{-1/(3+\beta)}\right) \le e^{-K\delta^2 N} + 2d\ e^{-2N}\ .$$

And we get (5) by replacing δ by $N^{-1/(3+\beta)}\delta$.

Acknowledgements

B. C. was partially funded by JSPS KAKENHI 17K18734, 17H04823, 15KK0162. F. P. benefited also from the aforementioned Kakenhi grants and a MEXT JASSO fellowship. A. G. and F. P. were partially supported by Labex Milyon (ANR-10-LABX-0070) of Université de Lyon. The authors would like to thank Narutaka Ozawa for supplying reference [31] for Lemma 4.2.

References

[1] G.W. Anderson, Convergence of the largest singular value of a polynomial in independent Wigner matrices. *Ann. Probab.* 41, no. 3B, 2103–2181, 2013.

[2] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[3] Z. D. Bai and Y. Q. Yin. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. *Ann. Probab.*, 16(4):1729?1741, 1988.

[4] P. Biane and R. Speicher, Free diffusions, free entropy and free Fisher information, *Ann. de l'inst. Henry Poinc.* B 37, 581, 2001.

[5] Cabanal-Duvillard, Thierry, Fluctuations de la loi empirique de grandes matrices aléatoires, *Ann. Inst. H. Poincaré Probab. Statist.*, 37 (3), 373–402, 2001.

[6] M. Capitaine and C. Donati-Martin, Strong asymptotic freeness for Wigner and Wishart matrices. *Indiana Univ. Math. J.*, 56(2):767–803, 2007.

[7] S. Belinschi and M. Capitaine, Spectral properties of polynomials in independent Wigner and deterministic matrices, *Journal of Functional Analysis*, 273, 3901–3973, 2016.

[8] 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.

[9] Erdős, László and Krüger Torben, and Nemish Yuriy, Local laws for polynomials of Wigner matrices, *arXiv :1804.11340* 2018.

[10] L. Erdős, B. Schlein, and H.T. Yau, Wegner estimate and level repulsion for Wigner random matrices, *Int. Math. Res. Not. IMRN*,436–479, 2010.

[11] L. Erdős and H.T. Yau, *A Dynamical Approach to Random Matrix Theory*, volume 28 of *Courant Lecture Notes*. American Mathematical Soc., 2017.

[12] A. Figalli, and A. Guionnet, Universality in several-matrix models via approximate transport maps, *Acta Math.*, 217 (1), 81–176, 2016.

[13] Z. Füredi and J. Komlós, The eigenvalues of random symmetric matrices, Combinatorica, vol. 1, no. 3, 233–241, 1981.

[14] A. Guionnet, *Large Random Matrices: Lectures on Macroscopic Asymptotics: École d'Été de Probabilités de Saint-Flour XXXVI – 2006*. In *Lecture Notes in Mathematics*. Springer, 2009.

[15] A. Guionnet and E. Maurel-Segala, Second order asymptotics for matrix models, *Ann. Probab.* 35, 2160–2212, 2007.

[16] A. Guionnet, Large deviations upper bounds and central limit theorems for non-commutative functionals of Gaussian large random matrices, *Ann. Inst. H. Poincaré Probab. Statist.*, 38 (3),341–384, 2002.

[17] U. Haagerup and S. Thorbjørnsen, A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(\mathbb{F}_2))$ is not a group. *Ann. of Math.*, 162(2):711–775, 2005.

[18] J.O. Lee, and J. Yin, A necessary and sufficient condition for edge universality of Wigner matrices, *Duke Math. J.*, 163(1),117–173, 2014.

[19] C. Male, The norm of polynomials in large random and deterministic matrices. With an appendix by Dimitri Shlyakhtenko. *Probab. Theory Related Fields* 154, no. 3-4, 477-532, 2012.

[20] J. Mingo and R. Speicher, Second order freeness and fluctuations of random matrices: I. Gaussian and Wishart matrices and cyclic Fock spaces. *J. Funct. Anal.*, 235:226–270, 2006.

[21] G.J. Murphy, *C*-Algebras and Operator Theory*. Elsevier Science, 1990.

[22] 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.

[23] G. Pisier, Random Matrices and Subexponential Operator Spaces. Israel Journal of Mathematics, 203, 2012.

[24] G. Pisier, *Introduction to Operator Space Theory* . In *London Mathematical Society Lecture Note Series*. Cambridge University Press, 2003.

[25] H. Schultz, Non-commutative polynomials of independent Gaussian random matrices. The real and symplectic cases. *Probab. Theory Related Fields*, 131(2):261–309, 2005.

[26] D. Shlyakhtenko and P. Skoufranis, Freely independent random variables with non-atomic distributions, *Trans. Am. Math. Soc.* 367, no. 9, 6267–6291, 2015.

[27] A. Soshnikov, Universality at the edge of the spectrum in Wigner random matrices, *Comm. Math. Phys.*, 207,no. 3, 697–733, 1999.

[28] T. Tao, and V. Vu, Random matrices: universality of local eigenvalue statistics up to the edge, *Comm. Math. Phys.*, 298(2),549–572, 2010.

[29] C. A. Tracy and H. Widom, Level spacing distributions and the Airy kernel, *Comm. Math. Phys.*, 159, 151-174, 1994.

[30] D. Voiculescu. Limit laws for random matrices and free products. *Invent. Math.*, 104(1):201–220, 1991.

[31] N.P. Brown and N. Ozawa, *C^*-algebras and Finite-dimensional Approximations*. In *Graduate studies in mathematics*. American Mathematical Soc., 2008.