Probab. Theory Relat. Fields (2009) 143:481–516
DOI 10.1007/s00440-007-0133-7

Universality results for the largest eigenvalues of some sample covariance matrix ensembles

Sandrine Péché

Received: 3 October 2007 / Revised: 5 December 2007 / Published online: 8 January 2008

Abstract For sample covariance matrices with i.i.d. entries with sub-Gaussian tails, when both the number of samples and the number of variables become large and the ratio approaches one, it is a well-known result of Soshnikov that the limiting distribution of the largest eigenvalue is same that of Gaussian samples. In this paper, we extend this result to two cases. The first case is when the ratio approaches an arbitrary finite value. The second case is when the ratio becomes infinite or arbitrarily small.

1 Introduction

The scope of this paper is to study the limiting behavior of the largest eigenvalues of real and complex sample covariance matrices with independent identically distributed (i.i.d.), but non necessarily Gaussian, entries. Consider a sample of size p of i.i.d. $N \times 1$ random vectors $\vec{y}_1, \ldots, \vec{y}_p$. We further assume that the sample vectors $\vec{y}_k$ have mean zero and *covariance* $\Sigma = Id$. We use $X = [\vec{y}_1, \ldots, \vec{y}_p]$ to denote the $N \times p$ data matrix and $M_N = \frac{1}{N} XX^*$ to denote the *sample covariance matrix*. Random sample covariance matrices have been first studied in mathematical statistics ([5,11,13]). A huge literature deals with the case where $p \to \infty$, N being fixed, which is now quite well understood. Contrary to the traditional assumptions, it is currently of strong interest to study the case where N is of the same order as p,

S. Péché (✉)
Institut Fourier BP 74, 100 Rue des maths, 38402 Saint Martin d'Heres, France
e-mail: Sandrine.Peche@ujf-grenoble.fr

Present Address:
S. Péché
Department of Mathematics, University of California at Davis, One Shields Ave., Davis, CA 95616, USA

due to the large amount of data available. In particular, the limiting behavior of the largest eigenvalues is important for testing hypotheses on the covariance matrix Σ. Here we focus on the simple case, $H_o : \Sigma = Id$ versus $H_a : \Sigma \neq Id$, and study the asymptotic distribution of extreme eigenvalues under the H_o. The study of extreme eigenvalues is also of interest in principal component analysis. We refer the reader to [15] and [8] for a review of statistical applications. Other examples of applications include genetics [21], mathematical finance [24,18,19], wireless communication [35], physics of mixture [26], and statistical learning [12]. We point out that the spectral properties of M_N readily translate to the companion matrix $W_N = \frac{1}{N} X^* X$. Indeed, W_N is a $p \times p$ matrix, of rank N, with the same non-zero eigenvalues as M_N. Thus, it is enough to study the spectral properties of M_N to give a complete picture of the spectrum of such sample covariance matrices.

1.1 Model and results

We consider both real and complex random sample covariance matrices

$$M_N = \frac{1}{N} X X^*,$$

where X is a $N \times p$, $p = p(N) \geq N$, random matrix satisfying certain "moment conditions". In the whole paper, we set $\gamma_N = \frac{p}{N}$. We assume that the entries X_{ij}, $1 \leq i \leq N$, $1 \leq j \leq p$, of the sequence of random matrices $X = X_N$ are non-necessarily Gaussian random variables satisfying the following conditions. First, in the complex case,

(i) $\{\Re e X_{i,j},\ \Im m X_{i,j}\ :\ 1 \leq i \leq N,\ 1 \leq j \leq p\}$ are real independent random variables,
(ii) all these real variables have symmetric laws (thus, $\mathbb{E}[X_{i,j}^{2k+1}] = 0$ for all $k \in \mathbb{N}^*$),
(iii) $\forall\, i, j$, $\mathbb{E}[(\Re e X_{i,j})^2] = \mathbb{E}[(\Im m X_{i,j})^2] = \frac{1}{2}$,
(iv) all their other moments are assumed to be sub-Gaussian, i.e. there exists a constant $\tau > 0$ such that uniformly in i, j and k,

$$\mathbb{E}[|X_{i,j}|^{2k}] \leq (\tau\, k)^k.$$

In the real setting, $X = (X_{i,j})_{1 \leq i,j \leq N}$ is a random matrix such that

(i′) the $\{X_{i,j},\ 1 \leq i \leq N\,,\ 1 \leq j \leq p\}$ are independent random variables,
(ii′) the laws of the $X_{i,j}$ are symmetric (in particular, $\mathbb{E}[X_{i,j}^{2k+1}] = 0$),
(iii′) for all i, j, $\mathbb{E}[X_{i,j}^2] = 1$,
(iv′) all the other moments of the X_{ij} are sub-Gaussian, i.e. there exists a constant $\tau > 0$ such that, uniformly in i, j and k, $\mathbb{E}[X_{i,j}^{2k}] \leq (\tau\, k)^k$.

When the entries of X are further assumed to be Gaussian random variables, we denote by X_G the corresponding model. If the entries of X are complex Gaussian random variables, M_N^G is the so-called Laguerre unitary ensemble (LUE), which is also called the complex Wishart ensemble. In the real setting, M_N^G is the Laguerre orthogonal ensemble (LOE) or real Wishart ensemble.

The goal of this paper is to describe the large-N-limiting distribution of the K largest eigenvalues induced by any such ensemble, for any fixed integer K independent of N. Two regimes are investigated in this paper. In the first part, we assume that there exists a constant $\gamma \geq 1$ such that $\lim_{N\to\infty} \gamma_N = \gamma$. In the second part, we consider the case where $\lim_{N\to\infty} \gamma_N = \infty$.

Before stating our results, we recall some known results about sample covariance matrices. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ be the ordered eigenvalues induced by any ensemble of the above type. We first focus on the case where $\lim_{N\to\infty} \gamma_N = \gamma < \infty$. The first fundamental result for the limiting spectral behavior of such random matrix ensembles has been obtained by Marchenko and Pastur in [20] (in a much more general context than here). It is in particular proved therein that the spectral measure $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{\lambda_i}$ a.s. converges as N goes to infinity. Set $u_\pm^c = (1 \pm \sqrt{\gamma})^2$. Then one has that

$$\lim_{N\to\infty} \mu_N = \rho_{MP} \text{ a.s.,} \quad \text{where } \frac{d\rho_{MP}(x)}{dx} = \frac{\sqrt{(u_+^c - x)(x - u_-^c)}}{2\pi x} 1_{[u_-^c, u_+^c]}(x). \quad (1)$$

The limiting probability distribution ρ_{MP} is the so-called Marchenko–Pastur distribution.

The above result gives no insight about the behavior of the largest eigenvalues. The first study of the asymptotic behavior of the largest eigenvalue goes back to Geman [10]. It was later refined in [2] and [27]. In particular, it is well known that $\lim_{N\to\infty} \lambda_1 = u_+^c$ a.s. if the entries of the random matrix X admit moments up to order 4. Significant results about fluctuations of the largest eigenvalues around u_+^c are much more recent and are essentially established for Wishart ensembles only. In particular, the limiting distribution of the largest eigenvalue has been obtained by Johansson [14] for complex Wishart matrices and Johnstone [15] for real Wishart matrices. Soshnikov [31] has derived for both ensembles the limiting distribution of the K largest eigenvalues, for any fixed integer K. Their results, proved in the case where $\gamma < \infty$, was later extended by El Karoui [7] to the case where p, N go to infinity and $p/N \to \infty$. Before recalling all these results, we need a few definitions. We here define the limiting Tracy–Widom distribution for the largest eigenvalue. Let Ai denote the standard Airy function and q denote the solution of the Painlevé II differential equation $\frac{\partial^2}{\partial x^2} q = xq(x) + 2q^3(x)$, with boundary condition $q(x) \sim Ai(x)$ as $x \to +\infty$.

Definition 1.1 The GUE (resp. GOE) Tracy–Widom distribution for the largest eigenvalue is defined by the cumulative distribution function $F_2(x) = \exp\{\int_x^\infty (x-t)q^2(t)dt\}$ (resp. $F_1(x) = \exp\{\int_x^\infty \frac{-q(t)}{2} + \frac{(x-t)}{2} q^2(t)dt\}$).

The GUE (resp. GOE) Tracy–Widom distribution for the joint distribution of the K largest eigenvalues (for any fixed integer K) has been also defined. We refer the reader to [36] and [37] for a precise definition.

We then rescale the eigenvalues as follows. For any ensemble satisfying (i) to (iν)

(resp. (i') to (iν')), we set for $i = 1, \ldots, N$,

$$\mu_i = \frac{\gamma_N^{1/6}}{(1+\sqrt{\gamma_N})^{4/3}} N^{2/3} \left(\lambda_i - (1+\sqrt{\gamma_N})^2 \right). \tag{2}$$

When the entries of X are futher assumed to be Gaussian random variables, we denote by $\mu_i^{G,\beta}$ the above rescaled eigenvalues where $\beta = 2$ (resp. $\beta = 1$) in the complex (resp. real) case.

Theorem 1.1 [7,14,15,31]. *The joint distribution of the K rescaled largest eigenvalues $\mu_i^{G,2}$ (resp. $\mu_i^{G,1}$), $1 \le i \le K$, of the LUE (resp. LOE) converges, as $N \to \infty$, to the joint distribution defined by the GUE (resp. GOE) Tracy–Widom law. This holds true if $\gamma < \infty$ as well as if $\gamma_N \to \infty$.*

The proof of Theorem 1.1 relies on the crucial fact that the joint eigenvalue density of the Wishart ensembles can be explicitly computed. Starting from numerical simulations, it was then conjectured, in [15], e.g., that Theorem 1.1 actually holds for a class of random sample covariance matrices much wider than the Wishart ensembles. Such a universality result was later proved for some quite general ensembles by Soshnikov [31], yet under some restriction on the sample size, as we now recall.

Theorem 1.2 [31] *Assume that $p - N = O(N^{1/3})$. The joint distribution of the K rescaled largest eigenvalues μ_i, $i \le K$, induced by any ensemble satisfying* (i) *to* (iν) *(resp.* (i') *to* (iν')) *converges, as N goes to infinity, to the joint distribution defined by the GUE (resp. GOE) Tracy–Widom law.*

In this paper, we prove that such a universality result holds for any value of the parameter γ. This is the main result of this note.

Theorem 1.3 *The joint distribution of the K rescaled largest eigenvalues μ_i, $i \le K$, induced by any ensemble satisfying* (i) *to* (iν) *(resp.* (i') *to* (iν')) *converges, as N goes to infinity, to the joint distribution defined by the GUE (resp. GOE) Tracy–Widom law. This holds true in both the cases where $\gamma < \infty$ and $\gamma_N \to \infty$.*

Remark 1.1 Assumptions (iν) and (iν') can actually be relaxed if $\gamma < \infty$. This relaxation is discussed in the second paragraph of Sect. 1.2.

Before giving secondary results, we give a few comments on the way we proceed to prove Theorem 1.3. In Theorem 1.2, the reason for the restriction on $p - N$ follows from the idea of the proof used therein. When $\gamma = 1$, the eigenvalues of a random sample covariance matrix roughly behave as the squares of those of a typical Wigner random matrix. This adequacy still works for the largest eigenvalues, but fails if γ is not close enough to one. Theorem 1.2 has been proved using universality results established for classical Wigner random matrices. Here, we revisit the problem of computing the asymptotics of $\mathbb{E}\left[\mathrm{Tr} M_N^L\right]$ for some powers L that may go to infinity, using combinatorial tools specifically well suited for the study of spectral functions of sample covariance matrices. It is known that Dyck paths and Catalan numbers are associated to standard Wigner matrices (see [1]). Suitable combinatorial tools in the case of sample covariance matrices are the so-called *Narayana numbers* and some particular Dyck paths. Using those, we can extend the universality result of [31] to any value of the ratio γ.

The case where $\gamma \leq 1$ can also be considered thanks to the companion matrix W_N. Let λ'_i, $1 \leq i \leq p$, be the eigenvalues of $\frac{1}{p}X^*X$, ordered in decreasing order and let $\delta_N = \gamma_N^{-1}$, so that $\lim_{N\to\infty}\delta_N = \gamma^{-1} \leq 1$. We set:

$$\mu'_i = \frac{{\delta_N}^{1/6}}{(1+\sqrt{\delta_N})^{4/3}}\, p^{2/3}\left(\lambda'_1 - (1+\sqrt{\delta_N})^2\right), \quad i = 1,\ldots,p.$$

Corollary 1.1 *Under the assumptions* (i) *to* (iv) (*resp.* (i') *to* (iv')), *the joint distribution of* $(\mu'_1,\ldots,\mu'_K)$ *converges as* $N\to\infty$ *to the GUE* (*resp. GOE*) *Tracy–Widom joint distribution of the* K *largest eigenvalues.*

1.2 Discussion on some implications of the result

Our result has some statistical flavor. Testing homogeneity of a population has long been of interest in mathematical statistics. It is often a preliminary step in discriminant analysis and cluster analysis. We consider here the test of the null hypothesis $H_o : \Sigma = Id$ versus the alternative hypothesis $H_a : \Sigma \neq Id$, assuming high dimensionality. The result stated here for the largest eigenvalue can be formulated as

$$\lim_{N\to\infty} P(\mu_1 \leq x | H_o) = F_{2(1)}(x). \tag{3}$$

This theoretical result was established for Gaussian samples only so far (cf. [16] for a review). Removing the Gaussian assumption is fundamental for various statistical problems. For instance, samples in genetic data are usually drawn from a distribution with compact support and the size of matrices encountered therein is typically large enough so that (3) should be observed for appropriate models (see [21]).

Regarding the assumptions made on the distribution of the entries, they may appear strong. It is indeed believed (and observed numerically) that the universal Tracy–Widom picture holds as soon as the entries X_{ij} admit a fourth moment. The moment assumptions (iv) and (iv') can actually be relaxed, using truncation techniques and ideas from [25]. We can show that Theorem 1.3 holds under the assumption that $P(|X_{ij}| > x) \leq C(1+x)^{m_o}, \forall i, j$, for some $m_o > 36$. We do not consider this case here, which would increase the technicalities of the paper. Our result differs from that in [25], since we do not understand Formula 4.7 (which roughly gives universality if $m_o > 18$). In Remarks 2.3 and 2.5, we indicate the changes to be made to consider such a case (and also justify a weaker version of the above cited Formula 4.7).

The symmetry assumption is also a technical assumption for the proof. Indeed, it is expected that the lack of symmetry has no impact on the limiting distribution of the largest eigenvalues (provided the distribution is centered). Yet, analytical tools to prove such a result are not established (see e.g. [22] for some recent progress).

The method we develop is also a first step towards considering samples with non-Identity covariance. Such results are of practical importance for understanding the behavior of principal component analysis and dimension reduction in high dimensional setting. It is therefore important to consider covariance matrices with more complex structures. In particular (in progress), the moment approach developed here seems to be well suited in the case where the population covariance is a so-called "spiked" diagonal matrix. That is, $\Sigma = Id + D$, where the deformation D is a finite rank diagonal matrix. This is important, since the test based on (3) may not reject H_o if the largest eigenvalue of D is not large enough, because of a phase transition phenomenon described, e.g. in [4] and [3].

It also seems possible to refine the combinatorics used in this paper to consider sample covariance matrices in the form commonly used in statistics, that is $\frac{1}{N}(X - \overline{X})(X - \overline{X})^*$ where $\overline{X}$ is the empirical sample mean. It is believed that the above centering, used in the case of where the entries are i.i.d. with unknown mean value, does not have any impact on the limiting behavior of the largest eigenvalues. The study of such sample covariance matrix ensembles (in progress) is deferred to another paper.

1.3 Sketch of the proof

We here give the main ideas of the proof of Theorem 1.3. The proof follows essentially the strategy introduced in [31] and we refer to this paper for most of the detail. We focus on the case where $\gamma = \lim_{N\to\infty} p/N < \infty$. Basically, we compute the leading term in the asymptotic expansion of expectations of traces of high powers of M_N:

$$\mathbb{E}\left[\mathrm{Tr}\left(\frac{1}{N}XX^*\right)^{s_N}\right], \tag{4}$$

where $\mathrm{Tr}(A) = \sum_{i=1}^N a_{ii}$ denotes throughout the paper the un-normalized trace of a matrix $A = (a_{ij})_{1\le i,j\le N}$. Here s_N is a sequence such that there exists some constant $c > 0$ with $\lim_{N\to\infty} \frac{s_N}{N^{2/3}} = c$. It is indeed expected that the largest eigenvalues exhibit fluctuations in the scale $N^{-2/3}$ around

$$u_+ := (\sqrt{\gamma_N} + 1)^2. \tag{5}$$

The core of the proof is to show the following result. Let K be some given integer and $c_1, \ldots, c_K$ be constants chosen in some compact interval $J \subset (0, +\infty)$ (independent of N). Let $s_N^{(i)}$, $i = 1, \ldots, K$, be sequences such that $\lim_{N\to\infty} \frac{s_N^{(i)}}{N^{2/3}} = c_i$ (or $\lim_{N\to\infty} \frac{s_N^{(i)}}{\sqrt{\gamma_N}N^{2/3}} = c_i$ if $\gamma_N \to \infty$). We show that

$$\exists \tilde{C}_1 = \tilde{C}_1(K, J) > 0 \quad \text{such that} \left|\mathbb{E}\left[\prod_{i=1}^K Tr\left(\frac{XX^*}{Nu_+}\right)^{s_N^{(i)}}\right]\right| \le \tilde{C}_1; \tag{6}$$

$$\left|\mathbb{E}\left[\prod_{i=1}^K Tr\left(\frac{XX^*}{Nu_+}\right)^{s_N^{(i)}}\right] - \mathbb{E}\left[\prod_{i=1}^K Tr\left(\frac{X_G X_G^*}{Nu_+}\right)^{s_N^{(i)}}\right]\right| = o(1). \tag{7}$$

Formula (7) claims universality of moments of traces of powers of M_N in the scale $(\sqrt{\gamma_N})N^{2/3}$. Using the machinery developed in [30] (Sects. 2 and 5) and [31] (Sect. 2), we can then deduce that the limiting joint distribution of any fixed number of largest eigenvalues for sample covariance matrices satisfying (i) to (iν) (resp. (i') to (iν')) is the same as for complex (resp. real) Wishart ensembles. We roughly give the main idea. On the one hand, the Laplace transform of the joint distribution of a finite number of the rescaled eigenvalues μ_i can be conveniently expressed in terms of joint moments of traces as in (4). On the other hand, the asymptotic distribution of these rescaled largest eigenvalues (and also the corresponding Laplace transform) is well-known in the Wishart setting. One can then deduce from universality of moments of traces that the asymptotic joint distribution of the largest eigenvalues for any ensemble considered here is the same as for the corresponding Wishart ensemble. The detail of the derivation of such a result from formula (7), including the required asymptotics of correlation functions for Wishart ensembles, can be found in [30,31] and [7]. The improvement we obtain with respect to [31] is that Formula (7) holds for any value γ. Our result is due to a refinement in the counting procedure of [31].

The paper is organized as follows. In Sect. 2, we introduce the so-called Narayana numbers. These numbers are the major combinatorial tools needed to adapt the computations of [31] to sample covariance matrices of any sample size to dimension ratio γ_N. We also establish a central limit theorem for traces of high powers of M_N. Section 3 essentially yields formulas (6) and (7) and is is based on the computations made in [31]. Finally, in Sect. 4, we consider the case where $\gamma_N \to \infty$, which requires some minor modifications.

2 Combinatorics

In this section, we define the combinatorial objects suitable for the computation of moments of the spectral measure of random sample covariance matrices. These combinatorial objects are the *Narayana paths*, that is Dyck paths with a prescribed number of up steps at the odd instants, and are directly related to the so-called Marchenko–Pastur distribution. Then, we give the basic technical estimates needed to compute the moments of traces of powers of sample covariance matrices. We illustrate our counting strategy by giving a refinement of the Marchenko–Pastur theorem and also obtain a Central Limit Theorem.

2.1 Dyck paths and Narayana numbers

Let s_N be some integer that may depend on N. Developing (4), we obtain that

$$\mathbb{E}\left[\mathrm{Tr}\left(XX^*\right)^{s_N}\right] = \sum_{i_o,\ldots,i_{s_N-1}} \sum_{j_o,\ldots,j_{s_N-1}} \mathbb{E}\left(X_{i_o j_o}\overline{X_{i_1 j_o}}\cdots X_{i_{s_N-1} j_{s_N-1}}\overline{X_{i_o j_{s_N-1}}}\right), \tag{8}$$

$$\text{where } i_l \in \{1, 2, \ldots, N\} \quad \text{and} \quad j_l \in \{1, \ldots, p\}, \quad 0 \le l \le s_N - 1. \tag{9}$$

In the whole paper, we denote by R the rule (9) for the choice of indices in (8). We shall later prove that such a rule plays a fundamental role in the asymptotics of (4). To each term in the expectation (8), we associate three combinatorial objects that will be needed in the following.

First, to each term $X_{i_o j_o}\overline{X_{i_1 j_o}}\cdots X_{i_{s_N-1} j_{s_N-1}}\overline{X_{i_o j_{s_N-1}}}$ occurring in (8), we associate the following " edge path" $\mathcal{P}_E$, formed with oriented edges (read from bottom to top)

$$\binom{j_o}{i_o}\binom{j_o}{i_1}\binom{j_1}{i_1}\cdots\binom{j_{s_N-1}}{i_{s_N-1}}\binom{j_{s_N-1}}{i_o}. \tag{10}$$

Due to the symmetry assumption on the entries of X, the sole paths giving a non zero contribution in (8) are such that each oriented edge appears an even number of times. From now on, we consider only such even edge paths.

To such an even edge path, we also associate a so-called Dyck path, which is a trajectory $x(t), 0 \leq t \leq 2s_N$, of a simple random walk on the positive half-lattice such that

$$x(0) = 0, \quad x(2s_N) = 0; \quad \forall t \in [0, 2s_N], \ x(t) \geq 0 \quad \text{and} \quad x(t) - x(t-1) = \pm 1.$$

We start the path at the origin and draw up steps $(1, +1)$ and down steps $(1, -1)$ as follows. We read successively the $2s_N$ edges of (10), reading each edge from bottom to top. Then if the (oriented) edge is read for an odd number of times, we draw an up step. Otherwise we draw a down step. We obtain in this way a trajectory with s_N up and s_N down steps, which is clearly a Dyck path. We shall now estimate the number of possible trajectories associated to the edge path. Due to the constraint (9) on the choices for vertices, we shall distinguish trajectories with respect to the number of up steps performed at the odd instants. Indeed, they are the moments of time where the vertices can be chosen in the set $\{1, \ldots, p\}$.

In the whole paper, we denote by $k = k(x)$ the number of up steps performed at an odd instant in a Dyck path x. We also call $\mathcal{X}_{s_N,k}$ the set of Dyck paths of length $2s_N$ with k odd up steps.

Proposition 2.1 [6] *Let* $\mathbf{N}(s_N, k)$ *be the so-called kth Narayana number defined by*

$$\mathbf{N}(s_N, k) = \frac{1}{s_N} C_{s_N}^{k} C_{s_N}^{k-1}. \tag{11}$$

Then $\mathbf{N}(s_N, k) = \sharp\mathcal{X}_{s_N,k}$.

Remark 2.1 For more details about Narayana numbers and their occurrences in various combinatorial problems, we refer the reader to the work of Sulanke [33,34] as well as Stanley [32].

Narayana numbers are intimately linked with Dyck paths. Let $D(2s_N) = \frac{1}{s_N} C_{2s_N}^{s_N+1}$ be the Catalan number counting the number of Dyck paths of length $2s_N$. It is obvious that $\sum_{k=1}^{s_N} \mathbf{N}(s_N, k) = D(2s_N)$. Narayana numbers are also linked to the moments of the Marchenko–Pastur distribution defined in (1), since the following was proved by Jonsson [17] (see also [23] and [1]).

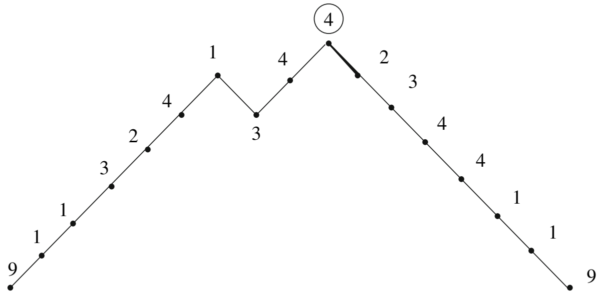

Fig. 1 The path P_k with $k = 4$, $s_N = 8$

Proposition 2.2 *For any integer L, one has that*

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\mathrm{Tr} M_N^L\right] = \sum_{k=1}^{L} \gamma^k \mathbf{N}(L,k) = \int x^L d\rho_{MP}(x). \tag{12}$$

Remark 2.2 Proposition 2.2 was actually proved for a broader class of sample covariance matrices than that considered in this paper.

Last, we associate to the edge path $\mathcal{P}_E$ a “usual” path, which we denote by P_k, as follows. We mark on the underlying Dyck path x the successive vertices met in the edge path. The path P_k associated to (10) is then $i_o\, j_o\, i_1\, j_1 \dots j_{s_N-1}\, i_o$. For instance, the path associated to the path $\mathcal{P}_E = \binom{1}{9}\binom{1}{1}\binom{3}{1}\binom{3}{2}\binom{4}{2}\binom{4}{1}\binom{3}{1}\binom{3}{4}\binom{4}{4}\binom{4}{2}\binom{3}{2}\binom{3}{4}\binom{4}{4}\binom{4}{1}\binom{1}{1}\binom{1}{9}$ is given on Fig. 1.

The three structures $\mathcal{P}_E$, x and P_k introduced here will now be used to compute the moments of traces of (high) powers of M_N. Our counting strategy is as follows. Given a Dyck path x, we shall estimate the number of edge paths that can be associated to this Dyck path. We shall also estimate their contribution to the expectation (8). This is the object of the next two sections.

2.2 Marked vertices

In this section, we bring out the connection between Narayana paths and the restriction for the choices of vertices occurring in the path imposed by the rule R. Given a Dyck path $\{x(t), 0 \le t \le 2s_N\} \in \mathcal{X}_{s_N,k}$, we shall now count the number of ways to mark the vertices using the rule R. In this way, we count the number of paths P_k associated to a given Dyck path. The terminology we use is close to the one used in [28–30] and [31]. We recall the main definitions that will be needed here and also assume that the reader is acquainted with most of the techniques used in the above papers.

The first task is to choose the pairwise distinct vertices occurring in the path. There are at most p^{s_N+1} such vertices. We shall now define the "marked vertices", separating the cases where they are marked at odd or even instants.

Definition 2.1 The instant $t \in \{1, 2, \dots, 2s_n\}$ is said to be marked if the tth step of the Dyck path x is up. A vertex from $\{1, 2, \dots, N, \dots, p\}$ occurring in P_k at a marked instant is said to be a marked vertex.

Marked instants correspond to the moments of time (apart from $t = 0$) where, considering the top and bottom lines separately, one can possibly "discover" some vertex not already encountered. Consider first the vertices on the top line of the edge path $\mathcal{P}_E$, that is, vertices occurring at the odd instants in the path P_k. For $0 \le i \le s_N$, call $\mathcal{T}_i$ the class of vertices of $\{1, \dots, p\}$ occurring i times as a marked vertex at an odd instant. Then if we set $p_i = \sharp\mathcal{T}_i$, one has

$$p = \sum_{i=0}^{s_N} p_i \quad \text{and} \quad \sum_{i\ge 1} i p_i = k.$$

Note that each time we "discover" on the top line some new vertex, the corresponding instant is necessarily marked. Consider also the vertices on the bottom line. For $0 \le i \le s_N$, denote by $\mathcal{N}_i$ the class of vertices of $\{1, \dots, N\}$ occurring i times as a marked vertex at an even instant. Then one has, if $n_i = \sharp\mathcal{N}_i$,

$$N = \sum_{i=0}^{s_N} n_i \quad \text{and} \quad \sum_{i\ge 1} i n_i = s_N - k.$$

Note that a vertex from the set $\{1, 2, \dots, N\}$ can occur as a marked vertex on both lines. Yet it is the type of the vertex on each line which is here taken into account. Thanks to the above definitions, we characterize a path P_k by its associated Dyck path x and its type:

$$(n_o, n_1, \dots, n_{s_N})\,(p_o, p_1, \dots, p_{s_N}), \quad \text{with}\, n_i = 0, \forall\, i > s_N - k,\, p_i = 0\, \forall i > k.$$

For short, we denote by $(\tilde{n}, \tilde{p})$ the type of a path. We also use the following notations. A vertex $v \in \mathcal{T}_i$ (resp. $v \in \mathcal{N}_i$) is said to be of type i on the top (resp. bottom) line. Any vertex $v \in \cup_{i\ge 2}\mathcal{T}_i$ (resp. $v \in \cup_{i\ge 2}\mathcal{N}_i$) is said to be a vertex of self-intersection on the top (resp. bottom) line. An odd (resp. even) marked instant $0 \le t' \le 2s_N$ is an instant of self-intersection if there exists an odd (resp. even) marked instant $0 \le t < t'$ such that the vertices occurring at t and t' are equal.

The choice of marked vertices is enough to determine the distinct vertices of the path P_k, if the origin of the path also occurs as a marked vertex. We will see that for typical paths, this is not the case and $i_o \in \mathcal{N}_0$. Thus, given the type $(\tilde{n}, \tilde{p})$ of the path, the number of ways to assign vertices at the marked instants and choose the origin is then at most:

$$N\frac{N!}{\prod_{i=0}^{s_N} n_i!}\frac{p!}{\prod_{i=0}^{s_N} p_i!}\frac{k!}{\prod_{i\ge 2}(i!)^{p_i}}\frac{(s_N-k)!}{\prod_{i\ge 2}(i!)^{n_i}}. \tag{13}$$

Indeed, one distributes the vertices of $\{1, \ldots, N\}$ and $\{1, \ldots, p\}$ into the possible classes $\mathcal{N}_i, \mathcal{T}_i, 1 \leq i \leq s_N$, choose the corresponding marked occurrences of each vertex and fix the origin. Once the marked vertices and the origin of the path are chosen, there remains to fill in the blanks of the path, i.e. assign vertices at the unmarked instants. Due to self-intersections, there are multiple ways to do so. We investigate this numbering in the sequel and consider at the same time the expectation $\mathbb{E}\left[\prod_{j=0}^{s_N-1} M_{i_j i_{j+1}}\right]$ of the whole path P_k.

2.3 Filling in the blanks of the path

Assume that the Dyck path x and the type $(\tilde{n}, \tilde{p})$ of a path are given. Call $\Omega_{k,(\tilde{n},\tilde{p})}$ the number of ways to fill in the blanks of the path once the marked vertices and the origin are known.

Proposition 2.3 *Set* $(\Omega\mathbb{E})_{max} := \Omega_{k,(\tilde{n},\tilde{p})} \left|\mathbb{E}\left[\prod_{j=0}^{s_N-1} M_{i_j i_{j+1}}\right]\right|$. *There exists* $\tilde{C} > 0$ *independent of* p, N, k *and* s_N *such that*

$$(\Omega\mathbb{E})_{\max} \leq \frac{2}{N^{s_N}} \prod_{l=2}^{s_N-k} \left(\tilde{C}l\right)^{ln_l} \prod_{m=2}^{k} \left(\tilde{C}m\right)^{mp_m}. \tag{14}$$

Proof of Proposition 2.3 We only sketch the proof which follows essentially the same steps as in Lemma 1 of [29]. Assume that in P_k, at the unmarked instant t, one makes a down step with left vertex i. If i is of type 1, then there is no ambiguity in the choice of the right endpoint of such an edge. In general, the maximal number of possible right endpoints depends on the multiplicity of i as a marked vertex. Now, the parity of t specifies whether i is a vertex on the top or on the bottom line of $\mathcal{P}_E$. Thus, the sole top or bottom multiplicity of i has to be taken into account to estimate the number of ways to close the edge. Then, it is not hard to see that the number of ways to fill in the blanks of the path is at most

$$2 \prod_{l=2}^{s_N-k} (2l)^{ln_l} \prod_{m=2}^{k} (2m)^{mp_m}.$$

The extra factor 2 comes from the case (negligible) where the origin is of type 1. To consider simultaneously the expectation of the path, one also has to take into account the number of times each oriented edge is read. Assume that an edge $\begin{pmatrix} v \\ w \end{pmatrix}$ is read $2q$ times, with $q \geq 2$. Call $l_u(v; w)$ (resp. $l_d(w; v)$) the number of times v (resp. w) is a marked vertex of this edge. Then, if $l_d(w; v)l_u(v; w) > 0$, one has that $\mathbb{E}|M_{vw}|^{2q} \leq (\tau q)^q \leq (2\tau l_u(v; w))^{l_u(v;w)} (2\tau l_d(w; v))^{l_d(w;v)}$. Now if an oriented edge is read $2l$ times then it is closed l times along the same edge. That is, we overcount the number of ways to fill in the blanks of the path. Thus, if we let $2l(ij)$ be the number of times the oriented edge (ij) is read in the path, we obtain that

$$\frac{(\Omega\mathbb{E})_{\max}}{2} \leq \prod_{l(ij)>1} \frac{(C_\tau l(ij))^{l(ij)}}{l(ij)!} \prod_{l=2}^{s_N-k} (2l)^{ln_l} \prod_{m=2}^{k} (2m)^{mp_m}$$
$$\leq \prod_{l=2}^{s_N-k} \left(\tilde{C}l\right)^{ln_l} \prod_{m=2}^{k} \left(\tilde{C}m\right)^{mp_m},$$

where $C_\tau, \tilde{C}$ are some constants independent of p, k, N and s_N. □

Remark 2.3 In the case where the entries X_{ij} have polynomial tails, with $P(|X_{ij}| \geq x) \leq (1+x)^{m_o}$ for some $m_o > 36$, one can first consider (up to a set of negligible probability) that all the entries of X are smaller in absolute value than $\Gamma_N := N^{2/m_o+\epsilon_o}$ for some $\epsilon_o > 0$ small enough. This is true if $\gamma < \infty$. Then Proposition 2.3 has to be replaced with

$$\frac{(\Omega\mathbb{E})_{max}}{2N^{-s_N}} \leq \prod_{l=2}^{s_N-k} \left(lC(1+\Gamma_N^4 1_{l\geq \frac{m_o}{4}})\right)^{ln_l} \prod_{m=2}^{k} \left(mC(1+\Gamma_N^4 1_{m\geq \frac{m_o}{4}})\right)^{mp_m}.$$

This follows from the fact that to each edge seen $2l \geq 2m_o$ times there corresponds at least $l/2$ (and not l) marked occurrences of one of its endpoints. This is the reason why our formula differs from Formula 4.7 in [25].

In the next two sections, we investigate moments of Traces of powers of M_N in scales $s_N \ll \sqrt{N}$. This will give the foundations for the asymptotics of higher moments.

2.4 Narayana numbers and Marchenko–Pastur distribution

In this section, we illustrate our counting strategy and present a refinement of (12) allowing to consider higher moments than in Proposition 2.2.

Proposition 2.4 *If $s_N \ll \sqrt{N}$, one has that*

$$\frac{1}{N}\mathbb{E}\left[\mathrm{Tr} M_N^{s_N}\right] = \sum_{k=1}^{s_N} \gamma_N^k \mathbf{N}(s_N, k)(1+o(1)).$$

Proof of Proposition 2.4 The proof is similar to that of the classical Wigner theorem using Dyck paths (see e.g. [1]). It is divided into two steps. First, we show that paths for which $\sum_{i\geq 2} n_i + p_i > 0$ yield a negligible contribution to $\mathbb{E}\left[\mathrm{Tr} M_N^{s_N}\right]$. Then we estimate the contribution of paths with vertices of type 1 at most, which give the leading term in the asymptotic expansion of $\mathbb{E}\left[\mathrm{Tr} M_N^{s_N}\right]$, as long as $s_N \ll \sqrt{N}$.

Denote by $Z(k, (\tilde{n}, \tilde{p}))$ the contribution of paths with k odd marked instants and of type $(\tilde{n}, \tilde{p})$. Using Proposition 2.3 and (13), we deduce that

$$
\begin{aligned}
&Z(k,(\tilde{n},\tilde{p}))\\
&\quad\leq \mathbf{N}(s_N,k)\frac{2}{N^{s_N}}N\frac{p!}{p_o!}\frac{N!}{n_o!}\frac{k!}{p_1!}\frac{(s_N-k)!}{n_1!}\prod_{i=2}^{s_N}\frac{\left(\tilde{C}i\right)^{in_i+ip_i}}{p_i!n_i!(i!)^{p_i}(i!)^{n_i}}\\
&\quad\leq \mathbf{N}(s_N,k)N2\prod_{i=2}^{s_N}\frac{(Ck)^{ip_i}\;(C(s_N-k))^{in_i}}{p_i!n_i!}\;\frac{N^{\sum_{i\geq 1}n_i}\;p^{\sum_{i\geq 1}p_i}}{N^{\sum_{i\geq 1}in_i+ip_i}}\\
&\quad\leq \mathbf{N}(s_N,k)N2\gamma_N^k\prod_{i\geq 2}\frac{1}{n_i!}\left(\frac{C^i(s_N-k)^i}{N^{i-1}}\right)^{n_i}\prod_{i\geq 2}\frac{1}{p_i!}\left(\frac{C^ik^i}{p^{i-1}}\right)^{p_i}, \qquad (15)
\end{aligned}
$$

where in (15) we use that $\gamma_N^{\sum_{i\geq 1}p_i}=\gamma_N^{k-\sum_{i\geq 2}(i-1)p_i}$ and $C>0$ is a constant independent of N, p, s_N and k (whose value may change from line to line).

We denote by Z_2 the contribution of paths for which $\sum_{i\geq 2}n_i+p_i>0$. By Proposition 2.3, and using summation, one has that

$$
\frac{Z_2}{2N}\leq\sum_{k=1}^{s_N}\mathbf{N}(s_N,k)\gamma_N^k\sum_{\tilde{M}_1,\tilde{M}_2:\tilde{M}_1+\tilde{M}_2>0}\frac{1}{\tilde{M}_1!\tilde{M}_2!}\left(\frac{2Cs_N^2}{N}\right)^{\tilde{M}_1}\left(\frac{2Cs_N^2}{p}\right)^{\tilde{M}_2},
$$

where $\tilde{M}_1=\sum_{i\geq 2}n_i$ and $\tilde{M}_2=\sum_{i\geq 2}p_i$. Thus, it is straightforward to see that there exists some constant $B>0$ independent of N such that

$$
\frac{Z_2}{N}\leq B\frac{s_N^2}{N}\times\sum_{k=1}^{s_N}\gamma_N^k\mathbf{N}(s_N,k). \qquad (16)
$$

From this, we can deduce that $Z_2/N=o((1+\sqrt{\gamma_N})^{2s_N})$.

We now show that only the paths with vertices of type 1 at most (the origin being unmarked) have to be taken into account. In this case, once the vertices occurring in the path have been chosen, there is no choice for filling in the blanks of the path. Furthermore, each edge is passed only twice in the path $\mathcal{P}_E$, once at an odd instant and once at an even instant. Thus, denoting by $Z_1:=\sum_{k=1}^{s_N}Z_1(k)$ the contribution of such paths, one has that

$$
Z_1=\sum_{k=1}^{s_N}N\mathbf{N}(s_N,k)\gamma_N^k\prod_{i=1}^{s_N}\frac{N-i}{N}=(1+o(1))\sum_{k=1}^{s_N}N\mathbf{N}(s_N,k)\gamma_N^k.
$$

Using (16), this finishes the proof that $Z_2=o(1)Z_1$. The contribution of paths with marked origin and vertices of type 1 at most is of order Z_1s_N/N and is thus negligible. This finishes the proof of Proposition 2.4. □

Remark 2.4 Set $\hat{k}=\left[\frac{\sqrt{\gamma_N}}{1+\sqrt{\gamma_N}}s_N\right]+1$. For any sequence $s_N\ll 1$, one can show that $\sum_{1\leq k\leq s_N}\frac{N}{s_N}C_{s_N}^kC_{s_N}^{k-1}\gamma_N^k=O(u_+^{s_N})$ and that the main contribution to the expectation

(4) should come from paths with $\hat{k}(1+o(1))$ odd marked instants. Indeed, using Stirling's formula, one has that

$$\max_{1\leq k\leq s_N}\left(C^k_{s_N}\right)^2\gamma_N^k\sim\left(C^{\hat{k}}_{s_N}\right)^2\gamma_N^{\hat{k}}\sim\left(1+\sqrt{\gamma_N}\right)^{2s_N}\frac{(1+\sqrt{\gamma_N})^2}{s_N\sqrt{\gamma_N}}\frac{1}{2\pi}.$$

It is also easy to check that, for any $l>0$, one has that

$$\mathbf{N}(s_N,\hat{k}+l)\gamma_N^{\hat{k}+l}\leq\mathbf{N}(s_N,\hat{k})\gamma_N^{\hat{k}}\exp\left\{-\frac{C_\gamma l^2}{(s_N-\hat{k})}\right\},\tag{17}$$

for some constant C_γ depending on γ only. In the case where $l<0$, we fix some $\Delta>0$ large. Then one can show that, for any $-\Delta(s_N-\hat{k}+1)<l<0$, $\frac{\mathbf{N}(s_N,\hat{k}-1+l)\gamma_N^{\hat{k}-1+l}}{\mathbf{N}(s_N,\hat{k}-1)\gamma_N^{\hat{k}-1}}\leq\exp\left\{-\frac{l^2}{(2\Delta+2)(s_N-\hat{k}+1)}\right\}$. One can also check that for any $l\leq-\Delta(s_N-\hat{k}+1)$, $\frac{\mathbf{N}(s_N,\hat{k}-1+l)\gamma_N^{\hat{k}-1+l}}{\mathbf{N}(s_N,\hat{k}-1)\gamma_N^{\hat{k}-1}}\leq e^{-\Delta(s_N-\hat{k})/3}$. We thus find that $\sum_{1\leq k\leq s_N}\frac{N}{s_N}C^k_{s_N}C^{k-1}_{s_N}\gamma_N^k=O(u_+^{s_N})$, yielding Remark 2.4.

Remark 2.5 In the case of polynomial tails, that is if $P(|X_{ij}|\geq x)\leq(1+x)^{m_o}$ for some $m_o>36$, (15) has to be multiplied by $\prod_{i\geq m_o/4}\Gamma_N^{4i(n_i+p_i)}$. If $\epsilon_o<\frac{m_o-36}{12m_o}$ and $s_N=O(N^{2/3})$, (which is the largest scale considered in this paper), this has no impact on the computations as $\frac{(s_N\Gamma_N^4)^i}{N^{i-1}}\ll 1$ for any $i\geq m_o/4$. All the results stated in the following can be proved in the case of polynomial tails up to minor technical modifications (which amounts essentially to considering apart vertices of type at least $m_o/4$).

2.5 A central limit theorem

The main result of this section is the following Proposition. Let u_+ be defined by (5). We show that all the moments of $\mathrm{Tr}(M_N/u_+)^{s_N}$ are bounded and universal, as long as $1\ll s_N\ll\sqrt{N}$. Assume that $\lim_{N\to\infty}\gamma_N=\gamma<\infty$ and set $l_\beta=1/(\beta\pi)$ where $\beta=1$ (resp. $\beta=2$) in the case where M_N is real (resp. complex).

Proposition 2.5 *Assume that* $1\ll s_N\ll\sqrt{N}$ *and set* $\tilde{M}_N=\frac{M_N}{u_+}$. *Then, there exists* $D>0$ *such that* $Var\left(\mathrm{Tr}\tilde{M}_N^{s_N}\right)\leq D$, *for any* N, *and* $\lim_{N\to\infty}Var\left(\mathrm{Tr}\tilde{M}_N^{s_N}\right)=l_\beta$. *Similarly, for any integer* k,

$$\mathbb{E}\left[\mathrm{Tr}\tilde{M}_N^{s_N}-\mathbb{E}[\mathrm{Tr}\tilde{M}_N^{s_N}]\right]^{2k}=(2k-1)!!\,l_\beta^k(1+o(1)),$$
$$\mathbb{E}\left[\mathrm{Tr}\tilde{M}_N^{s_N}-\mathbb{E}[\mathrm{Tr}\tilde{M}_N^{s_N}]\right]^{2k+1}=o(1).$$

Remark 2.6 In [17], a central limit theorem (CLT) is also established for traces of fixed (independent of N) moments of M_N. In this case, the limiting Gaussian distribution does depend on the fourth moment of the law of the entries. The above CLT is also stated in Remark 6 of [28] (a factor $1/\beta$ is missing) in the case where $\gamma = 1$.

Proof of Proposition 2.5 We only give the proof for the variance. The proof of the asymptotics for higher moments is a rewriting of pp. 128–129 in [29] (see also [9] Sect. 6) and is skipped. In the following, $C_1, \ldots, C_6, C_2', C_3'$ denote some positive constants independent of N. One has that

$$\mathrm{Var}\,(\mathrm{Tr}\, M_N^{s_N}) = \frac{1}{N^{2s_N}} \sum_{\mathcal{P}_E, \mathcal{P}_E'} \Bigg[\mathbb{E}\Bigg(\prod_{e_i \in \mathcal{P}_E} \prod_{e_i' \in \mathcal{P}_E'} \hat{X}_{e_i} \hat{X}_{e_i'} \Bigg) - \mathbb{E}\Bigg(\prod_{e_i \in \mathcal{P}_E} \hat{X}_{e_i} \Bigg) \mathbb{E}\Bigg(\prod_{e_i' \in \mathcal{P}_E'} \hat{X}_{e_i'} \Bigg) \Bigg].$$

Here, given an edge $e = (v_1, v_2)$, $\hat{X}_e$ stands for $X_{v_1 v_2}$ if e occurs at an odd instant of $\mathcal{P}_E$ or for $\overline{X_{v_2 v_1}}$ if it occurs at an even instant. Now the non zero terms in the above sum come from pairs of paths $\mathcal{P}_E$, $\mathcal{P}_E'$ sharing at least one oriented edge and such that each edge appears an even number of times in the union of the two paths. We say that such paths are correlated. To estimate the number of correlated paths and their contribution to the variance, we use the *construction procedure* defined in Sect. 3 of [28]. This construction associates a path of length $4s_N - 2$ to a pair of correlated paths. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two correlated paths of length $2s_N$. When reading the edges of $\mathcal{P}_1$, let e denote the first oriented edge common to the two paths. Let also t_e and t_e' be the instants of the first occurrence of this edge in $\mathcal{P}_1$ and $\mathcal{P}_2$. Then we are going to glue the two paths $\mathcal{P}_1$ and $\mathcal{P}_2$, in such a way that we erase the two first occurrences of e in each of these paths. The glued path, denoted $\mathcal{P}_1 \vee \mathcal{P}_2$, is obtained as follows. We first read $\mathcal{P}_1$ until we meet the left endpoint of e at the instant t_e. Then we switch to $\mathcal{P}_2$ as follows. Assume first that t_e and t_e' are of the same parity. We then read the path $\mathcal{P}_2$, starting from t_e', in the reverse direction to the origin and restart from the end of $\mathcal{P}_2$ until we come back to the instant $t_e' + 1$. If t_e and t_e' are not of the same parity, we read the edges of $\mathcal{P}_2$ in the usual direction starting from $t_e' + 1$ and until we come back to the instant t_e'. We have then read all the edges of $\mathcal{P}_2$ except the edge e occurring between t_e' and t_{e+1}'. We then read the end of $\mathcal{P}_1$, starting from $t_e + 1$. Having done so, we obtain a path $\mathcal{P}_1 \vee \mathcal{P}_2$ which is of length $4s_N - 2$. One can also note that the Dyck path x associated to $\mathcal{P}_1 \vee \mathcal{P}_2$ does not descent lower than the level $x(t_e)$ during the time interval $[t_e, t_e + 2s_N - 1]$, by the definition of e and t_e.

Now, to reconstruct the paths $\mathcal{P}_1$ and $\mathcal{P}_2$ from $\mathcal{P}_1 \vee \mathcal{P}_2$, it is enough to determine the instant at which one has switched from one path to the other, the origin of the path $\mathcal{P}_2$ and the direction in which $\mathcal{P}_2$ is read. There are at most $4s_N$ ways to determine the origin and the direction once the instant of switch is known. To estimate the number of preimages of a given path $\mathcal{P}_1 \vee \mathcal{P}_2$ of length $4s_N - 2$ and with k odd up steps, one has to give an upper bound for the number of instants t_e in $\mathcal{P}_1 \vee \mathcal{P}_2$, which can be the instants of switch. To this aim, fix some $t_e \in [0, 2s_N - 1]$ and assume that the Dyck path of $\mathcal{P}_1 \vee \mathcal{P}_2$ does not go below the level $x(t_e)$ during an interval of time of length greater than or equal to $2s_N - 1$. Assume that $x(t_e) > 0$. Set then

$$l = \inf\{t \geq t_e, x(t) = x(t_e), x(t+1) = x(t_e) - 1\} - 2s_N + 1.$$

Denote by T_2 the sub-trajectory in the interval $[t_e, t_e + 2s_N - 1 + l]$. It is a Dyck path. Denote also by T_1 the remaining part of the trajectory: it is also a Dyck path, along which the instant t_e has been chosen. We denote by k_1 the number of the odd up steps of T_1. As the Dyck path of $\mathcal{P}_1 \vee \mathcal{P}_2$ is obtained by inserting T_2 at the instant t_e in T_1, and using the fact that $\mathcal{P}_1 \vee \mathcal{P}_2$ and $\mathcal{P}_1 \cup \mathcal{P}_2$ have all the same edges but one, one can then deduce (see [28], pp. 11–13, for the detail) that the contribution of correlated pairs is at most of order

$$\sum_{k=1}^{2s_N-1} \sum_{l=0}^{2s_N-1} \sum_{k_1 \leq k \wedge 2s_N - 1 - l} \frac{\mathbf{N}\left(s_N - \frac{(1+l)}{2}, k_1\right) \mathbf{N}\left(s_N + \frac{l-1}{2}, k - k_1\right)}{\mathbf{N}(2s_N - 1, k)}$$
$$\times (2s_N - 1 - l) \frac{4s_N}{N} Z_1(4s_N - 2, k) \tag{18}$$
$$+ \sum_{k=1}^{2s_N-1} \sum_{l=0}^{2s_N-1} \sum_{k_1 \leq k \wedge 2s_N - 1 - l} \frac{\mathbf{N}\left(s_N - \frac{(1+l)}{2}, k_1\right) \mathbf{N}\left(s_N + \frac{l-1}{2}, s_N + \frac{l-1}{2} + k_1 - k\right)}{\mathbf{N}(2s_N - 1, k)}$$
$$\times (2s_N - 1 - l) \frac{4s_N}{N} Z_1(4s_N - 2, k). \tag{19}$$

Here $Z_1(4s_N - 2, k)$ is the contribution of paths of length $4s_N - 2$ with k odd up steps to the expectation $\mathbb{E}[\mathrm{Tr} M_N^{2s_N-1}]$ and (18) [resp. (19)] corresponds to the case where t_e is even (resp. odd). The term $(2s_N - l - 1)$ in (18) comes from the determination of t_e and where $1/N = \mathbb{E}(|X_e|^2)/N$, if e is the edge erased from $\mathcal{P}_1 \vee \mathcal{P}_2$. It can indeed be shown that paths for which such an edge occurs also in $\mathcal{P}_1 \vee \mathcal{P}_2$ yield a contribution of order s_N/N that of typical paths and are thus negligible.

We first show that the variance is bounded. In the following, we set $s_1(l) = s_N - \frac{1+l}{2}$ and $s_2(l) = s_N + \frac{l-1}{2}$. Considering for instance (18), [(19) is similar], it is enough to prove that there exists a constant $C_1 > 0$ such that

$$\sum_{l=0}^{2s_N-1} \sum_{k_1 \leq k \wedge 2s_N - 1 - l} \frac{\mathbf{N}(s_1(l), k_1)\mathbf{N}(s_2(l), k - k_1)}{\mathbf{N}(2s_N - 1, k)} (2s_N - 1 - l) \leq C_1 \sqrt{s_N}.$$

One can easily see that it is enough to consider the case where $2s_N - 1 - l \geq \sqrt{s_N}$. It is also straightforward by Remark 2.4 and Proposition 2.4 to see that one can choose $0 < 2\beta' < \frac{\sqrt{\gamma}}{1+\sqrt{\gamma}} < 2\beta'' < 2$ such that

$$\sum_{k \leq \beta' s_N \text{ or } k \geq \beta'' s_N} Z_1(4s_N - 2, k) \ll s_N^{-5} N u_+^{2s_N - 1}.$$

This is enough to ensure that the contribution of correlated pairs such that the corresponding glued path has k odd up steps for some $k \leq \beta' s_N$ or $k \geq \beta'' s_N$ is negligible in the large-N-limit. We now set

$$f(k_1) := \frac{\mathbf{N}(s_1(l), k_1)\mathbf{N}(s_2(l), k-k_1)}{\mathbf{N}(2s_N - 1, k)}. \tag{20}$$

Then, l and k being fixed, f is maximal at $\tilde{k}_1 = [k\frac{2s_N-1-l}{4s_N-2}](+1)$. Furthermore, one can check that there exist constants $C_2, C_2' > 0$ such that $f(\tilde{k}_1 + j) \leq C_2 \exp\left\{-C_2' j^2/\tilde{k}_1\right\}$ for any j. From this we deduce that

$$\sum_{k_1 \leq k \wedge 2s_N - 1 - l} \frac{\mathbf{N}(s_1(l), k_1)\mathbf{N}(s_2(l), k-k_1)}{\mathbf{N}(2s_N - 1, k)}(2s_N - 1 - l)$$
$$\leq C_3'(2s_N - 1 - l)^{3/2}\frac{\mathbf{N}(s_1(l), \tilde{k}_1)\mathbf{N}(s_2(l), k-\tilde{k}_1)}{\mathbf{N}(2s_N - 1, k)}. \tag{21}$$

It is now an easy consequence of Stirling's formula that

$$\sum_{l=0}^{2s_N-1} C_3(2s_N - 1 - l)^{3/2}\frac{\mathbf{N}(s_1(l), \tilde{k}_1)\mathbf{N}(s_2(l), k-\tilde{k}_1)}{\mathbf{N}(2s_N - 1, k)} \leq C_4 \frac{\sqrt{s_N}}{(1-\alpha_N)^2}, \tag{22}$$

where $\alpha_N = k/(2s_N - 1)$. Using Proposition 2.4, one can also show that there exists $C_5 > 0$ such that

$$\sum_{\beta' s_N \leq k \leq \beta'' s_N} \frac{1}{\alpha_N(1-\alpha_N)} Z_1(4s_N - 2, k) \sim C_5 s_N^{-3/2} N u_+^{2s_N - 1}.$$

Combining the whole yields that there exists $C_6 > 0$ such that (18)+(19)$\leq C_6 u_+^{2s_N}$. In the case where $x(t_e) = 0$, t_e is chosen amongst the returns to the level 0 of the Dyck path. It can be shown that the number of such instants is negligible with respect to $\sqrt{s_N}$ in typical paths. This follows from arguments already used above and in [28] p. 13.

To compute the variance, we notice that in (18), the term $(2s_N - 1 - l)$ can actually be replaced with $s_1(l) - k_1$. Indeed as t_e is even, the first step after $t_e + 2s_N - 1 + l$ is a down step occurring at an odd instant. Also, there are only s_N choices for the origin of $\mathcal{P}_2$, since one knows the parity of e in $\mathcal{P}_2$ once the orientation of $\mathcal{P}_2$ is fixed. Then, using (18), (19), Remark 2.4, the exponential decay of $f(k_1)$, and Proposition 2.4, one can deduce that (for the real case)

$$\lim_{N\to\infty} \mathrm{Var}\,\mathrm{Tr}\tilde{M}_N^{s_N} = \lim_{N\to\infty}\left(1+\frac{1}{\sqrt{\gamma}}\right)\frac{2s_N}{1+\sqrt{\gamma}}\sum_{l\leq 2s_N-1}\frac{s_1(l)}{(1+\sqrt{\gamma})^{4s_N}}$$

$$\times \sum_{k\geq 1}\sum_{k_1\leq k}\gamma_N^k \mathbf{N}(s_1(l),k_1)\mathbf{N}(s_2(l),k-k_1) \tag{23}$$

$$= \lim_{N\to\infty}\frac{2s_N}{\sqrt{\gamma}}\sum_{l\leq 2s_N-1}\frac{s_1(l)}{(1+\sqrt{\gamma})^{4s_N}}E_{s_1(l)}E_{s_2(l)} \tag{24}$$

$$= l_1.$$

In (24), we have set $E_k = \int x^k \frac{\sqrt{((1+\sqrt{\gamma})^2-x)(x-(1-\sqrt{\gamma})^2)}}{2\pi x}dx$, and the equality follows from the fact that (23) is a Cauchy product. The value of l_1 can be deduced from Formulas 4.7 in [31] and 3.6 in [28]. The computation of l_2 follows from the fact that, in the complex case, the occurrences e in $\mathcal{P}_1$ and $\mathcal{P}_2$ cannot have the same parity (in typical paths). □

3 The case where $\gamma_N \to \gamma$, $1 \leq \gamma < \infty$

The aim of this section is to prove the following universality result. Let $c_1, \dots, c_k$ be positive real numbers and $s_N^{(i)}$, $i{=}1,\dots,K$, be sequences such that $\lim_{N\to\infty}\frac{s_N^{(i)}}{N^{2/3}}{=}c_i$.

Theorem 3.1 *Assume that* $M_N = \frac{1}{N}XX^*$ *satisfies* (i) *to* (iv) *(resp.* (i') *to* (iv')*). Formulas* (6) *and* (7) *hold true.*

The proof of Theorem 3.1 is the object of this section. We actually focus on the case where $K=1$. The proof of Theorem 3.1 for $K>1$ is a rewriting of the arguments used in [30] (p. 41), Sect. 2.5 and of those used in the case where $K=1$. It is not developed further here. Then, we essentially show that typical paths (i.e. those having a non negligible contribution to the expectation) have no oriented edge read more than twice. This ensures that the expectation (4) only depends on the variance of the entries X_{ij}, $1\leq i\leq N$, $1\leq j\leq p$. Universality of the expectation then follows.

3.1 Number of self-intersections and odd marked instants in typical paths

We first give a technical Proposition which bounds the number of self intersections and gives the approximate number of odd marked instants in typical paths. In the following, we denote by $Z(k)$ the contribution of paths with k odd marked instants. We also denote the number of self-intersections on each line by $M_1 = \sum_{i\geq 2}(i-1)n_i$ and $M_2 = \sum_{i\geq 2}(i-1)p_i$.

Proposition 3.1 *There exists a positive constant* d_1 *such that the contribution of paths for which* $M_1+M_2\geq d_1\sqrt{s_N}$ *is negligible in the large-*N*-limit, whatever* $1\leq k\leq s_N$ *is.*
And for any α, α' *such that* $0<\alpha'<\frac{\sqrt{\gamma}}{1+\sqrt{\gamma}}<\alpha<1$*, one has that*

$$\sum_{k\leq \alpha' s_N} Z(k) + \sum_{k\geq \alpha s_N} Z(k) = o(1)u_+^{s_N}.$$

Proof of Proposition 3.1 We first give the proof of the first point of Proposition 3.1. Denote by $Z(k,(\tilde{n},\tilde{p}))$ the contribution of paths with k odd marked instants and of type $(\tilde{n},\tilde{p})$. Using (15), Remark 2.4 (and exactly the same arguments as in [30] p. 34), one can see that, for d_1 large enough,

$$\sum_{k=1}^{s_N}\sum_{(\tilde{n},\tilde{p})/\sum_{i\geq 2}(i-1)(n_i+p_i)\geq d_1\sqrt{s_N}} Z(k,(\tilde{n},\tilde{p}))=o(1)u_+^{s_N}.$$

We now turn to the second statement. Let then α and α' be chosen as in Proposition 3.1. We assume that N is large enough so that $\alpha'<\frac{\sqrt{\gamma_N}}{1+\sqrt{\gamma_N}}<\alpha$. We now show that $\sum_{k\geq\alpha s_N}Z(k)\ll u_+^{s_N}$. Given any integer $k\leq s_N$, and using (15), one can show that there exists a constant $C_8>0$ independent of N and k such that $Z(k)\leq N\gamma_N^k\exp\left\{C_8N^{1/3}\right\}\mathbf{N}(s_N,k)$. Thus

$$\begin{aligned}\sum_{k\geq\alpha s_N}Z(k)&\leq\sum_{k\geq\alpha s_N}\mathbf{N}(s_N,k)\gamma_N^kN\exp\left\{C_8N^{1/3}\right\}\\&\leq NC_{s_N}^{\hat{k}}C_{s_N}^{\hat{k}-1}\gamma_N^{\hat{k}}\exp\left\{C_8N^{1/3}-C_7s_N\left(\alpha-\frac{\sqrt{\gamma_N}}{(1+\sqrt{\gamma_N})}\right)^2\right\}\ll u_+^{s_N},\end{aligned}$$

for N large enough. Similarly, one can show that the contribution of paths for which $k\leq\alpha's_N$ is negligible in the large-N-limit. □

3.2 Asymptotics of $\mathbb{E}[\mathrm{Tr}M_N^{s_N}]$

In this section, we refine the estimate (15) and in particular deal with vertices of type 2. Indeed, when summing (15) over $n_i, i\leq s_N$ and $p_i, i\leq s_N$, one can note that terms associated to vertices of type 2 make the summation go to infinity. To this aim, we shall control the number of vertices for which there is an ambiguity to continue the path at an unmarked instant. We shall also control the number of such vertices associated to edges passed four times or more. Finally, we shall also show that amongst vertices of type 3, none belongs to edges passed more than twice, while there are no more complex self-intersections in typical paths.

From now on, given $\alpha's_N\leq k\leq\alpha s_N$, we consider paths of type $(\tilde{n},\tilde{p})$ with $M_1=\sum_{i\geq 2}(i-1)n_i\leq d_1\sqrt{s_N}$ (resp. $M_2=\sum_{i\geq 2}(i-1)p_i\leq d_1\sqrt{s_N}$) self-intersections on the bottom (resp. top line). Our counting strategy is refined as follows. We first choose the moments of self-intersection and the vertices occurring at the remaining marked moments and the origin. One fills in the blanks of the path until the first instant of self-intersection. At that moment, one chooses the vertex to be repeated among the preceding ones (and repeat it if needed at the moment of second self-intersection and so on). We then proceed in the same way for subsequent vertices.

Let us choose the instants of self-intersection on each line: let $t_{jb,1}<t_{jb,2}<\cdots<t_{jb,n_2}$ (resp. $t_{ju,1}<t_{ju,2}<\cdots<t_{ju,p_2}$) be the instants of self-intersection corresponding to vertices of type 2 on the bottom (resp. top) line, $t_{jb,1}^{3,1}<t_{jb,2}^{3,1}<$

$\cdots < t^{3,1}_{jb,n_3}$ those corresponding to the first repetition of a vertex of type 3 on the bottom line, $t^{3,2}_{jb,1} > t^{3,1}_{jb,1}$, $t^{3,2}_{jb,2} > t^{3,1}_{jb,2} \dots$ for the second repetition of a vertex of type 3 on the bottom line. We do not detail the list of instants since it is the same as in [30] p. 724, except that we make a distinction between instants marked at the odd or even instants (even if it is the same vertex). We then choose the origin and the vertices occurring at the remaining marked instants. The number of pairwise distinct vertices in the order of appearance occurring in the path on the top and bottom lines is

$$N \prod_{i=1}^{s_N-k-M_1} (N-i) \prod_{i=1}^{k-M_2} (p+1-i). \tag{25}$$

Note that (25)$\sim N^{s_N-M_1-M_2+1}\gamma_N^{k-M_2} \exp\left\{-\frac{(s_N-k)^2}{2N} - \frac{k^2}{2p}\right\}$.

We now turn to the determination of vertices occurring at the instants of self-intersection. First, we focus on vertices of type 2. In the general case, there are $ju, i - i = O(s_N)$ choices for the vertex occurring at the instant $t_{ju,i}$, since one chooses vertices occurring twice as marked instants. We first consider the simplest case where the path is such that there is no choice for closing any edge from vertices of type 2 at unmarked instants and where no vertex of type 2 belongs to edges passed four times or more. Then the number of possible choices for the instants of self-intersection and the corresponding vertices of type 2 is at most

$$\begin{aligned}
&\sum_{1\le t_{ju,1}<t_{ju,2}<\cdots<t_{ju,p_2}\le s_N-k} \prod_{i=1}^{p_2}(j_{u,i}-i) \sum_{1\le t_{jb,1}<t_{jb,2}<\cdots<t_{jb,n_2}\le k} \prod_{i=1}^{n_2}(j_{b,i}-i)\\
&\quad\le \frac{1}{n_2!}\left(\frac{(s_N-k)^2}{2}\right)^{n_2} \frac{1}{p_2!}\left(\frac{k^2}{2}\right)^{p_2}.
\end{aligned}$$

Such an estimate combined with formula (25) and Remark 2.4 then ensures that the contribution of such paths to $\mathbb{E}[\mathrm{Tr}(M_N/u_+)^{s_N}]$ is bounded.

We now consider the general case. In general, given a vertex v of type 2 (on the bottom or on the top line), there might be multiple ways to close an edge with v as its left endpoint at an unmarked instant. Note that there are at most three possible ways to close the edge. An example of such a vertex is the distinguished vertex 4 in Fig. 1, as the distinguished edge could have been $(4,4)$. Indeed, the two up edges with 4 as a marked vertex on the top line are read before any down edge is closed starting from 4. This leads to the notion of non-closed vertex.

Definition 3.1 A vertex v of type 2 is said to be non-MP-closed if it is an odd (resp. even) marked instant and if there are more than one choice for closing an edge at an unmarked instant starting from this vertex on the top (resp. bottom) line.

Remark 3.1 The definition of non-MP-closed vertices differs from that of non-closed vertices in [30], essentially due to the distinction which is made between the top and bottom lines.

Let t be a given marked instant. Assume that the marked vertices before t have been chosen and that, at the instant t, there is a non-MP-closed vertex. Then, by the definition of the Dyck path x and that of non-MP-closed vertices, there are at most $x(t)$ possible choices for this vertex. This can be checked as in [29], p 122. In Lemma 3.1 below, we show that $\max_t x(t) \sim \sqrt{s_N}$ in typical paths.

Apart from non-MP-closed vertices, a vertex of type 2 can also belong to an edge that is read four times or more in the path. To consider such vertices, we need to introduce other characteristics of the path. Let $\nu_N(P)$ be the maximal number of vertices that can be visited at marked instants from a given vertex of the path P. Let also $T_N(P)$ be the maximal type of a vertex in P. Then, if at the instant t, one reads for the second time an oriented up edge e, there are at most $2(\nu_N(P)+T_N(P))$ choices for the vertex occurring at the instant t. Indeed, one shall look among the oriented edges already encountered in the path and for which one endpoint is the vertex occurring at the instant $t-1$ (see the Appendix in [29] and [9] Sect. 5.1.2 e.g.). It is an easy fact that paths for which $T_N(P) \geq AN^{1/3}(\ln N)^{-1}$ lead to a negligible contribution, if A is large enough (independently of k). Using Lemma 3.2 stated below, we prove at the end of this section that there exists $\epsilon > 0$ small enough such that, for typical paths, $\nu_N(P) \leq s_N^{1/2-\epsilon}$ for any $\alpha' s_N \leq k \leq \alpha s_N$.

For vertices of type $i > 2$, once the $i-1$ moments of self-intersection are fixed, one chooses at the first moment of self-intersection the vertex to be repeated amongst those already occurred in the path.

Assuming the above estimates on $\max x(t)$ and ν_N hold, we consider paths P_k of type $(\tilde{n}, \tilde{p})$ with $M_1 := \sum_{i\geq 2}(i-1)n_i \leq d_1\sqrt{s_N}$ and $M_2 := \sum_{i\geq 2}(i-1)p_i \leq d_1\sqrt{s_N}$ self-intersections respectively on the bottom or on the top line, r_i non-MP-closed vertices of type 2 ($i=1,2$) on the bottom and top lines and q_i ($i=1,2$) vertices of type 2 on the bottom or top line visited at the second marked instant along an oriented edge already seen in the path. Using Lemma 1 in [29], one can check that Proposition 2.3 can now be refined to

$$(\Omega\mathbb{E})_{\max} \leq \frac{2}{N^{s_N}} \prod_{l=3}^{s_N-k} \left(\tilde{C}l\right)^{ln_l} \prod_{m=3}^{k} \left(\tilde{C}m\right)^{mp_m} 3^{r_1+r_2} D_3^{q_1} D_4^{q_2},$$

where D_3, D_4 are positive constants independent of k, p, N, and s_N. Using (25) and the above, the contribution of such paths to $\mathbb{E}[\mathrm{Tr} M_N^{s_N}]$ is then at most of order (see also [30], p. 725)

$$C\mathbf{N}(s_N, k) N \gamma_N^k e^{\left\{-\frac{(s_N-k)^2}{2N} - \frac{k^2}{2p}\right\}} \mathbb{E}_k$$
$$\times \left[\frac{1}{(n_2-r_1-q_1)!} \left(\frac{(s_N-k)^2}{2N}\right)^{n_2-r_1-q_1} \frac{1}{r_1!} \left(\frac{3(s_N-k)\max\, x(t)}{N}\right)^{r_1} \right.$$
$$\times \quad \frac{1}{q_1!} \left(\frac{D_3(s_N-k)(\nu_N+T_N)}{N}\right)^{q_1} \prod_{i\geq 3} \frac{1}{n_i!} \left(\frac{C^i (s_N-k)^i}{N^{i-1}}\right)^{n_i}$$

$$\times \quad \frac{1}{(p_2-r_2-q_2)!}\left(\frac{k^2}{2p}\right)^{p_2-r_2-q_2}\frac{1}{r_2!}\left(\frac{3k\max\ x(t)}{p}\right)^{r_2}$$
$$\times \quad \frac{1}{q_2!}\left(\frac{D_4 k(\nu_N+T_N)}{p}\right)^{q_2}\prod_{i\geq 3}\frac{1}{p_i!}\left(\frac{C^i k^i}{p^{i-1}}\right)^{p_i}\Bigg]. \tag{26}$$

Here $\mathbb{E}_k$ denotes the expectation with respect to the uniform distribution on $\mathcal{X}_{s_N,k}$ and we have used the fact that $\sum_{x\in\mathcal{X}_{s_N,k}} f(x) = \mathbf{N}(s_N,k)\mathbb{E}_k(f(x))$ for any function $f\geq 0$.

Before considering paths in complete generality, we first restrict to paths with less than $d_1\sqrt{s_N}$ self-intersections and no vertex of type strictly greater than 3, that is $\sum_{i\geq 4} p_i+n_i=0$. Let $Z_3(k)$ denote the total contribution of paths with k odd marked instants such that $q:=q_1+q_2=0$, with no oriented edges read more than twice and satisfying the above conditions.

Proposition 3.2 *There exists a constant $B_1>0$ independent of N such that $Z_3 := \sum_{k=\alpha' s_N}^{\alpha s_N} Z_3(k)\leq B_1 u_+^{s_N}$.*

Proof of Proposition 3.2 From (26), we deduce that there exists a constant $D_o>0$ independent of N, p, k and s_N, such that

$$Z_3(k)\leq \mathbf{N}(s_N,k)N\gamma_N^k\mathbb{E}_k\left(\exp\left\{6\frac{\max x(t)s_N}{N}\right\}\right)\exp\left\{D_o\frac{s_N^3}{N^2}\right\}. \tag{27}$$

In Lemma 3.1 proved below, we show that, given $a>0$, there exists $b>0$, independent of N, such that $\mathbb{E}_k\left(e^{\{\frac{a\max x(t)}{\sqrt{s_N}}\}}\right)\leq b$, $\forall \alpha' s_N\leq k\leq \alpha s_N$. This yields that (27)$\leq D_5\mathbf{N}(s_N,k)N\gamma_N^k$, for some constant D_5 independent of k. Remark 2.4 ensures that $Z_3:=\sum_k Z_3(k)=O(u_+^{s_N})$. □

Assuming that there are no self-intersections of type greater than 3, we can then show that paths for which $q=q_1+q_2\geq 1$ give a contribution of order $u_+^{s_N}\nu_N/\sqrt{s_N}=o(u_+^{s_N})$ and thus there are no edges read more than twice (associated to vertices of type 2). We then proceed in the same way to show that there are no more than $\ln\ln N$ vertices of type 3 in typical paths and that there are no oriented edges read more than twice associated to vertices of type 3. It is then easy to deduce from the above result that paths with self-intersections of type 4 or greater, or a marked origin, lead to a contribution of order $u_+^{s_N}s_N/N=o(1)u_+^{s_N}$. The detail is skipped.

Finally, we investigate the total contribution of the paths for which $\nu_N\geq s_N^{1/2-\epsilon}$ where $\epsilon>0$ is fixed (small). Denote by Z_4 such a contribution. We only indicate the tools needed to prove that

$$Z_4:=\sum_{k=\alpha' s_N}^{\alpha s_N} Z_4(k)=o(1)u_+^{s_N},$$

since the detail of the proof is a rewriting of the arguments of the proof of Lemma 7.8 in [9]. To consider such paths, we introduce the following characteristic of the path, namely $N_o := r_1 + r_2 + \sum_{i \geq 3} in_i + ip_i$. Assume then that k, N_o, q_1 and q_2 are given. We can then divide the interval $[0, 2s_N]$ into N_o sub-intervals, so that inside an interval, there are no non-MP-closed vertices of type 2 and no vertices of type at least 3. Then there is no choice for closing the edges inside these sub-intervals. Assume that a vertex v is the starting point of ν_N up edges. Then, there is a time interval $[s_1, s_2]$ during which the Dyck path of P_k comes $\nu'_N := \frac{\nu_N}{2N_o}$ times to the level x_o (of v) and never goes below. Denote by $\Gamma(\nu'_N)$ the event that there exists such an interval in a Dyck path and let $\mathbb{P}_k$ denote the uniform distribution on $\mathcal{X}_{s_N,k}$. In Lemma 3.2, we show that there exist positive constants A_1, A_2 independent of s_N such that

$$\max_{1 \leq k \leq s_N} \mathbb{P}_k \left(\Gamma(\nu'_N) \right) \leq A_1 s_N^2 \exp\left\{ -A_2 \nu'_N \right\}.$$

Using Lemma 3.1, one can also show that there exists a constant $A > 0$ such that $\max x(t) \leq AN^{1/6}\sqrt{s_N}$ in any non-negligible path. Using these estimates and formula (26), one can then copy the arguments used in [9] Lemma 7.8 to deduce that $Z_4 = o(1)u_+^{s_N}$.
This finishes the proof that typical paths have a non-marked origin, vertices of type 3 at most (and less than $\ln \ln N$ of type 3), less than $d_1\sqrt{s_N}$ self-intersections and no edges read more than twice. The proof of Theorem 3.1 is completed once Lemmas 3.1 and 3.2 are proved. □

3.3 Technical Lemmas

In this section, we prove the results used in the previous section on characteristics of typical paths. This will complete the proof of Theorem 3.1. The first quantity of interest here is the maximum level reached by the Dyck path x. We shall show that it roughly behaves as $\sqrt{s_N}$ in typical paths. The second one is the maximal number of vertices visited from a given vertex, $\nu_N(P)$, which should not grow faster than s_N^δ for any power δ (but we get a weaker bound).

We shall now prove the announced estimate for $\max_{t \in [0, 2s_N]} x(t)$, where $x(t)$ denotes the level at time t of the Dyck path x associated to a path $P = P_k$. Let $a > 0$ be some constant independent of N and k and denote by $\mathbb{E}_k$ the expectation with respect to the uniform distribution $\mathbb{P}_k$ on $\mathcal{X}_{s_N,k}$.

Lemma 3.1 *There exists $b = b(a) > 0$ independent of s_N and N such that*

$$\max_{\alpha' s_N \leq k \leq \alpha s_N} \mathbb{E}_k \exp\left\{ \frac{a \max x(t)}{\sqrt{s_N}} \right\} \leq b.$$

Proof of Lemma 3.1 It is proved in [29] that the above result holds true if one replaces $\mathbb{E}_k$ with the expectation with respect to the uniform distribution on Dyck paths (no constraint on k) of length $2s_N$. We will call on this result to prove Lemma 3.1. To this aim, we cut the Dyck path x into 2-steps, so that there are 4 types of basic 2-steps :

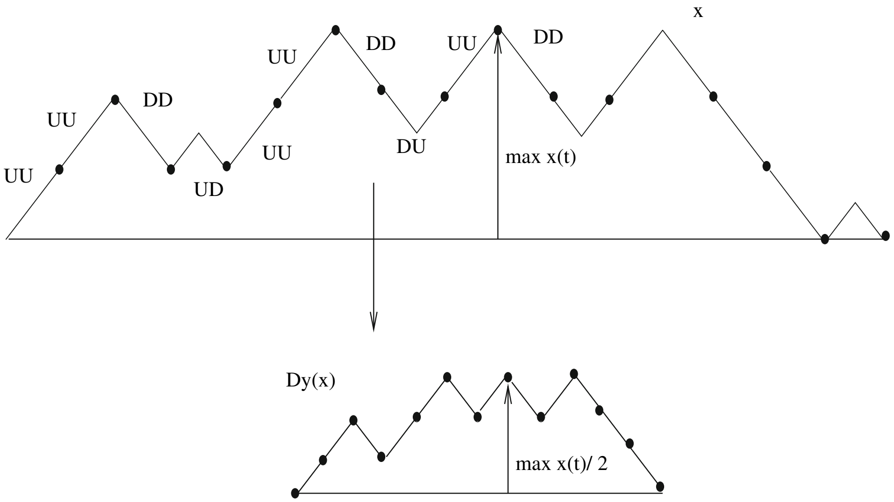

Fig. 2 A Dyck path x and the associated trajectory $Dy(x)$

UU, UD, DD and DU (D stands here for down, U for up). It is an easy fact that the number of UU steps equals that of DD steps. Let then l be the number of UU steps (and DD steps), k_2 be those of DU and k_3 be those of UD steps. Then,

$$2l + k_3 + k_2 = s_N, \quad l + k_2 = s_N - k, \quad l + k_3 = k. \tag{28}$$

As a step UD or DU brings the path to the same level, it is easy to see that the steps UU and DD are arranged in such a way that they form a Dyck path (if we identify a UU step with an up step and a DD step with a down step) of length $2l$. We denote by $Dy(x)$ this sub-Dyck path associated to the Dyck path x (see Fig. 2).

We now explain how to build a Dyck path x given $Dy(x)$ of length $2l$ and $s_N - k - l$ (resp. $k - l$) DU (resp. UD) steps. To construct x from $Dy(x)$, one has to insert "horizontal" steps, namely DU and UD steps, in a particular way. Note that two distinct insertions lead to two different trajectories. The sole constraint is to insert steps DU when the path $Dy(x)$ is at a level greater than or equal to one. This is the reason why we enumerate Dyck paths with $2l$ steps according to the number of times they come back to the level 0. Call $\sharp Dyck(l, Q)$ the number of Dyck paths with $2l$ steps and Q returns to 0. We then have to insert $s - k - l$ horizontal DU steps into $2l - Q$ boxes. Then we can insert the UD steps arbitrarily. This yields that

$$\mathbf{N}(s_N, k) = \sum_{l=0}^{k \wedge (s_N - k)} \sum_{Q=0}^{l} \sharp Dyck(l, Q) C_{l-Q+s-k-1}^{s-k-l} C_s^{k-l}. \tag{29}$$

From this construction, it is easy to see that the maximum level reached by the Dyck path x is twice the maximum $(+1)$ reached by the sub-path $Dy(x)$. Let then $\mathcal{Y}_{l,Q}$

denote the set of Dyck paths of length $2l$ with Q returns to 0. We do not consider the degenerate case where $Dy(x) = \emptyset$, which corresponds to the Dyck path obtained with UD steps only. Then one has that

$$\begin{aligned}
&\mathbb{P}_k(\max x(t) = r) \\
&\leq \sum_{l=1}^{s_N/2} \sum_{Q=0}^{s_N} \mathbb{P}_k\left(\max Dy(x) = r/2 \,\middle|\, Dy(x) \in \mathcal{Y}_{l,Q}\right) \mathbb{P}_k\left(Dy(x) \in \mathcal{Y}_{l,Q}\right) \\
&\quad + \sum_{l=1}^{s_N/2} \sum_{Q=0}^{s_N} \mathbb{P}_k\left(\max Dy(x) = \frac{r-1}{2} \,\middle|\, Dy(x) \in \mathcal{Y}_{l,Q}\right) \mathbb{P}_k\left(Dy(x) \in \mathcal{Y}_{l,Q}\right).
\end{aligned} \tag{30}$$

Let $\mathbb{P}_{l,Q}$ denote the uniform distribution on $\mathcal{Y}_{l,Q}$. Let also $a_o > 0$ (small) be given. It can easily be inferred from [29], p. 11 (see also [30] and [9], Lemma 7.10) that there exist positive constants a_1, a_2, independent of l and Q such that, if $r' \geq a_o\sqrt{l}$, one has that

$$\mathbb{P}_{l,Q}(\max x(t) = r') \leq \frac{a_1}{\sqrt{l}} \exp\left\{-\frac{a_2 r'^2}{l}\right\}. \tag{31}$$

Thus, inserting (31) in (30), we deduce that there exist some positive constants a_3, a_4 independent of k, s_N and N such that, provided $r \geq a_o\sqrt{s_N}$ and for any $\alpha' s_N \leq k \leq \alpha s_N$, $\mathbb{P}_k(\max x(t) = r) \leq \frac{a_3}{\sqrt{s_N}} \exp\left\{-\frac{a_4 r^2}{s_N}\right\}$. This yields Lemma 3.1. □

The second estimate we need is a suitable bound on $\nu_N(P_k)$. Recall that $\Gamma(\nu'_N)$ denotes the event that a Dyck path x comes back from above ν'_N times to some level x_o.

Lemma 3.2 *There exist positive constants A_1, A_2 independent of k, N and p such that*

$$\max_{1 \leq k \leq s_N} \mathbb{P}_k\left(\Gamma(\nu'_N)\right) \leq A_1 s_N^2 \exp\left\{-A_2 \nu'_N\right\}. \tag{32}$$

Proof of Lemma 3.2 Let $[s_1, s_2]$ be an interval such that $x(t_1) = x(t_2) = x_o$ for some $x_o \geq 0$ and $x(t) \geq x_o, \forall t \in [s_1, s_2]$ and for which there exists $s_1 < t_1 < t_2 < \cdots < t_{\nu'_N} \leq s_2$ such that $x(t_i) = x_o$. We first consider the case where s_1 and s_2 are even instants (then x_o is also even). Modifications to be made in the case where they are odd will be indicated at the end of the proof. The instants t_i are then called instants of returns from above to x_o of the Dyck path x. Set now Y_o to be the Dyck path of length $s_2 - s_1$ defined by $y_o(t) = x(t + s_1) - x_o, t \in [0, s_2 - s_1]$. Then the returns from above to x_o correspond to returns to 0 of Y_o. Now, the returns to 0 of Y_o can either be made using UD steps or correspond to a return of the sub-trajectory $Dy(Y_o)$ to this level. Thus, either the number of UD steps is large or the number of returns of $Dy(Y_o)$ to level 0 is large. We shall show that in both cases, (32) holds. Thanks to Proposition 3.1, it is enough to consider trajectories x for which $\alpha' s_N \leq k \leq \alpha s_N$. The proof of Lemma 3.2 is divided into three steps.

Step 1 We first show that there exist positive constants C_7', C_8' independent of N and k such that, provided $\alpha' s_N \leq k \leq \alpha s_N$,

$$\mathbb{P}_k\,(\text{ } x \text{ has } \eta_N \text{ consecutive } UD \text{ steps }) \leq C_7' s_N^2 \exp\left\{-C_8' \eta_N\right\}. \tag{33}$$

Assume that there exists a time interval $[s_1', s_2']$ with η_N consecutive UD steps only. Given even instants s_1' and s_2' (with $s_2' - s_1' = 2\eta_N$), the proportion of trajectories x that have η_N steps UD in $[s_1', s_2']$ is at most

$$\frac{\frac{1}{s_N - \eta_N}\left(C_{s_N-\eta_N}^{k-\eta_N}\right)^2}{\frac{1}{s_N}\left(C_{s_N}^{k}\right)^2} \leq C_9 \left(\frac{k(k-1)\cdots(k-\eta_N+1)}{s_N(s_N-1)\cdots(s_N-\eta_N+1)}\right)^2 \leq C_9 \alpha^{2\eta_N}, \tag{34}$$

for some constant $C_9 > 0$. This readily yields (33).

Step 2 We consider the case where the number of returns to level 0 made by the associated path $Dy(Y_o)$ is large. Let $\mathbb{P}_s$ denote the uniform distribution on the set of Dyck paths Y with length $2s$. It is proved in [29] that there exist constants C_{10}, C_{11} independent of s such that

$$\begin{aligned}&\mathbb{P}_s(\exists s_1', s_2' : Y \text{ has } \eta_N \text{ returns from above to the level } x_o \text{ in } [s_1', s_2'])\\ &\quad \leq C_{10} s^2 \exp\{-C_{11}\eta_N\}.\end{aligned} \tag{35}$$

Denote by Q the number of sub-Dyck paths $\tilde{Y}_i, i = 1, \ldots Q$, of $Dy(Y_o)$ starting and ending at level 0. From the above result and (29), one can deduce that there exists constants $C_{12}, C_{13} > 0$, independent of s_N and k, such that

$$P(Q = \eta_N) \leq C_{12} s_N^2 \exp\{-C_{13}\eta_N\}. \tag{36}$$

Step 3 We can now turn to the proof of Lemma 3.2. A Dyck path x coming back ν_N' times to the level x_o during $[s_1, s_2]$ can be described as follows. Denote by Q the number of sub-Dyck paths $Y_i, i = 1, \ldots Q$, going from level x_o to x_o and starting with a UU step and ending with a DD step. Denote by $l_i, i = 1, \ldots Q$, the respective length of these sub-Dyck paths. Then these sub-Dyck paths are interspaced by $\nu_N' - Q$ UD steps that split in at most $Q + 1$ sequences. Let ν_N^i $(i \leq Q + 1)$ be the respective lengths of these disjoint sequences of UD steps from x_o to x_o. Using the estimates of Step 1 and Step 2, there exist constants $C_{12}', C_{13}' > 0$ such that for any constants $A, A' > 0$ (fixed later),

$$\begin{aligned}\mathbb{P}_k\left(\exists\, \nu_N^i \geq \nu_N'/A\right) &\leq C_{12}' s_N^2 \exp\left\{-C_{13}'\nu_N'/A\right\}, \quad \forall\, 1 \leq k \leq s_N,\\ \mathbb{P}_k\left(Q \geq \nu_N'/A'\right) &\leq C_{12}' s_N^2 \exp\left\{-C_{13}'\nu_N'/A'\right\}, \quad \forall\, 1 \leq k \leq s_N.\end{aligned} \tag{37}$$

Thus it is enough to consider trajectories such that $A \leq Q \leq \nu_N'/A'$. To count these trajectories, we study their structure in more detail. Set $L_o = (s_2 - s_1)/2$ and let then $k_o < k$ be the number of odd marked instants of the sub-trajectory inside the interval

$[s_1, s_2]$. The remaining trajectory $x(t)$, $t \in [0, 2s_N] \setminus [s_1, s_2]$, is then a Dyck path of length $2s_N - 2L_o$ with $k - k_o$ odd up steps. Assume s_1 and s_2 are known. In order to count the number of such trajectories x, we first re-order the paths Y_i and UD steps inside the interval $[s_1, s_2]$ as follows. We first read the Dyck paths $Y_i, i = 1, \dots Q$, and then read all the UD steps.

Fix some $0 < \epsilon < (1 - \alpha)/2$ (small). Assume first that

$$L_o - k_o < (1 - \epsilon)s_N + \epsilon(L_o - (\nu'_N - Q)) - k. \tag{38}$$

As $k \leq \alpha s_N$ for some $\alpha < 1$, (38) ensures that $k - k_o + \nu'_N - Q \leq (1 - \epsilon)(s_N - L_o + \nu'_N - Q)$. Thus, we can apply Step 1 to the sub-trajectory obtained from x by erasing the sub-paths $Y_i, i = 1, \dots, Q$. Then, given Q, k_o, s_1 and s_2, the number of trajectories of length $L = 2s_N - 2L_o + 2(\nu'_N - Q)$ with $k - k_o + \nu'_N - Q$ odd up steps and that have $\nu'_N - Q$ UD steps between $[L - (s_N - s_2) - 2(\nu'_N - Q), L - (s_N - s_2)]$ is of order

$$C_{14} \exp\left\{-C'_o(\nu'_N - Q)\right\} \times N_2, \quad \text{if } N_2 = \sharp\{\mathcal{X}_{2s_n - 2L_o + 2(\nu'_N - Q),\, k - k_o + \nu'_N - Q}\}.$$

Here C_{14}, C'_o are some positive constants independent of s_N, k and N. Note that the constant C'_o depends only on ϵ and α. Then, the number of Dyck paths of length $\sum_{i=1}^{Q} l_i = 2L_o - 2(\nu'_N - Q)$, with $k_o - (\nu'_N - Q)$ odd up steps and coming back Q times to the level 0 using DD steps is at most of order

$$C_{15} \exp\left\{-C''_o Q\right\} \times N_1, \quad \text{if } N_1 = \sharp\left\{\mathcal{X}_{\sum_{i=1}^{Q} l_i,\, k_o - (\nu'_N - Q)}\right\}.$$

As above C_{15}, C''_o are positive constants independent of N, s_N and k. Finally, the number of ways to order the paths Y_i and the UD steps inside the interval $[s_1, s_2]$ is equal to the number of ways to write $\nu'_N - Q$ as a sum of $Q + 1$ integers. There are $C_{\nu'_N}^{Q}$ such ways. Thus the number of trajectories x coming ν'_N times to some level x_o never falling below is at most

$$\begin{aligned}
&\sum_{0 \leq s_1 < s_2 \leq s_N} \sum_{Q=A}^{\nu'_N/A'} \sum_{k_o \leq k} C_{\nu'_N}^{Q} C_{16} \exp\left\{-C_o \nu'_N\right\} N_1 N_2 \\
&\quad \leq C_{16} s_N^2 \sum_{Q=A}^{\nu'_N/A'} C_{\nu'_N}^{Q} \exp\left\{-C_o \nu'_N\right\} \mathbf{N}(s_N, k),
\end{aligned} \tag{39}$$

if $C_o = \min\{C'_o, C''_o\}$. Indeed, L_o, Q, k and ν'_N being fixed, one has that $\sum_{k_o} N_1 N_2 \leq \mathbf{N}(s_N, k)$. This yields the following estimate:

$$\mathbb{P}_k\left(x \text{ has } \nu'_N \text{ returns to } 0,\ A \leq Q \leq \frac{\nu'_N}{A'}\right) \leq \nu'_N s_N^2 e^{-C_o \nu'_N} C_{\nu'_N}^{\nu'_N/A'}.$$

We can then choose A' large enough so that there exists a constant $C_{18} > 0$, independent of N, k and s_N, such that

$$\nu'_N e^{-C_o \nu'_N} C_{\nu'_N}^{\nu'_N/A'} \leq C_{18} \exp\left\{-C_o \nu'_N/2\right\}.$$

This yields Lemma 3.2 if (38) is satisfied.
Assume now that (38) is not satisfied. Then necessarily $k_o \leq (\alpha+\epsilon)L_o \leq (\alpha+1)L_o/2$. Thus the number of trajectories Y_o coming ν'_N times to some level x_o with Q returns made using DD steps is at most

$$C_{\nu'_N}^{Q} \tilde{\mathbf{N}}(L_o - (\nu'_N - Q), k_o - (\nu'_N - Q), Q),$$

where $\tilde{\mathbf{N}}(L_o - (\nu'_N - Q), k_o - (\nu'_N - Q), Q)$ is the number of Dyck paths of length $2L_o - 2(\nu'_N - Q)$, with Q returns to 0 made only with DD steps and admitting $k_o - (\nu'_N - Q)$ odd up steps. From Step 2, one deduces that there exists a constant C_{20} (independent of k_o, L_o) such that

$$\tilde{\mathbf{N}}(L_o - (\nu'_N - Q), k_o - (\nu'_N - Q), Q) \leq e^{\{-C_{20} Q\}} \mathbf{N}(L_o - (\nu'_N - Q), k_o - (\nu'_N - Q)).$$

As $k_o \leq (\alpha + \epsilon)L_o$, there also exists $C_{21} > 0$ (depending on ϵ and α only) such that $\mathbf{N}(L_o - (\nu'_N - Q), k_o - (\nu'_N - Q)) \leq \exp\left\{-C_{21}(\nu'_N - Q)\right\}\mathbf{N}(L_o, k_o)$. The end of the proof is as above. This finishes the proof of Lemma 3.2 in the case where s_1 and s_2 are even.
To consider the case where s_1 and s_2 are odd, one can then use exactly the same arguments as above, up to the following modifications. An odd marked instant of Y_o simply defines an even marked instant of x. Then it is an easy task to show that Step 1 holds if one replaces α with $1 - \alpha'$ in (34). Step 2 and Step 3 can then be obtained by using arguments as above. □

4 The case where $p, N \to \infty$ and $N/p \to 0$

In this section, we prove the following universality result. Let $c_1, \dots, c_k$ be positive real numbers and $s_N^{(i)}$, $i = 1, \dots, K$, be sequences such that $\lim_{N\to\infty} \frac{s_N^{(i)}}{\sqrt{\gamma_N} N^{2/3}} = c_i$.

Theorem 4.1 *Assume that* $M_N = \frac{1}{N} X X^*$ *satisfies* (i) *to* (iv) *(resp.* (i') *to* (iv')*). Formulas* (6) *and* (7) *hold true.*

The proof of Theorem 4.1 is the object of the whole section. We only consider the case where $K = 1$ and where s_N is a sequence such that $\lim_{N\to\infty} \frac{s_N}{\sqrt{\gamma_N} N^{2/3}} = c$, for some real $c > 0$. We also choose to consider traces of the sequence of random matrices $M_p := \frac{1}{p} X X^*$ instead of M_N. One can check that

$$\mathbb{E}\left(\mathrm{Tr}\left(\frac{M_p}{v_+}\right)^{s_N}\right) = \mathbb{E}\left(\mathrm{Tr}\left(\frac{M_N}{u_+}\right)^{s_N}\right), \quad \text{where } v_+ = \left(1 + \frac{1}{\sqrt{\gamma_N}}\right)^2.$$

As in the preceding section, we establish that the typical paths have no edges read more than twice. This ensures that the leading term in the asymptotic expansion of $\mathbb{E}\left[\mathrm{Tr} M_p^{s_N}\right]$ is the same as that for Wishart ensembles. The idea of the proof is very similar to that of the preceding section, but requires some minor modifications. This is essentially due to the discrepancy between marked vertices on the bottom and top lines, due to the fact that $p \gg N$.
In this section, $C, C_i, C_i', D_i, B_i, i = 0, \ldots, 9$, denote some positive constants independent of N, p, k and s_N whose value may vary from line to line (and from the preceding sections).

4.1 Typical paths

We now state the counterpart of the second point in Proposition 3.1 in the following two Propositions. Let $Z_{\infty,0} = Z_{\infty,0}(A)$ be the sub-sum corresponding to the contribution to $\mathbb{E}\left(\mathrm{Tr} M_p^{s_N}\right)$ of the paths for which $k \leq s_N(1 - \frac{A}{\sqrt{\gamma_N}})$, for some $A > 0$ to be fixed.

Proposition 4.1 *There exists $A > 0$ such that $Z_{\infty,0} = o(1)v_+^{s_N}$.*

Proof of Proposition 4.1 By (15), one has that

$$Z_{\infty,0} \leq \sum_{k=1}^{s_N(1-\frac{A}{\sqrt{\gamma_N}})} \frac{N}{s_N} \frac{k}{s_N - k + 1} \left(C_{s_N}^k\right)^2 \frac{2}{\gamma_N^{s_N - k}} \sum_{(\tilde{n}, \tilde{p})} \prod_{i \geq 2} \frac{1}{n_i!} \left(\frac{C^i (s_N - k)^i}{N^{i-1}}\right)^{n_i} \prod_{i \geq 2} \frac{1}{p_i!} \left(\frac{C^i k^i}{p^{i-1}}\right)^{p_i}. \tag{40}$$

Now there exists a constant $C_1 > 0$ such that, for any $1 \leq k \leq s_N$,

$$\sum_{p_i, 1 \leq i \leq k} \prod_{i \geq 2} \frac{1}{p_i!} \left(\frac{C^i s_N^i}{p^{i-1}}\right)^{p_i} \leq \exp\left\{C_1 N^{1/3}\right\}. \tag{41}$$

Similarly, using the fact that $\sum_{i=1}^{s_N - k} i n_i = s_N - k$, we find that

$$\begin{aligned} &\frac{1}{\sqrt{\gamma_N}^{s_N-k}} \sum_{n_i, 1 \leq i \leq s_N - k} \prod_{i \geq 2} \frac{1}{n_i!} \left(\frac{C^i (s_N - k)^i}{N^{i-1}}\right)^{n_i} \\ &\leq \frac{1}{\sqrt{\gamma_N}^{s_N-k}} \sum_{n_i, 1 \leq i \leq s_N - k} \prod_{i \geq 2} \frac{1}{n_i!} \left(\frac{C^i s_N^i}{N^{i-1}}\right)^{n_i} \left(\frac{s_N - k}{s_N}\right)^{i n_i} \\ &\leq \exp\left\{C_2 N^{1/3}\right\} \left(\frac{s_N - k}{s_N}\right)^{s_N - k}, \end{aligned} \tag{42}$$

as $s_N - k > A s_N \gamma_N^{-1/2}$ and provided $A > 2$. Inserting (42) and (41) in (40) yields that

$$Z_{\infty,0} \leq \sum_{k=1}^{s_N(1-\frac{A}{\sqrt{\gamma_N}})} \frac{N}{s_N} \frac{k}{s_N - k + 1} \left(C_{s_N}^k\right)^2 \left(\frac{s_N - k}{\sqrt{\gamma_N} s_N}\right)^{s_N - k} e^{\{C_3 N^{1/3}\}}. \tag{43}$$

We deduce the following upper bound. For N large enough, one has that

$$\begin{aligned}
(43) &\leq C_4 N \sum_{k \leq s_N(1-\frac{A}{\sqrt{\gamma_N}})} \frac{e^{\{C_3 N^{1/3}\}}}{\sqrt{\gamma_N}^{s_N - k}} \left(\frac{2 s_N^2 e^2}{(s_N - k)^2}\right)^{s_N - k} \left(\frac{s_N - k}{s_N}\right)^{s_N - k} \\
&\leq C_4 \exp\left\{C_4 N^{1/3}\right\} \sum_{k \leq s_N(1-\frac{A}{\sqrt{\gamma_N}})} \left(\frac{2e^2}{A}\right)^{s_N - k} = o(1) v_+^{s_N},
\end{aligned}$$

where in the last line we have chosen $A > 4e^2$. This finishes the proof of Proposition 4.1. □

Given $0 < \epsilon < 1/2$, we also consider the contribution $Z_{\infty,1} = Z_{\infty,1}(\epsilon)$ of the paths for which $k \geq s_N(1 - \frac{\epsilon}{\sqrt{\gamma_N}})$.

Proposition 4.2 *There exists* $0 < \epsilon < 1/2$ *such that* $Z_{\infty,1} = o(1) v_+^{s_N}$.

Proof of Proposition 4.2 By (15) and as $s_N - k = O(N^{2/3})$, one has that

$$\begin{aligned}
Z_{\infty,1} &\leq 2 \sum_{k = s_N(1-\frac{\epsilon}{\sqrt{\gamma_N}})}^{s_N} N \mathbf{N}(s_N, k) \exp\left\{(C_1 + C_2') N^{1/3}\right\} \gamma_N^{-(s_N - k)} \\
&\leq C_4' N \exp\left\{C_4' N^{1/3}\right\} \mathbf{N}\left(s_N, [s_N - \left[\frac{\epsilon s_N}{\sqrt{\gamma_N}} + 1\right]\right) \gamma_N^{-\left[\frac{\epsilon s_N}{\sqrt{\gamma_N}}\right] - 1} \\
&\leq C_4' N \exp\left\{C_4' N^{1/3}\right\} \exp\left\{-C_5 N^{2/3}/8\right\} v_+^{s_N},
\end{aligned} \tag{44}$$

provided $\epsilon < 1/2$. This is enough to ensure Proposition 4.2. □

Set now $I_N = [s_N(1 - \frac{A}{\sqrt{\gamma_N}}), s_N(1 - \frac{\epsilon}{\sqrt{\gamma_N}})]$. Thanks to Propositions 4.1 and 4.2, typical paths are such that $k \in I_N$. This implies in particular that $\frac{(s_N - k)^2}{N} = O(N^{1/3})$ and $\frac{k^2}{p} = O(N^{1/3})$. Using the fact that $\sum_{k \in I_N} \mathbf{N}(s_N, k) N \gamma_N^{k - s_N} = O(v_+^{s_N})$, it is easy to deduce from (15), that it is enough to consider paths for which $M_1 + M_2 = \sum_{i \geq 2} (i-1)(n_i + p_i) \leq d_1 N^{1/3}$ for some constant d_1 independent of $k \in I_N$, N and p. For such paths, denote by $Z_\infty(k)$ the contribution of paths with k odd marked instants. We now prove the following Proposition yielding Theorem 4.1.

Proposition 4.3 *There exists* $D_1 > 0$ *such that* $\sum_{k \in I_N} Z_\infty(k) \leq D_1 v_+^{s_N}$. *Furthermore, the contribution of paths admitting either an edge read more than twice, or more than*

$\ln\ln N$ vertices of type 3, or a vertex of type 4 or greater, or a marked origin is negligible in the large-N-limit.

Remark 4.1 Considering the paths $P_k, k \in I_N$, admitting only vertices of type 2 at most and no non-MP-closed vertices, one can also deduce from the subsequent proof of Proposition 4.3, that there exists $D_2 > 0$ such that $\sum_{k\in I_N} Z_\infty(k) \geq D_2 v_+^{s_N}$.

Proof of Proposition 4.3 First, one can state the counterpart of Formula (26). Due to the different scales $s_N - k = O(N^{2/3})$, while $k = O(\sqrt{\gamma_N} N^{2/3})$, we need in this section to distinguish vertices being the left endpoint of an up edge according to the parity of the corresponding instant. Define $\nu_{N,o}(P_k)$ (resp. $\nu_{N,e}(P_k)$ to be the maximum number of vertices visited (at marked instants) from a vertex of the path occurring at odd instants (resp. even instants). Then,

$$\begin{aligned}
Z_\infty(k) \leq\ & C\mathbf{N}(s_N, k) N \gamma_N^{k-s_N} \mathrm{e}^{\left\{-\frac{(s_N-k)^2}{2N} - \frac{k^2}{2p}\right\}} \\
& \sum_{n_2, r_1, q_1, n_3, \ldots, n_{s_N-k}} \ \sum_{p_2, r_2, q_2, p_3, \ldots, p_k} \mathbb{E}_k \left[\frac{\left((s_N-k)^2/(2N)\right)^{n_2-r_1-q_1}}{(n_2-r_1-q_1)!} \right. \\
& \times \frac{1}{r_1!} \left(\frac{3(s_N-k) \max\ x(t)}{N} \right)^{r_1} \frac{1}{q_1!} \left(\frac{D_3(s_N-k)(\nu_{N,o}+T_N)}{N} \right)^{q_1} \\
& \times \frac{\left(k^2/(2p)\right)^{p_2-r_2-q_2}}{(p_2-r_2-q_2)!} \frac{1}{r_2!} \left(\frac{3k \max\ x(t)}{p} \right)^{r_2} \frac{1}{q_2!} \left(\frac{D_4 k(\nu_{N,e}+T_N)}{p} \right)^{q_2} \\
& \left. \times \prod_{i\geq 3} \frac{1}{n_i!} \left(\frac{C^i (s_N-k)^i}{N^{i-1}} \right)^{n_i} \prod_{i\geq 3} \frac{1}{p_i!} \left(\frac{C^i k^i}{p^{i-1}} \right)^{p_i} \right].
\end{aligned} \tag{45}$$

One still has that $T_N < A'' N^{1/3}/\ln N$ for some $A'' > 0$ in typical paths (independently of $k \in I_N$). This ensures that the analysis performed in the case where $\lim_{N\to\infty} \gamma_N < \infty$ can be copied, provided the counterparts of the lemmas of Sect. 3.3 hold. Let $a > 0$ be given. Assume for a while that typical paths are such that there exists $\epsilon' > 0$ such that, $\forall k \in I_N$,

$$\begin{aligned}
& \max_{k\in I_N} \mathbb{E}_k \left(\exp\left\{ a \frac{\max x(t)}{N^{1/3}} \right\} \right) < b, \ \text{ for some } b > 0, \\
& \nu_{N,o} < N^{1/3-\epsilon'} \quad \text{and} \quad \nu_{N,e} < \sqrt{\gamma_N} N^{1/3-\epsilon'}.
\end{aligned} \tag{46}$$

The above statement will be proved in the subsequent section (Lemmas 4.1 and 4.2) and using exactly the same arguments as in Lemma 7.8 in [9]. We then copy the arguments of Proposition 3.2 and the sequel. Then it is easy to deduce that typical paths have a non-marked origin, vertices of type 3 at most (and a number of vertices of type 3 smaller than $\ln\ln N$) and no edge passed more than twice. The other paths lead to a negligible contribution. We can also deduce that non-MP-closed vertices of type 2 as well as vertices of type 3 occur only on the bottom line in typical paths.

In particular, let $Z_{3,\infty}$ denote the contribution of paths for which $q_1 + q_2 = 0$, $r_2 = 0$, $\sum_{i\geq 4} n_i + \sum_{i\geq 3} p_i = 0$, $n_3 \leq \ln\ln N$ and no edges read more than twice. Then $\mathbb{E}\left[\mathrm{Tr} M_p^{s_N}\right] = Z_{3,\infty}(1+o(1))$ and

$$Z_{3,\infty} \leq C_6 \sum_{k\in I_N} N\gamma_N^{k-s_N}\mathbf{N}(s_N,k) \leq \frac{D_1}{2} v_+^{s_N}.$$

This ensures that the limiting expectation depends only on the variance of the entries and has the same behavior as for Wishart ensembles. The proof of Theorem 4.1 is complete, provided we prove the announced Lemmas. □

4.2 Technical Lemmas

We now state the counterpart of Lemma 3.1.

Lemma 4.1 *Let $a > 0$ be a given real number. There exists $b = b(a) > 0$ such that* $\max_{k\in I_N} \mathbb{E}_k\left(e^{\{\frac{a\max x(t)(s_N-k)}{N}\}}\right) < b.$

Remark 4.2 Lemma 4.1 also yields that $\max_{k\in I_N} \mathbb{E}_k\left(e^{\{a\frac{\max x(t)k}{p}\}}\right) - 1 \ll 1.$

Proof of Lemma 4.1 The proof refers to the proof of Lemma 3.1 in Sect. 3.3. Let l be the number of UU steps of a Dyck path x for which $k \in I_N$. From (28), one deduces that $l \leq s_N - k \leq A\frac{s_N}{\sqrt{\gamma_N}}$. Thus by (31), we deduce that, if $r \geq a_o\sqrt{As_N\gamma_N}^{-1/2}$, and for any $k \in I_N$, $\mathbb{P}_k(\max x(t) = r) \leq \frac{a_3}{\sqrt{s_N-k}}\exp\left\{-\frac{a_4 r^2}{(s_N-k)}\right\}$. This readily proves Lemma 4.1. □

One next turns to establishing the counterpart of Lemma 3.2. We denote by $\Gamma_{\nu_{N,o}(x)}$ (resp. $\Gamma_{\nu_{N,e}(x)}$) the event that the maximal number of times the Dyck path x comes from above (without falling below) to some level x_o at even instants (resp. odd instants) is $\nu_{N,e}$ (resp. $\nu_{N,o}$).

Lemma 4.2 *There exist positive constants B_1, B_2, B_3, B_4, independent of N and p such that*

$$\max_{s_N(1-\frac{A}{\sqrt{\gamma_N}})\leq k\leq s_N(1-\frac{\epsilon}{\sqrt{\gamma_N}})} \mathbb{P}_k(\Gamma_{\nu_{N,o}(x)}) \leq B_1\frac{s_N^2}{\gamma_N}\exp\left\{-B_2\nu_{N,o}\right\}. \tag{47}$$

$$\max_{s_N(1-\frac{A}{\sqrt{\gamma_N}})\leq k\leq s_N(1-\frac{\epsilon}{\sqrt{\gamma_N}})} \mathbb{P}_k(\Gamma_{\nu_{N,e}(x)}) \leq B_3\frac{s_N^2}{\gamma_N}\exp\left\{-\frac{B_4\nu_{N,e}}{\sqrt{\gamma_N}}\right\}. \tag{48}$$

Proof of Lemma 4.2 As in Sect. 3.3, we have to estimate the probability that the Dyck path x comes ν_N' times to some level x_o without falling below in some (maximal) time interval $[s_1, s_2]$. Note that the two steps leading and starting at s_1 (resp. s_2) are up

(resp. down) steps. This is because $[s_1, s_2]$ is a maximal interval. Thus the number of possible choices for s_1 and s_2 is at most of order s_N^2/γ_N, as $k \in I_N$.
We first prove (47) and thus assume that s_1 is odd. Then the returns to x_o occur at odd instants. The counterpart of formula (34) states

$$\mathbb{P}_k\left(x|_{t\in[s_1',s_2']} \text{ has only } \eta_N\ UD \text{ steps}\right) \leq C_9'\left(\frac{2A}{\sqrt{\gamma_N}}\right)^{\eta_N}. \tag{49}$$

The two steps preceding (and following) $[s_1', s_2']$ are either both up steps or both down steps (regardless of the fact that s_1', s_2' are even or odd). The estimate (35) still holds (up to the change $s_N^2 \to s_N^2/\gamma_N$) so that formula (47) is proved, copying the proof of Lemma 3.2.
We now turn to the proof of (48) which is more involved than in Lemma 3.2. Formula (34) translates to

$$\mathbb{P}_k\left(x \text{ has } \eta_N \text{ UD steps in between } [s_1', s_2']\right) \leq C_9 \exp\left\{-\epsilon\frac{\eta_N}{\sqrt{\gamma_N}}\right\}.$$

Step 1 and Step 2 are then obtained as in Lemma 3.2 (with $s_N^2 \to s_N^2/\gamma_N$). From that, we can deduce that we can consider in Step 3 only the paths for which $A_1 \leq Q \leq A_o\nu_N'/\sqrt{\gamma_N}$ for some constants $A_1, A_o > 0$. We need to refine the estimate for Step 3. Let then $[s_1, s_2]$ be the interval where ν_N' returns to some level x_o occur. We call Y_o the trajectory defined by $x(t) - x_o$, $t \in [s_1, s_2]$. We then define k_o to be its number of odd up steps, Q to be its number of returns to 0 using DD steps, l (resp. μ_o, ν_o'') to be its number of UU steps (resp. of DU steps and of UD steps occurring at some positive level). Assume that l, Q, μ_o, k_o are given and observe that $k_o = l + \nu_N' - Q + \nu_o''$. Let then $\mathbb{P}_{l,Q,k_o,\mu_o}$ denote the conditional probability on the event that Y_o has k_o odd up steps, μ_o DU steps and $Dy(Y_o)$ has $2l$ steps and Q returns to 0. Then, one has that

$$\mathbb{P}_{l,Q,k_o,\mu_o}(\ Y_o \text{ has } \nu_N' - Q\ UD \text{ steps at level } 0\ ) \leq \frac{C_{\nu_N'}^{Q}\, C_{(s_2-s_1)/2-\nu_N'}^{\nu_o''}}{C_{(s_2-s_1)/2}^{k_o-l}}.$$

One first shows that it is enough to consider the subpaths Y_o such that $\frac{2k_o}{s_2-s_1} = \frac{k}{s_N}(1 + o(1)) \in \left[1-\epsilon_1^{-1}\gamma_N^{-1/2}, 1-\epsilon_1\gamma_N^{-1/2}\right]$ for some $\epsilon_1 > 0$ small enough. This follows from the fact that $k \in I_N$ and arguments already used in Sect. 2.5 [see also (17)]. This yields that

$$\begin{aligned}
&\mathbb{P}_k\left(\Gamma_{\nu_{N,e}(x)} \cap \left\{\frac{2k_o}{(s_2-s_1)} \leq 1-\epsilon_1^{-1}\gamma_N^{-1/2}\right\}\right)\\
&\leq \sum_{s_1\leq s_2}\ \sum_{\frac{2k_o}{(s_2-s_1)}\leq 1-\epsilon_1^{-1}\gamma_N^{-1/2}} \frac{\mathbf{N}(\frac{s_2-s_1}{2}, k_o)\mathbf{N}(s_N - \frac{s_2-s_1}{2}, k-k_o)}{\mathbf{N}(s_N, k)}\\
&\leq \frac{s_N^2}{\gamma_N}\exp\left\{-\eta N^{2/3}\right\},
\end{aligned}$$

for some $\eta > 0$ provided $\epsilon_1^{-1} > 2A$, where A has been fixed in Proposition 4.1. The analysis of the case where $2k_o/(s_2 - s_1)) \geq 1 - \epsilon_1 \gamma_N^{-1/2}$ is similar. We can assume that B_4 in (48) is small enough so that $B_4 s_N < \sqrt{\gamma_N} \eta N^{2/3}$. This yields (48) and ensures that it is enough to consider the case where $\frac{2k_o}{s_2 - s_1} = \frac{k}{s_N}(1 + o(1)) \in [1 - \epsilon_1^{-1}\gamma_N^{-1/2}, 1 - \epsilon_1 \gamma_N^{-1/2}]$. Fixing l and k_o we set $Q_T := (1 - \frac{k_o - l}{(s_2 - s_1)/2})\nu'_N$. As $l \leq \frac{s_2 - s_1}{2} - k_o \leq \frac{\epsilon_1^{-1}}{\sqrt{\gamma_N}} \frac{s_2 - s_1}{2}$, one has that $\frac{\epsilon_1}{\sqrt{\gamma_N}} \leq \frac{Q_T}{\nu'_N} \leq \frac{2\epsilon_1^{-1}}{\sqrt{\gamma_N}}$. One can check that there exists a constant C_o independent of s_1, s_2, k_o and l such that, for a given constant $A_2 > 4$,

$$\frac{C^{Q}_{\nu'_N} C^{\nu''_o}_{(s_2 - s_1)/2 - \nu'_N}}{C^{k_o - l}_{(s_2 - s_1)/2}} \leq C_o \exp\left\{ -\frac{(Q - Q_T)^2}{(A_2 + 1) Q_T} \right\}, \text{ if } \quad Q \geq Q_{A_2} := Q_T(1 - A_2),$$

$$\frac{C^{Q}_{\nu'_N} C^{\nu''_o}_{(s_2 - s_1)/2 - \nu'_N}}{C^{k_o - l}_{(s_2 - s_1)/2}} \leq C_o \exp\left\{ -\frac{A_2 Q_T}{2} \right\} \left(\frac{1}{5}\right)^{Q_T(1 - A_2) - Q}, \text{ if } \quad Q \leq Q_{A_2}.$$

Thus, it is clear that the proportion of paths coming back ν'_N times from above to some level x_o and for which $Q \leq Q_T(1 - \epsilon_1)$ is at most of order $s_N^2/\gamma_N \exp\left\{ -\frac{\epsilon_1^3 \nu'_N}{(A_2 + 1)\sqrt{\gamma_N}} \right\}$. Paths for which $Q \geq Q_T(1 - \epsilon_1) \geq \frac{\epsilon_1 \nu'_N}{2\sqrt{\gamma_N}}$ are considered as in Step 2. This is enough to ensure (48). □

Remark 4.3 The investigation of higher moments follows the same steps as in Sect. 2.5. In particular, considering $\mathrm{Var}(\mathrm{Tr} M_p^{s_N})$, only pairs of correlated paths such that $\mathcal{P}_1 \vee \mathcal{P}_2$ has a number of odd up steps of order $2s_N \left(1 - O(\sqrt{\gamma_N^{-1}})\right)$ are non-negligible. In (18), one can also replace the term $(2s_N - 1 - l)$ with $s_N - (1 + l)/2 - k_1$. Considering as above the exponential decay of (20), one can also show that $\sum_{k_1 \leq k \wedge 2s_N - 1 - l} f(k_1) \leq C'_3(2s_N - 1 - l - \tilde{k}_1)^{1/2} f(\tilde{k}_1)$ where $\tilde{k}_1 = \left[k \frac{2s_N - 1 - l}{4s_N - 2}\right](+1)$. Thus (22) can be replaced with

$$\sum_{l=0}^{2s_N - 1} C_3\left(s_N - \frac{l+1}{2} - \tilde{k}_1\right) \sqrt{\frac{(2s_N - 1 - l)}{\sqrt{\gamma_N}}} f(\tilde{k}_1) \leq C_4 \frac{\sqrt{s_N} \gamma_N^{-3/4}}{(1 - \alpha_N)^2},$$

where $\alpha_N = k/(2s_N - 1) \sim 1 - 1/\sqrt{\gamma_N}$. One can readily deduce from the above that the contribution of (19) is negligible. The case where $x(t_e) = 0$ yields a negligible contribution, as readily seen from (46) and Lemma 4.2. The latter is then enough to ensure that $\mathrm{Var}(\mathrm{Tr} M_p^{s_N})$ is bounded and only depends on the variance of the entries. The investigation of higher moments is similar.

Acknowledgments I thank C. Tracy, A. Soshnikov, D. Paul, N. Patterson, R. Cont, L. Choup and C. Semadeni for their great help in the improvement which lead to the final version of this paper. This work was done while visiting UC Davis.

References

1. Bai, Z.D.: Methodologies in spectral analysis of large-dimensional random matrices, a review. Stat. Sin. **9**(3), 611–677 (1999)
2. Bai, Z.D., Krishnaiah, P.R., Yin, Y.Q.: On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. Probab. Theor. Relat. Fields **78**, 509–521 (1988)
3. Baik, J., Ben Arous, G., Péché, S.: Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. Ann. Probab. **33**(5), 1643–1697 (2005)
4. Baik, J., Silverstein, J.: Eigenvalues of large sample covariance matrices of spiked population models. J. Mult. Anal. **97**, 1382–1408 (2006)
5. Bronk, B.V.: Exponential ensembles for random matrices. J. Math. Phys. **6**, 228–237 (1965)
6. Chen, W., Yan, S., Yang, L.: Identities from Weighted 2 Motzkin paths. http://www.billchen.org/publications/identit/identit.pdf
7. El Karoui, N.: On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. ArXiv math.ST/0309355 (2003)
8. El Karoui, N.: Recent results about the largest eigenvalue of random covariance matrices and statistical application. Acta Phys. Polon. B **36**(9), 2681–2697 (2005)
9. Féral, D., Péché, S.: The largest eigenvalue of some rank one deformation of large Wigner matrices. ArXiv math.PR/0605624. Comm. Math. Phys. (to appear) (2006)
10. Geman, S.: A limit theorem for the norm of random matrices. Ann. Prob. **8**, 252–261 (1980)
11. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**, 417–441 (1933)
12. Hoyle, D., Rattray, M.: Limiting form of the sample covariance eigenspectrum in PCA and kernel PCA. Advances in Neural Information Processing Systems NIPS 16 (2003)
13. James, A.: Distributions of matrix variates and latent roots derived from normal samples. Ann. Math. Stat. **31**, 151–158 (1960)
14. Johansson, K.: Shape fluctuations and random matrices. Comm. Math. Phys. **209**, 437–476 (2000)
15. Johnstone, I.M.: On the distribution of the largest principal component. Ann. Stat. **29**, 295–327 (2001)
16. Johnstone, I.M.: High dimensional statistical inference and random matrices. arXiv:math/0611589. In: Proceedings of the International Congress of Mathematicians, 2006 (to appear) (2007)
17. Jonsson, D.: Some limit theorems for the eigenvalues of sample covariance matrices. J. Mult. Anal. **12**, 1–38 (1982)
18. Laloux, L., Cizeau, P., Potters, M., Bouchaud, J.: Random matrix theory and financial correlations. Int. J. Theor. Appl. Financ. **3**(3), 391–397 (2000)
19. Malevergne, Y., Sornette, D.: Collective origin of the coexisence of apparent RMT noise and factors in large sample correlation matrices. Physica A **331**(3–4), 660–668 (2004)
20. Marcenko, V.A., Pastur, L.A.: Distribution of eigenvalues forsome sets of random matrices. Math. USSR-Sbornik **1**, 457–486 (1967)
21. Patterson, N., Price, A.L., Reich, D.: Population structure and eigenanalysis. PLoS Genet **2**(12), e190 (2006) DOI:10.1371/journal.pgen.0020190
22. Péché, S., Soshnikov, A.: Wigner random matrices with non-symmetrically distributed entries. arXiv:math/0702035. Applications of random matrices, determinants and pfaffians to problems in statistical mechanics. J. Stat. Phys (to appear) (2007)
23. Oravecz, F., Petz, D.: On the eigenvalue distribution of some symmetric random matrices. Acta Sci. Math. **63**, 383–395 (1997)
24. Plerous, V., Gopikrishnan, P., Rosenow, B., Amaral, L., Guhr, T., Stanley, H.: Random matrix approach to cross correlations in financial data. Phys. Rev. E **65**(6), 66–126 (2002)
25. Ruzmaikina, A.: Universality of the edge distribution of eigenvalues of Wigner random matrices with polynomially decaying distributions of entries. Commun. Math. Phys. **261**, 277–296 (2006)
26. Sear, R., Cuesta, J.: Instabilities in complex mixtures with a large number of components. Phys. Rev. Lett. **91**(24), 245–701 (2003)
27. Silverstein, J.: On the weak limit of the largest eigenvalue of a large-dimensional sample covariance matrix. J. Multivariate Anal. **30**(2), 307–311 (1989)
28. Sinai, Y., Soshnikov, A.: Central limit theorem for traces of large random symmetric matrices with independent matrix elements. Bol. Soc. Brasil. Mat. (N.S.) **29**(1), 1–24 (1998)
29. Sinai, Y., Soshnikov, A.: A refinement of Wigner's semicircle law in a neighborhood of the spectrum edge for random symmetric matrices. Funct. Anal. Appl. **32**, 114–131 (1998)

30. Soshnikov, A.: Universality at the edge of the spectrum in Wigner random matrices. Commun. Math. Phys. **207**, 697–733 (1999)
31. Soshnikov, A.: A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. J. Stat. Phys. **108**, 1033–1056 (2002)
32. Stanley, R.: Enumerative combinatorics, vol. 2. Cambridge Studies in Advanced Mathematics, 62. Cambridge University Press, Cambridge (1999)
33. Sulanke, R.: Counting lattice paths by Narayana polynomials. Electron. J. Combin. **7**, 9p (electronic) (2000)
34. Sulanke, R.: Catalan path statistics having the Narayana distribution. In: Proceedings of the 7th Conference on Formal Power Series and Algebraic Combinatorics (Noisy-le-Grand, 1995). Discrete Math. **180**(1–3), 369–389 (1998)
35. Telatar, E.: Capacity of multi-antenna Gaussian channels. Eur. Trans. Telecomm. **10**(6), 585–595 (1999)
36. Tracy, C., Widom, H.: Level-spacing distribution and Airy kernel. Commun. Math. Phys. **159**, 151–174 (1994)
37. Tracy, C., Widom, H.: On orthogonal and symplectic matrix ensembles. Commun. Math. Phys. **177**, 727–754 (1996)
38. Wachter, K.: The strong limits of random matrix spectra for sample covariance matrices of independent elements. Ann. Probab. **6**, 1–18 (1978)