

Large Deviations of the Maximum Eigenvalue for Wishart and Gaussian Random Matrices

Satya N. Majumdar¹ and Massimo Vergassola²

¹ *Laboratoire de Physique Théorique et Modèles Statistiques (UMR 8626 du CNRS), Université Paris-Sud, Bâtiment 100, 91405 Orsay Cedex (France)*

² *Institut Pasteur, CNRS URA 2171, F-75724 Paris 15, France.*

(Dated: November 17, 2008)

We present a simple Coulomb gas method to calculate analytically the probability of rare events where the maximum eigenvalue of a random matrix is much larger than its typical value. The large deviation function that characterizes this probability is computed explicitly for Wishart and Gaussian ensembles. The method is quite general and applies to other related problems, e.g. the joint large deviation function for large fluctuations of top eigenvalues. Our results are relevant to widely employed data compression techniques, namely the principal components analysis. Analytical predictions are verified by extensive numerical simulations.

PACS numbers: 02.50.-r, 02.50.Sk, 02.10.Yn, 24.60.-k, 21.10.Ft

Rare events where one of the eigenvalues of a random matrix is much larger than the others play an important role in data compression techniques such as the “Principal Components Analysis” (PCA). PCA is a very useful method to detect hidden patterns or correlations in complex, high-dimensional datasets. A non-exhaustive list of applications includes image processing [1, 2, 3], biological microarrays [4, 5], population genetics [6, 7], finance [8, 9], meteorology and oceanography [10]. The main idea behind PCA is very simple. Consider a rectangular ($M \times N$) matrix X whose entries X_{ij} represent some data. For instance, X_{ij} might represent examination marks of the i -th student ($1 \leq i \leq M$) in the j -th subject (physics, mathematics, chemistry, etc., with $1 \leq j \leq N$). The product ($N \times N$) symmetric matrix $W = X^T X$ represents the covariance matrix of the data and it contains information about correlations. In PCA, one first identifies eigenvalues and eigenvectors of W . The data are maximally scattered and correlated along the eigenvector (“principal component”) associated with the largest eigenvalue λ_{\max} . The scatter progressively reduces as lower and lower eigenvalues are considered. The subsequent step is the reduction of data dimensionality, achieved by setting to zero those components corresponding to low eigenvalues. The *rationale* is that retaining the largest components will preserve the major patterns in the data and only minor variations are filtered out.

The above description of PCA makes it clear that the efficiency of the method crucially depends upon the gap between the top eigenvalues and the “sea” of intermediate and small eigenvalues. In particular, the further is the maximum eigenvalue λ_{\max} spaced from all the others, the more effective the projection and the compression procedure will be. A question then naturally arises: how good is PCA for random data? This issue has a twofold interest. First, in many situations, the data are high-dimensional and have random components. Second, random ensembles provide null models needed to gauge the statistical significance of results obtained for non-random datasets. To address the question just formulated, one needs to compute the probability of rare events where the largest eigenvalue λ_{\max} has atypically large fluctuations. The purpose of this Letter is to provide a simple physical method, based on

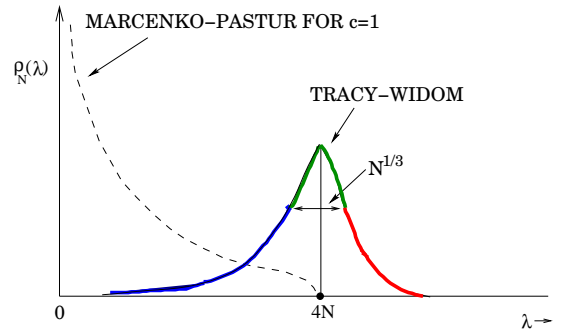


FIG. 1: The dashed line shows schematically the Marcenko-Pastur average density of states for Wishart matrices with the aspect-ratio parameter $c \equiv N/M = 1$ and the full line is the distribution of the maximum eigenvalue λ_{\max} . The PDF is centered around the mean value $\langle \lambda_{\max} \rangle = 4N$ and typically fluctuates over a scale of width $N^{1/3}$. The probability of fluctuations on this scale is described by the known Tracy-Widom distribution (green curve). The red (blue) line on the right (left) describes the right (left) large deviation tail of the PDF, which is the object of interest in this paper.

the Coloumb gas method in statistical physics, that allows us to compute analytically the probability of these rare events for a general class of random matrices.

Let us start by considering Wishart matrices [11], which are directly related to PCA and multivariate statistics [12]. Wishart matrices are defined via the product $W = X^T X$ of a ($M \times N$) random matrix X having its elements drawn independently from a Gaussian distribution, $P[X] \propto \exp \left[-\frac{\beta}{2} \text{tr}(X^\dagger X) \right]$. The Dyson indices $\beta = 1, 2$ correspond respectively to real and complex X [13]. Without any loss of generality, we will assume hereafter that $M \geq N$. In addition to the aforementioned PCA applications, Wishart matrices appear in antenna selection in communication technology [14], nuclear physics [15], quantum chromodynamics [16], statistical physics of directed polymers in random media [17] and nonintersecting Brownian motions [18].

The spectral properties of $W = X^T X$ are well-known. For example, N eigenvalues $\{\lambda_i\}$'s of W are nonnegative random variables with a joint probability density function (PDF) [19]

$$P[\{\lambda_i\}] \propto e^{-\frac{\beta}{2} \sum_{i=1}^N \lambda_i} \prod_{i=1}^N \lambda_i^{\alpha \beta/2} \prod_{j < k} |\lambda_j - \lambda_k|^\beta, \quad (1)$$

where $\alpha = (1 + M - N) - 2/\beta$. This can be written as $P[\{\lambda_i\}] \propto \exp[-\beta E(\{\lambda_i\})/2]$, with the energy

$$E[\{\lambda_i\}] = \sum_{i=1}^N (\lambda_i - \alpha \log \lambda_i) - \sum_{j \neq k} \ln |\lambda_j - \lambda_k|, \quad (2)$$

coinciding with that of a 2-d Coulomb gas of charges with coordinates $\{\lambda_i\}$. Charges are confined to the positive half-line in the presence of an external linear+logarithmic potential. The external potential tends to push the charges towards the origin, whilst the Coulomb repulsion tends to spread them apart. A glance at (2) indicates that these two competing mechanisms balance for values of λ scaling as $\sim N$. Indeed, from the joint PDF (1), one can calculate the average density of eigenvalues, $\rho_N(\lambda) = \frac{1}{N} \sum_{i=1}^N \langle \delta(\lambda - \lambda_i) \rangle \approx \frac{1}{N} f(\frac{\lambda}{N})$, with the Marcenko-Pastur (MP) [20] scaling function:

$$f(x) = \frac{1}{2\pi x} \sqrt{(b-x)(x-a)}. \quad (3)$$

Here, $c = N/M$ (with $c \leq 1$) and the upper and lower edges are $b = (c^{-1/2} + 1)^2$ and $a = (c^{-1/2} - 1)^2$. For all $c < 1$, the average density vanishes at both edges of the MP sea. For the special case $c = 1$, we have $a = 0$, $b = 4$ and the average density $f(x) = \frac{1}{2\pi} \sqrt{(4-x)/x}$ for $0 \leq x \leq 4$, diverges as $x^{-1/2}$ at the lower edge (see Fig. 1).

The MP expression shows that the maximum eigenvalue λ_{\max} has the average value $\langle \lambda_{\max} \rangle \approx bN$ for large N . Typical fluctuations of λ_{\max} around its mean are known to be of $O(N^{1/3})$ [12, 17]. More specifically, $\lambda_{\max} = bN + c^{1/6} b^{2/3} N^{1/3} \chi$, where χ has an N -independent limiting PDF, $g_\beta(\chi)$, the well-known Tracy-Widom (TW) density [21]. The TW distribution for $\beta = 1, 2$ has asymmetric tails [21]

$$g_\beta(\chi) \sim \exp\left[-\frac{\beta}{24} |\chi|^3\right] \quad \text{as } \chi \rightarrow -\infty, \quad (4)$$

$$\sim \exp\left[-\frac{2\beta}{3} \chi^{3/2}\right] \quad \text{as } \chi \rightarrow \infty. \quad (5)$$

In contrast, the probability of *atypically large*, e.g. $\sim O(N)$, fluctuations of λ_{\max} from its mean bN are not captured by the TW distribution. Note that these configurations are precisely those relevant here, as they are ideally suited for the PCA to work accurately.

How does the PDF $P(\lambda_{\max}, N)$ look like for $|\lambda_{\max} - bN| \gg O(N^{1/3})$ where the TW form is no longer valid? Using general large deviation principles, Johansson [17] proved that for large fluctuations $\sim O(N)$ from its mean, the PDF

$P(\lambda_{\max} = t, N)$ has the form (for large N):

$$\begin{aligned} P(t, N) &\sim \exp\left[-\beta N^2 \Phi_-\left(\frac{bN-t}{N}\right)\right] \quad t \ll bN; \\ &\sim \exp\left[-\beta N \Phi_+\left(\frac{t-bN}{N}\right)\right] \quad t \gg bN; \end{aligned} \quad (6)$$

where $\Phi_\pm(x)$ are the right (left) rate functions for the large positive (negative) fluctuations of λ_{\max} . The challenge is to explicitly compute their functional forms. The approach developed for Gaussian matrices [22] allows to compute the left function $\Phi_-(x)$ [23] but it does not apply to the right tail. The problem of computing the right function $\Phi_+(x)$ is solved hereafter. This is followed by the application of the new method to Gaussian matrices and further generalizations.

The starting point of our method to compute $\Phi_+(x)$ is the energy expression (2). The Coulomb gas physics suggests that when the rightmost charge is moved to the right, $\lambda_{\max} - bN \sim O(N)$, the MP sea is *a priori* not subject to forces capable of macroscopic rearrangements. Following this physical picture, the right rate function is determined by the energy cost in pulling the rightmost charge in the external potential of the Coulomb gas and the interaction of the charge with the *unperturbed* MP sea. This energy cost for $\lambda_{\max} = t \gg bN$ can be estimated for large N using Eq. (2)

$$\Delta E(t) = t - \alpha \ln(t) - 2N \int \ln|t - \lambda| \rho_N(\lambda) d\lambda, \quad (7)$$

where $\rho_N(\lambda)$ is the MP average density of charges and the integral describes the Coulomb interaction of the rightmost charge with the MP sea. This energy cost expression is valid up to an additive constant, which is chosen such that $\Delta E(t = bN) = 0$ since our reference configuration is the one where $\lambda_{\max} = bN$. For large N , we scale $t = zN$, use the MP expression (3) and the energy cost finally takes the form

$$\frac{\Delta E(z)}{N} = z - \frac{1-c}{c} \ln(z) - 2 \int_a^b \ln(z-z') f(z') dz', \quad (8)$$

valid for $z \geq b$ and up to an additive constant. The probability of such a configuration is $P(z, N) \propto \exp[-\beta \Delta E(z)/2]$. Making a shift of variable $z = b + x$, it follows that $P(t, N)$ for large N and for $t - bN \sim O(N)$ agrees with the large deviation behavior in Eq. (6). Progress is that we also have derived the explicit expression of the right rate function $\Phi_+(x)$

$$\begin{aligned} \Phi_+(x) &= \frac{x}{2} - \frac{1-c}{2c} \ln\left(\frac{x+b}{b}\right) \\ &\quad - \int_a^b \ln\left(\frac{x+b-x'}{b-x'}\right) f(x') dx', \end{aligned} \quad (9)$$

where $x > 0$ and the additive constant was chosen to have $\Phi_+(0) = 0$. The integral can be performed exactly and expressed as a hypergeometric function. For $c = 1$ ($a = 0$ and $b = 4$), we obtain

$$\Phi_+(x) = \frac{x+2}{2} - \ln(x+4) + \frac{1}{x+4} G\left(\frac{4}{4+x}\right), \quad (10)$$

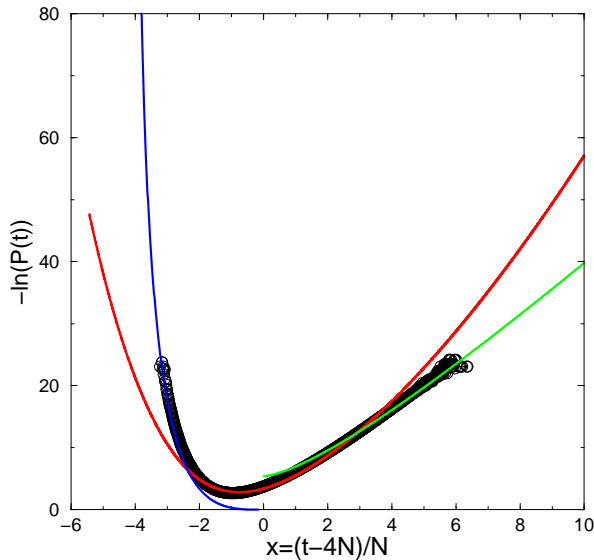


FIG. 2: Numerical results (circles) for the maximum eigenvalue distribution $-\ln P(\lambda_{\max} = t, N)$ vs the scaled variable $(t - 4N)/N$. Here, $N = 10$, Wishart matrices are real ($\beta = 1$) and $c = 1$. Comparisons are with the Tracy-Widom distribution (red line) and the exact right (green line) and left (blue line) large deviation predictions.

where $G(z) = {}_3F_2\left[\{1, 1, 3/2\}, \{2, 3\}, z\right]$ is a hypergeometric function (with a lengthy but explicit expression in terms of elementary functions). For the sake of comparison, we also report the simpler expression of the left rate function [23]: $\Phi_-(x) = \ln(2/\sqrt{4-x}) - x/8 - x^2/64$ for $x \geq 0$.

The asymptotics of $\Phi_+(x)$ can be easily worked out from Eq. (9). For large x , $\Phi_+(x) \sim x/2$ independently of c , while the function has a nonanalytic behavior for small x :

$$\Phi_+(x) \approx \frac{\sqrt{b-a}}{3b} x^{3/2} \quad \text{as } x \rightarrow 0. \quad (11)$$

This shows that, as $\lambda_{\max} - bN \ll O(N)$ from the right side, the PDF of $\lambda_{\max} = t$ in Eq. (6) behaves as $\exp[-\beta N(\sqrt{b-a}/3b)(t/N - b)^{3/2}]$. Expressing the exponent in terms of the TW variable $\chi = c^{-1/6}b^{-2/3}N^{-1/3}(t - bN)$, we recover exactly the right tail behavior of the TW density in Eq. (5). Thus, the large deviation function $\Phi_+(x)$ matches, for small arguments x , the behavior of the TW density at large arguments. This is quite consistent with the fact that the scales of the fluctuations for TW and $\Phi_+(x)$ are $O(N^{1/3})$ and $O(N)$, respectively. In fact, our method provides, as a bonus, a physical derivation of the right tail behavior of the TW density (originally derived through the Painlevé differential equation satisfied by the TW distribution [21]).

We confirmed theoretical predictions by extensive numerical simulations. About 10^{11} realizations of real ($\beta = 1$) Wishart matrices of sizes $N = 10, 26, 50, 100$ and with different values of $c \leq 1$ were efficiently generated using the tridiagonal results in [24]. We find very good agreement with our analytical predictions for the right large deviations. For

example, in Fig. (2) we present the results for $c = 1$ and $N = 10$. MonteCarlo numerical results are compared to the TW density (obtained by numerically integrating the Painlevé equation satisfied by the TW distribution [21]) and $\Phi_+(x)$ in Eq. (10), multiplied by N . For comparison, we also show the corresponding left rate function $\Phi_-(-x)$ [23] multiplied by N^2 . It is clear that, while the numerical data are well described by the TW density near the peak of the distribution, they deviate considerably from TW as one moves into the tails, where our large deviation predictions work perfectly.

Our Coulomb gas method is quite general and it can be applied to other related problems. For example, we can compute the right large deviation function of λ_{\max} for Gaussian random matrices. For the latter, the eigenvalues can be positive or negative with joint PDF [25]:

$$P\{\{\lambda_i\}\} \propto e^{-\frac{\beta}{2} \sum_{i=1}^N \lambda_i^2} \prod_{j < k} |\lambda_j - \lambda_k|^\beta, \quad (12)$$

where the Dyson indices $\beta = 1, 2$ and 4 correspond to the orthogonal, unitary and symplectic ensembles. The quadratic nature of the potential in (12), in contrast to the linear term appearing in (1), makes that the amplitude of a typical eigenvalue scales as $\sim \sqrt{N}$. The average density of states for large N has the scaling form, $\rho_N(\lambda) \approx \frac{1}{\sqrt{N}} f_{\text{sc}}\left(\frac{\lambda}{\sqrt{N}}\right)$, where the famous Wigner semi-circular law $f_{\text{sc}}(x) = \sqrt{2-x^2}/\pi$ has compact support over $[-\sqrt{2}, \sqrt{2}]$. Thus, $\langle \lambda_{\max} \rangle = \sqrt{2N}$ and typical fluctuations of λ_{\max} around its mean are known [21] to be TW distributed over a scale of $\sim O(N^{-1/6})$. Specifically, $\lambda_{\max} = \sqrt{2N} + a_\beta N^{-1/6} \chi$, with $a_{1,2} = 1/\sqrt{2}$, $a_4 = 2^{-7/6}$ and χ is a random variable with the TW distribution $g_\beta(\chi)$. Again, the TW form describes the PDF $P(\lambda_{\max} = t, N)$ only in the vicinity of $t = \sqrt{2N}$ over a small scale of $\sim O(N^{-1/6})$, while deviations from the TW form appear in the tails.

Fluctuations of λ_{\max} over a scale $\sim O(\sqrt{N})$ are described by large deviation functions, analogous to the Wishart case in Eq. (6) but with a different scaling variable

$$P(t, N) \sim \exp\left[-\beta N^2 \Psi_-\left(\frac{\sqrt{2N}-t}{\sqrt{N}}\right)\right] \quad t \ll \sqrt{2N};$$

$$\sim \exp\left[-\beta N \Psi_+\left(\frac{t-\sqrt{2N}}{\sqrt{N}}\right)\right] \quad t \gg \sqrt{2N}.$$

As previously mentioned, the left rate function $\Psi_-(x)$ was recently computed exactly in Ref. [22], but the right rate function $\Psi_+(x)$ was yet unknown. Our Coulomb gas approach allows to solve this problem as well and gives for $\Psi_+(x)$:

$$\Psi_+(x) = \frac{z^2 - 1}{2} - \ln(z\sqrt{2}) + \frac{1}{4z^2} G\left(\frac{2}{z^2}\right). \quad (13)$$

Here, $z = \lambda_{\max}/N = x + \sqrt{2}$, the hypergeometric function G was defined earlier and the additive constant was chosen to have $\Psi_+(0) = 0$. The asymptotics of $\Psi_+(x)$ can be easily derived: for large x , $\Psi_+(x) \sim x^2/2$, while the non-analytic

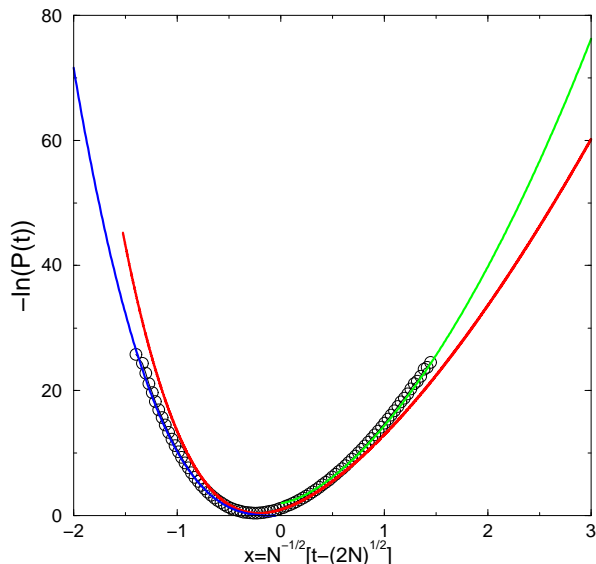


FIG. 3: Numerical results for the maximum eigenvalue distribution (circles) for $N = 10$ real ($\beta = 1$) Gaussian matrices, compared with the Tracy-Widom result (red line) and the exact right (green line) and left (blue line) large deviation functions.

behavior $\Psi_+(x) \approx 2^{7/4}x^{3/2}/3$ holds for small x . Using the TW scaling variable $\chi = (\lambda_{\max} - \sqrt{2N}) N^{1/6}/a_\beta$, with a_β defined after (12), it is easy to check that one recovers the correct TW right tails for all $\beta = 1, 2$ and 4 . This provides again a physical derivation of the TW right tail.

We have realized simulations for Gaussian matrices with sizes $N = 10, 25$ and 50 and for $\beta = 1$ and 2 . In Fig. (3) we present the data for the PDF of λ_{\max} (with $N = 10, \beta = 1$) and compare with the TW form and the exact left function Ψ_- [22] and right rate function $\Psi_+(x)$ derived in Eq. (13). As in the Wishart case, the TW form works well near the peak $t = \sqrt{2N}$, but it fails as we move into the tails, where the large deviation predictions are quite accurate.

Our Coulomb gas method lends to further generalizations that we only briefly mention here. For instance, we can compute the joint probability distribution for large fluctuations of n top eigenvalues in Wishart and Gaussian random matrices. If $n \ll N$, the energy will be given by their mutual charge interactions, the external potentials and their interaction with the unperturbed MP sea. Integrals are the same as those computed previously and yield the explicit expression for the joint PDF. It is also possible to use our method to compute the large deviation function for fluctuations of the smallest eigenvalue λ_{\min} for Wishart matrices with $c < 1$. Note that the MP sea remains unperturbed (and our method applies) for *small* fluctuations of λ_{\min} while the method in [22] applies for large fluctuations of λ_{\min} , which compress the MP sea.

In conclusion, we have presented a new Coulomb gas method to compute large deviation probabilities of top eigenvalues for a general class of random matrices. The physi-

cal picture that emerges is quite transparent: when the top eigenvalues are pulled to the right (towards large values) the Marcenko-Pastur (or Wigner) sea is simply pinched and the top eigenvalues do not drag all the other eigenvalues. In other words, no macroscopic rearrangement of the sea occurs and the top eigenvalues move in the effective potential defined by the external potential of the Coulomb gas and by the electrostatic potential generated by the charges in the sea. Our predictions are formally valid for large N yet our simulations indicate that they work for moderate N as well. This further adds to the relevance of the large deviation rate functions derived here to data compression methods and their applications.

Acknowledgments We are grateful to E. Aurell for the invitations to KTH, where this work was initiated.

-
- [1] S.S. Wilks, *Mathematical Statistics* (John Wiley & Sons, New York, 1962).
 - [2] K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Elsevier, New York, 1990).
 - [3] L.I. Smith, "A tutorial on Principal Components Analysis" (2002).
 - [4] N. Holter *et al.* Proc Natl Acad Sci USA **97**, 8409 (2000).
 - [5] O. Alter *et al.* Proc Natl Acad Sci USA **97**, 10101 (2000).
 - [6] L.-L. Cavalli-Sforza, P. Menozzi and A. Piazza, "The History and Geography of Human Genes", Princeton Univ. Press, 1994.
 - [7] J. Novembre and M. Stephens Nature Genetics **40**, 646 (2008).
 - [8] J.-P. Bouchaud and M. Potters, *Theory of Financial Risks* (Cambridge University Press, Cambridge, 2001).
 - [9] Z. Burda and J. Jurkiewicz, Physica A **344**, 67 (2004); Z. Burda, J. Jurkiewicz and B. Waclaw, Acta Physica Polonica **B 36**, 2641 (2005) and references therein.
 - [10] R.W. Preisendorfer, *Principal Component Analysis in Meteorology and Oceanography* (Elsevier, New York, 1988).
 - [11] J. Wishart, Biometrika **20**, 32 (1928).
 - [12] I.M. Johnstone, Ann. Statist. **29**, 295 (2001).
 - [13] F.J. Dyson, J. Math. Phys. **3** 140 (1962).
 - [14] M. Sadek, A. Tarighat and A.H. Sayed, IEEE Trans. Signal Processing, **55**, 1498 (2007).
 - [15] Y.V. Fyodorov and H.-J. Sommers, J. Math. Phys. **38**, 1918 (1997); Y.V. Fyodorov and B.A. Khoruzhenko, Phys. Rev. Lett. **83**, 66 (1999).
 - [16] J.J.M. Verbaarschot, Phys. Rev. Lett. **72**, 2531 (1994).
 - [17] K. Johansson, Comm. Math. Phys. **209**, 437 (2000).
 - [18] G. Schehr, S.N. Majumdar, A. Comtet and J. Randon-Furling, Phys. Rev. Lett. **101**, 150601 (2008).
 - [19] A.T. James, Ann. Math. Statistics **35**, 475 (1964).
 - [20] V.A. Marcenko and L.A. Pastur, Math. USSR-Sb, **1**, 457 (1967).
 - [21] C. Tracy and H. Widom, Comm. Math. Phys. **159**, 151 (1994); **177**, 727 (1996).
 - [22] D.S. Dean and S.N. Majumdar, Phys. Rev. Lett. **97**, 160201 (2006); Phys. Rev. E **77**, 041108 (2008).
 - [23] P. Vivo, S.N. Majumdar, and O. Bohigas, J. Phys. A: Math-Gen. **40**, 4317 (2007).
 - [24] I. Dumitriu and A. Edelman, J. Math. Phys., **43**, 5830, (2002).
 - [25] E.P. Wigner, Proc. Cambridge Philos. Soc. **47**, 790 (1951).