The QR-algorithm with Random Data: Complexity and Tracy-Widom Universality

Tibor Döme*

Department of Physics
ETH Zurich
8093 Zurich
Switzerland

September 23, 2018

To this very day, one of the most widely used eigenvalue algorithms for dense matrices is the QR-algorithm. Given the deterministic, non-ambiguous form of the QR-algorithm, the number of iterations needed to locate the smallest eigenvalue within a given tolerance is naturally a random variable, called the halting time. A weak form of universality of the distribution of the halting time for the QR algorithm can be proven mathematically, using some major results of the very young discipline of random matrix theory: In the limit of large matrices, the rescaled halting time is distributed as the Tracy-Widom-gap. In this review paper, the relation between complexity and universality for the QR algorithm is illuminated, including some open questions that were raised following the recent works of P. Deift and his collaborators.

1 Tracy-Widom Universality

The first to consider the running time fluctuations of standard eigenvalue algorithms in linear algebra have been P. Deift, G. Menon and C. Pfrang [12]. They conducted a statistical study of the performance of various eigenvalue algorithms $A \in \mathcal{A}$ applied on some real, $N \times N$ random symmetric matrix $M = (M_{ij}) \in \Sigma_N$ chosen from different ensembles $E \in \mathcal{E}$. In the following introductory arguments, the precise set of ensembles $\mathcal{E}$ is irrelevant. Such standard eigenvalue algorithms involve iterations of isospectral maps $\phi_A : \Sigma_N \to \Sigma_N$: $\mathrm{spec}(\phi_A(M)) = \mathrm{spec}(M)$. Under appropriate conditions, the sequence of matrices $M_{k+1} = \phi_A(M_k),\ k \geq 0$, with $M_0 = M$, converges to a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_N)$. Necessarily, the diagonal entries of Λ are the desired eigenvalues of M.

If an $N \times N$ matrix Y has block form

$$\begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix}, \tag{1}$$

where Y_{11} is $k \times k$ and Y_{22} is $(N-k) \times (N-k)$ for some $1 \leq k \leq N-1$, then the process of projecting $Y \to \hat{Y} = \mathrm{diag}(Y_{11}, Y_{22})$ is called *deflation*. Next, it is useful to introduce the *k-deflation time* $T^{(k)} \equiv T^{(k)}(M) \equiv T^{(k)}_{\epsilon,A,E,N}(M),\ 1 \leq k \leq N-1$: For a given precision $\epsilon > 0$, algorithm $A \in \mathcal{A}$ and matrix $M \in \Sigma_N$ sampled from the ensemble $E \in \mathcal{E}$, it is defined to be the smallest value of p such that M_p, the p-th iterate of algorithm A with $M_0 = M$, has an almost block-diagonal form

$$M_p = \begin{pmatrix} M^{(k)}_{11} & M^{(k)}_{12} \\ M^{(k)}_{21} & M^{(k)}_{22} \end{pmatrix}, \tag{2}$$

*This work was done during a *Scientific Internship* at IST Austria, supported by ERC Advanced Grant RANMAT No. 338804.

with $M_{11}^{(k)}$ of size $k \times k$, $M_{22}^{(k)}$ of size $(N-k) \times (N-k)$ and

$$||M_{12}^{(k)}||_F = ||M_{21}^{(k)}||_F \leq \epsilon. \tag{3}$$

Here, the norm introduced is the Frobenius-norm $|| \cdot ||_F$, which is defined for a $m \times n$ matrix $A = (a_{ij})_{\substack{i \in \{1,\dots,m\} \\ j \in \{1,\dots,n\}}} \in \mathbb{K}^{m \times n}$ over a field of real or complex numbers as

$$||A||_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}. \tag{4}$$

Owing to the notion of *deflation*, the name of the following object is intuitive:

Definition 1.1. (The Deflation Time)
For any given precision $\epsilon > 0$, for any $N \times N$ matrix M sampled from an ensemble E and any eigenvalue algorithm A, the deflation time is defined as

$$T \equiv T(M) \equiv T_{\epsilon,A,E,N}(M) := \min_{1 \leq k \leq N-1} T^{(k)}_{\epsilon,A,E,N}(M).$$

As soon as the algorithm reaches $T^{(\hat{k})}_{\epsilon,A,E,N}(M) := T(M)$, the eigenvalues of $\mathrm{diag}(M_{11}^{(\hat{k})}, M_{22}^{(\hat{k})})$ differ from the eigenvalues $\{\lambda_i\}$ of M by $\mathcal{O}(\epsilon)$. After running the algorithm $A \in \mathcal{A}$ to time $T(M)$, the matrix $M^{(\hat{k})}$ is *deflated* and the process restarts by applying the same iteration procedure A separately to the smaller matrices $M_{11}^{(\hat{k})}$ and $M_{22}^{(\hat{k})}$ until some off-diagonal blocks are again of order $\mathcal{O}(\epsilon)$. Generically, at worst $\mathcal{O}(N)$ deflations are needed to compute the eigenvalues $\{\lambda_i\}$ of M and at best, $\mathcal{O}(\log N)$, using parallelization as explained in [12]. However, not all eigenvalues are computed to precision $\mathcal{O}(\epsilon)$ by this procedure. In fact, the worst possible order to which any of the eigenvalues may be approximated is $\mathcal{O}((N-1)\epsilon)$: In the case where the deflation index is $\hat{k} = 1$ throughout, the algorithm needs $N-1$ steps to locate the eigenvalue λ_N, accumulating an error $\mathcal{O}(\epsilon)$ in each step.

The *fluctuation* of $T_{\epsilon,A,E,N}(M)$ with respect to T is defined by

$$\tau_{\epsilon,A,E,N}(M) = \frac{T_{\epsilon,A,E,N}(M) - \langle T_{\epsilon,A,E,N} \rangle}{\sigma_{\epsilon,A,E,N}}, \tag{5}$$

where $\langle T_{\epsilon,A,E,N} \rangle$ is the average of $T_{\epsilon,A,E,N}(M)$ over the ensemble $\mathcal{E}$ and $\sigma^2_{\epsilon,A,E,N}$ is the variance. In the context of numerical experiments, the average and the variance become the sample average and the sample variance over finitely many realizations of random matrices. In [12], the authors discovered the following phenomenon: For a given precision ϵ, a given matrix size N within an appropriate scaling range and a given algorithm A, the *fluctuations* in the time to compute the eigenvalues are universal, independent of the choice of the ensemble E. Hence, up to the two components $\langle T_{\epsilon,A,E,N} \rangle$ and $\sigma_{\epsilon,A,E,N}$, the statistics of the deflation time are universally prescribed. The majority of calculations in [12] concerned three eigenvalue algorithms: the QR algorithm, the QR algorithm with shifts, which is used in practice, and the Toda algorithm. However, it is conjectured that this *two-component universality*, i.e. universality given that the sample average and the sample variance are known, emerges on some appropriate scale whenever a standard eigenvalue algorithm with random data is run on a *physical system*. This *physical system* might be the electronic machine on our desk, but also the neuroscientists' version of the human brain which runs algorithms to, for instance, classify random visual objects (see [2]).

To the best of my knowledge, a rigorous proof of two-component universality does not yet exist! All existing proofs are restricted to calculating extremal eigenvalues of random matrices sampled from some given ensemble. In [7], P. Deift and T. Trogdon proved *weak* universality for the Toda algorithm, emerging when computing the largest eigenvalue of real symmetric and complex Hermitian random matrices. In this context, the property of *weak* universality of an algorithm shall denote that universality arises for the distribution of the rescaled running time for finding an extremal eigenvalue (the

smallest or the largest) of the random matrix. Correspondingly, *strong* universality corresponds to a (a priori not necessarily eigenvalue-independent) universal limiting distribution of the rescaled running time for all eigenvalues. It is important to note that an algorithm might have *strong* universality even though only the *weak* form has been proven.

For the QR algorithm and the QR algorithm with shifts, *weak* universality can be proven simultaneously since shifting does not affect universality, as shown in [12]. Again, this has been done by P. Deift and T. Trogdon, this time in another paper [6]. To understand their results, the non-ambiguous QR algorithm needs to be defined first, as it has been proposed independently by J. G. F. Francis and W. N. Kublanowskaja in 1961/1962 :

Definition 1.2. (QR algorithm)
Input: H *and* $\epsilon > 0$
Output: An approximation of the spectrum of H
iteration procedure:
set $X = H$*;*
while $f(X) > \epsilon$ *do*
set $(Q, R) := \mathrm{QR}(X)$*;*
set $X = RQ$*;*
end while
return $[X_{11}, X_{22}, ..., X_{NN}]^T$

Here, $f(H) \geq 0$ plays the role of a general error control function and $\mathrm{QR}(X)$ ($H = QR$) denotes the QR factorization where Q is unitary and R is upper-triangular with *positive* diagonal entries. In [6], the results about the QR algorithm to compute the smallest eigenvalue of H apply for the following set of matrices:

Definition 1.3. (Sample Covariance Matrix (SCM))
A sample covariance matrix is a real symmetric ($\beta = 1$) *or complex Hermitian* ($\beta = 2$) *matrix* $H = \frac{1}{M} V^* V$, $V = (V_{ij})_{\substack{1 \leq i \leq M \\ 1 \leq j \leq N}}$, *such that* V_{ij} *are independent random variables given by a probability measure* ν_{ij} *with*

$$\mathbb{E} V_{ij} = 0, \quad \mathbb{E}|V_{ij}|^2 = 1. \tag{6}$$

Next, $\exists\ \nu$ *(independent of* N, M, i, j*) such that* $|V_{ij}|$ *decays sub-exponentially:*

$$\mathbb{P}(|V_{ij}| > x) \leq \nu^{-1} \exp(-x^{\nu}),\ x > 1. \tag{7}$$

For $\beta = 2$ *the condition*

$$\mathbb{E} V_{ij}^2 = 0 \tag{8}$$

must also be satisfied.

SCMs have been analyzed within the framework of random matrix theory (RMT) several decades ago. Let us assume that all SCMs have $M \geq N$ such that $d_N := \frac{N}{M}$ shall satisfy $\lim_{N \to \infty} d_N =: d \in (0, 1)$. In 1967, V. A. Marcenko and L. A. Pastur proved in [11] that the averaged empirical spectral measure

$$\mu_N(z) = \mathbb{E} \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda_i - z) \tag{9}$$

converges to

$$\rho_d(x) dx := \frac{1}{2\pi d} \sqrt{\frac{[(\lambda_+ - x)(x - \lambda_-)]_+}{x^2}} dx, \quad \lambda_\pm = (1 \pm \sqrt{d})^2. \tag{10}$$

The *equilibrium measure* $\rho_d(x)$ has the following quantiles:

Definition 1.4. (Quantiles)
The quantile γ_n is defined to be the smallest value of t such that

$$\frac{n}{N} = \int_{-\infty}^{t} \rho_d(x)dx, \ n \in \{1, ..., N\}. \tag{11}$$

However, the *global law* (10) is only valid on the scale $\mathcal{O}(1)$ where one detects the cumulative effect of cN eigenvalues with $c > 0$. In fact, the global law (10) can be extended down to a mesoscopic scale where one detects $\mathcal{O}(N^{\epsilon})$ many eigenvalues for $\epsilon > 0$. This leads to the concept of *local laws*. For the case of SCMs, the interested reader is referred to [3]. The behavior of the eigenvalues on a much smaller scale $\mathcal{O}(\frac{1}{N})$, where individual eigenvalues are detected, has been a central topic in RMT for several years and it turned out that

$$N^{2/3}\lambda_{-}^{-2/3}d^{1/2}(\lambda_i - \lambda_{-}) \xrightarrow{N\to\infty} \Lambda_{i,\beta}, \ i \in \{1, 2, 3\}, \tag{12}$$

with $\Lambda_{i,\beta}$ being the smallest three eigenvalues of the stochastic Airy operator. The definition of the stochastic Airy operator and a modern proof of the above theorem (12) can be found in [13]. Using these results and many more, P. Deift and T. Trogdon proved the following form of weak universality for the QR algorithm in [6]:

Theorem 1.1. (Weak Universality for the QR algorithm on SCMs)
Let H be a real ($\beta = 1$) or complex ($\beta = 2$) $N \times N$ sample covariance matrix. Assuming ϵ satisfies

$$\frac{\alpha}{2} := \frac{\log \epsilon^{-1}}{\log N} \geq \frac{5}{3} + \frac{\sigma}{2} \tag{13}$$

for $0 < \sigma < \frac{1}{3}$ fixed, then for $t \in \mathbb{R}$, the $(N-1)$-deflation time $T^{(N-1)}_{\epsilon,QR,SCM,N}$ fulfills

$$F^{gap}_{\beta}(t) := \lim_{N\to\infty} \mathbb{P}\left(\frac{1}{\Lambda_{2,\beta} - \Lambda_{1,\beta}} \leq t\right) = \lim_{N\to\infty} \mathbb{P}\left(\frac{T^{(N-1)}_{\epsilon,QR,SCM,N}}{\lambda_{-}^{1/3}d^{1/2}N^{2/3}(\log \epsilon^{-1} - (2/3)\log N)} \leq t\right) \tag{14}$$

1.1 Remarks on Theorem 1.1

In simple terms, Theorem 1.1 states that after appropriate (and natural) rescalings, the $(N-1)$-deflation time $T^{(N-1)}_{\epsilon,QR,SCM,N}$ has the asymptotics

$$T^{(N-1)}_{\epsilon,QR,SCM,N} \sim N^{2/3}\log(N)\frac{1}{\xi}, \tag{15}$$

where $\xi = \xi_{\beta}$ is the random variable distributed as the Tracy-Widom-gap (TW-gap).

The $(N-1)$-deflation time is the central object of Theorem 1.1. It is interpreted as the running time of the QR algorithm to locate, up to precision ϵ, the eigenvalue λ_N in the matrix $\mathrm{diag}(\lambda_1, ..., \lambda_N)$ to which the iterates of the $N \times N$ SCM H converge. The reason why, for the QR algorithm, it delivers the smallest eigenvalue of the matrix H is that the QR algorithm orders the set of eigenvalues in the iterates $X_k \xrightarrow{N\to\infty} \Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_N)$, $X_0 = H$, in decreasing order. In other words, the non-negative $\{\lambda_i\}$ (recall that H is positive semidefinite) obey $\lambda_1 \geq ... \geq \lambda_N \geq 0$. This can be easily proven by reverting to the famous *power iteration* method that is used to find the largest eigenvalue of H and can be found in [10]. In fact, the QR algorithm is a power iteration method in disguise! Universality for the power method has also been worked out in [6], leading to a TW-gap distribution as well. The same analytical form of the limiting distribution of the running time fluctuations for both the QR algorithm and the power method is thus nothing but expected.

Fortunately, one can roughly predict the running time for the power method and hence, via the above argument, also for the QR algorithm: Let us assume that $A \in \mathbb{C}^{N\times N}$ is a complex matrix which is, for simplicity, diagonalizable, having eigenvectors $\vec{u}_i \in \mathbb{C}^N$, $i \in \{1, ..., N\}$. Without loss of generality all eigenvalues shall be simple. Next, the normalized starting vector $\vec{v}_0 \in \mathbb{C}^N$ should have a non-zero projection onto the space spanned by the eigenvector $\vec{u_1}$ corresponding to the largest eigenvalue λ_1, where $\lambda_1 \geq \lambda_2 ... \geq \lambda_N$, i.e. $\langle \vec{v}_0, \vec{u_1}\rangle \neq 0$. The interesting sequence $\vec{v}_k = A^k\vec{v}_0$, $k \in \mathbb{N}_0$, is defined via powers of A and it fulfills

$$\langle \vec{v}_0, A^n\vec{v}_0\rangle = \sum_{i=1}^{N} \lambda_i^n |\langle \vec{v}_0, \vec{u_1}\rangle|^2 = \lambda_1^n \left(|\langle \vec{v}_0, \vec{u_1}\rangle|^2 + \sum_{i=2}^{N} \left(\frac{\lambda_i}{\lambda_1}\right)^n |\langle \vec{v}_0, \vec{u_i}\rangle|^2 \right). \tag{16}$$

The second term can only be neglected iff $(N-1)\left(\frac{\lambda_i}{\lambda_1}\right)^n \ll 1$, which necessarily leads to

$$N\left(\frac{\lambda_2}{\lambda_1}\right)^n \ll 1 \iff n \gg \frac{\log N}{\log \lambda_1 - \log \lambda_2}. \tag{17}$$

Of course, the rough estimate (17) cannot predict the emergence of a TW-gap distributed random variable ξ as in (15), but the large N behavior of the running time is evident.

The idea of the proof given in [6] will be the focus of the next section. In a non-formal way, I also try to highlight some details that have been swept under the carpet in [6].

2 Proof Idea of Theorem 1.1

First of all, the QR algorithm in its general form is ambiguous since the QR decomposition $\mathrm{QR}(X)$ on which it is based is also ambiguous. To circumvent this, one needs to impose that the upper-triangular matrix R has positive diagonal entries, as has been done in Definition 1.2. This way, the QR factorization, which always exists, is unique. In particular, this property is used when developing an *interpolated version* of the QR algorithm. The latter procedure is conjectured to be the crucial step when proving universality for eigenvalue algorithms that are iterative and thus discrete in nature. In [8], the same two authors as above have even granted this conjecture a name, though, to the best of my knowledge, it has never been proven: *the stroboscope theorem*. For SCMs H in the case of the QR algorithm, one can find an interpolation of the iterates X_n, $n \in \mathbb{N}_0$, by letting H^t denote the t-th power of H, $t \geq 0$. One then defines $Q(t)$, $R(t)$ and $X(t)$ via

$$\begin{aligned} (Q(t), R(t)) &:= \mathrm{QR}(H^t), \\ H^t &= Q(t)R(t), \\ X(t) &:= Q^*(t)HQ(t). \end{aligned} \tag{18}$$

The interpolation character becomes evident in the following

Lemma 2.1. (stroboscope theorem for the QR algorithm) $\forall n \in \mathbb{N}_0$, $X(n) = X_n$.

Of course, the 0-th iterate is $X(0) = X_0 = H$. The above lemma can be proven in a straightforward way using the uniqueness of the QR factorization. It is important to note that $Q(t)$ and $R(t)$ are infinitely differentiable matrix-valued functions of t. The subsequent proof idea of Theorem 1.1 specifically uses the differentiability of $X(t)$.

The best choice for the error function f in Definition 1.2 is

$$f(X) = f(X(t)) = \sqrt{\sum_{n=1}^{N-1} |X_{Nn}(t)|^2}. \tag{19}$$

Instead of estimating the true error $\lambda_N - X_{NN}$ in computing λ_N, one uses

$$E_{QR}(t) := f^2(X(t)) \tag{20}$$

to determine a convergence criterion since it is easily observable: Indeed, if $E_{QR}(t) < \epsilon^2$, then $|X_{NN}(t) - \lambda_j| < \epsilon$ for some eigenvalue λ_j of $X(0)$. This follows from performing a perturbation of the matrix containing 0-s in the last row and last column except for an X_{NN} in the N, N-entry. In the limit $N \to \infty$, $\mathbb{P}(\lambda_j = \lambda_N) = 1$. The latter is not a trivial fact: The difference between the smallest two eigenvalues is known to fulfill $\mathbb{P}\left((\lambda_{N-1} - \lambda_N) \ll N^{-2/3}\right) = o(1)$. Hence for $\epsilon \ll N^{-2/3}$, with high probability there can only exist one eigenvalue in an ϵ-neighborhood of $X_{NN}(t)$, which has to be the smallest one λ_N.

One then defines the *halting time* $T_{QR,\epsilon}$ for the QR algorithm to be the interpolated version of the $(N-1)$-deflation time:

Definition 2.1. (The halting time) $T_{QR,\epsilon} \equiv T_{QR,\epsilon}(H) := \inf\{t : E_{QR}(t) \leq \epsilon^2\}$.

Here, one does not assume that t is an integer. The last step of the proof will consist of showing (though it is not done in [6]) that the "true" halting time $\lceil T_{QR,\epsilon} \rceil = T^{(N-1)}_{\epsilon,QR,SCM,N}$ has the same limiting distribution as $T_{QR,\epsilon}$.

Now, it is useful to introduce conditions on the SCM H that simplify the analysis: Given a SCM H, let v be a random (or deterministic) unit vector independent of H. Define $\mu_n = |\langle v, u_n \rangle|$, $n \in \{1, ..., N\}$, where u_n is the n-th eigenvector of H.

Condition 2.1. *For any fixed* $0 < s < \frac{\sigma}{40}$,

1) *Quantum Unique Ergodicity (QUE):* $\mu_n \leq N^{-1/2+s/2}$ $\forall$ n,

2) *Quantum Unique Ergodicity (QUE):* $N^{-1/2-s/2} \leq \mu_n$ *for* $n \in \{1, 2, N-1, N\}$,

3) *Right Edge Behavior:* $N^{-2/3-s/2} \leq \lambda_N - \lambda_{n-1} \leq N^{-2/3+s/2}$ *for* $n \in \{N-1, N\}$,

4) *Left Edge Behavior:* $N^{-2/3-s/2} \leq \lambda_n - \lambda_1 \leq N^{-2/3+s/2}$ *for* $n \in \{2, 3\}$, *and*

5) *Rigidity:* $|\lambda_n - \gamma_n| \leq N^{-2/3+s/2}(\min\{n, N-n+1\})^{-1/3}$ $\forall$ n.

Let $\mathcal{R}_{N,s}$ *denote the set of SCMs* H *that satisfy these conditions.*

These conditions are not arbitrary at all. It has been precisely the achievement of random matrix theorists in the last few decades to show that the following holds:

Theorem 2.1.

$$\lim_{N\to\infty} \mathbb{P}(\mathcal{R}_{N,s}) = 1. \tag{21}$$

The key point that is exploited in the proof is the connection between the $(N-1)$-deflation time and RMT: If the finite-sized $N \times N$ random SCM H of interest lies in the set $\mathcal{R}_{N,s}$ described by Condition 2.1, then its random eigenvalues and eigenvectors, which are related to the halting time, can be analyzed using RMT even without sending $N \to \infty$. A significant part of the proof in [6] requires expressions that are explicit in N. Only towards the end will indeed $N \to \infty$ and Theorem 2.1 be put to use.

Using the notation

$$\delta_n := \frac{\lambda_1^2}{\lambda_n^2},\ \Delta_n := \lambda_n - \lambda_1,\ \beta_n = |U_{Nn}(0)|,\ X(t) =: U(t)\Lambda U^*(t),\ \Lambda := \mathrm{diag}(\lambda_1, ..., \lambda_N),\ \nu_n := \frac{\beta_n^2}{\beta_1^2} \tag{22}$$

for $n \in \{1, ..., N\}$, the main steps of the proof can be formulated as follows:

1) **Lemma 2.2.** (estimate on $T_{QR,\epsilon}$)

Given Condition 2.1, N large: $\left(\alpha - \frac{4}{3} - 5s\right)\frac{\log N}{\log \delta_2^{-1}} \leq T_{QR,\epsilon} \leq \left(\alpha - \frac{4}{3} - 5s\right)\frac{\log N}{\log \delta_2^{-1}}$

Proof Idea:

(a) rigidity
(b) relevant edge scale: $N^{-2/3}$
(c) $\sqrt{}$-decay of $\rho_d(x)$ at both edges (see Eq. (10))

2) **Lemma 2.3.** (lower bound on $-E'_{QR}(t)$)

Given Condition 2.1, $t \in L_\alpha := \left[\left(\alpha - \frac{4}{3} - 5s\right)\frac{\log N}{\log \delta_2^{-1}}, \left(\alpha - \frac{4}{3} - 5s\right)\frac{\log N}{\log \delta_2^{-1}}\right]$, *N large, and* $C > 0$: $-E'_{QR}(t) \geq CN^{-12s-\alpha-2/3}$

Proof Idea:

(a) rigidity
(b) relevant edge scale: $N^{-2/3}$
(c) $\sqrt{}$-decay of $\rho_d(x)$ at both edges (see Eq. (10))
(d) differentiability of $X(t)$

3) **Lemma 2.4.** (estimate on $T^*_{QR,\epsilon}$)

Given Condition 2.1, *N large and* $T^*_{QR,\epsilon}$ *defined by* $\Delta_2^2 \delta_2^{T^*_{QR,\epsilon}} \nu_2 := \epsilon^2$:
$\left(\alpha - \frac{4}{3} - 4s\right)\frac{\log N}{\log \delta_2^{-1}} \leq T^*_{QR,\epsilon} \leq \left(\alpha - \frac{4}{3} + 4s\right)\frac{\log N}{\log \delta_2^{-1}}$, *i.e.* $T^*_{QR,\epsilon} \in L_\alpha$

Proof Idea: steps 1.) and 2.)

4) **Lemma 2.5.** ($T^*_{QR,\epsilon}$ is close to $T_{QR,\epsilon}$)

$N^{-2/3}|T^*_{QR,\epsilon} - T_{QR,\epsilon}| \xrightarrow{\text{in prob.}} 0$

Proof Idea: steps 1.), 2.) and 3.)

5) **Lemma 2.6.** ($T^*_{QR,\epsilon}$ converges in distribution)

$\left|\frac{T^*_{QR,\epsilon}}{N^{2/3}\log N} - \frac{\alpha - \frac{4}{3}}{N^{2/3}\log \delta_2^{-1}}\right| \xrightarrow{\text{in prob.}} 0$, *i.e.* $\lim_{N\to\infty} \mathbb{P}\left(\frac{T^*_{QR,\epsilon}}{(\alpha/2-2/3)\lambda_-^{1/3} d^{1/2} N^{2/3}\log N} \leq t\right) = F_\beta^{\text{gap}}(t)$

Proof Idea:

(a) Eq. (12)
(b) if $X_N \xrightarrow{\text{in distr.}} X$ and $|X_N - Y_N| \xrightarrow{\text{in prob.}}$, then $Y_N \xrightarrow{\text{in distr.}} X$; here: $X_N = \frac{(\alpha-4/3)\log N}{\log \delta_2^{-1}}$ and $Y_N = T^*_{QR,\epsilon}$
(c) continuity of $F_\beta^{\text{gap}}(t)$
(d) Theorem 2.1

6) **Lemma 2.7.** ($T_{QR,\epsilon}$ converges in distribution)

$\lim_{N\to\infty} \mathbb{P}\left(\frac{T_{QR,\epsilon}}{\lambda_-^{1/3} d^{1/2} N^{2/3}(\alpha/2-2/3)\log N} \leq t\right) = F_\beta^{\text{gap}}(t)$

Proof Idea:

(a) if $X_N \xrightarrow{\text{in distr.}} X$ and $|X_N - Y_N| \xrightarrow{\text{in prob.}}$, then $Y_N \xrightarrow{\text{in distr.}} X$; here: $X_N = T^*_{QR,\epsilon}$ and $Y_N = T_{QR,\epsilon}$
(b) step 4.)

7) **Theorem 1.1.** ($T^{(N-1)}_{\epsilon,QR,SCM,N}$ converges in distribution)

$\lim_{N\to\infty} \mathbb{P}\left(\frac{T^{(N-1)}_{\epsilon,QR,SCM,N}}{\lambda_-^{1/3} d^{1/2} N^{2/3}(\log \epsilon^{-1} - 2/3 \log N)} \leq t\right) = F_\beta^{\text{gap}}(t)$

Last Step: Since $\text{Var}[T_{QR,\epsilon}]$ diverges, $\lceil T_{QR,\epsilon} \rceil = T^{(N-1)}_{\epsilon,QR,SCM,N}$ has a limiting distribution and it is necessarily the same as that of $T_{QR,\epsilon}$.

The above steps are straightforward, for details see [6].

3 Computational Complexity of the QR algorithm

From the viewpoint of computational complexity theory, one would be naturally interested in the complexity of the QR algorithm. In terms of floating point operations (FLOPs), the answer is already known, but it depends on the numerical realization of the QR algorithm. The most efficient QR algorithm to find the eigenvalues of a real symmetric ($\beta = 1$) matrix $M \in \Sigma_N$ starts with the transformation of M into a tridiagonal matrix, preferably by using (at most) $N-2$ Householder reflections. It is well known that the reduction of M into tridiagonal form costs $\mathcal{O}(N^3)$ FLOPs. Tridiagonal matrices (or, more generally, Hessenberg matrices which have zero entries below the first subdiagonal) can be proven to be preserved by all versions of the QR algorithm (see [1]), i.e. if $\mathrm{QR}(M_k) = QR$, then $M_{k+1} = RQ$ is tridiagonal (or of Hessenberg form). An iteration step now consists of eliminating all $N-1$ elements of M_k that lie on the lower diagonal, this time by applying (at most) $N-1$ Givens rotations, to find R. M_{k+1} is then easily obtained by $M_{k+1} = RQ$, where Q is just a product of Givens rotations. Each iteration step from M_k to M_{k+1} costs $\mathcal{O}(N^2)$ FLOPs. In [1], there is a beautiful derivation of this result.

The above results on the scaling behavior indicate that the QR factorization with Householder reflections and Givens rotations lies in the complexity class **P**. In fact, it is **P-complete** (see [9]). Whether the QR algorithm itself is also **P-complete** cannot be deduced by that result. To the best of my knowledge, the complexity class of the QR algorithm is not known. Of course, one also wants to know the complexity of the algorithm on a smaller, *Law of Large Numbers*-like scale. In particular, one seeks the answer to the question how long it takes the algorithm to find *all* the eigenvalues λ_i given a certain joint precision ϵ. Theorem 1.1 does not give an answer to that, but it is an important first step towards its determination. It states that the $(N-1)$-deflation time scales like

$$T^{(N-1)}_{\epsilon,QR,SCM,N} \sim \left(\alpha - \frac{2}{3}\right) N^{2/3} \log N \tag{23}$$

in order to obtain a precision of $\epsilon = N^{-\alpha/2}$. In other words, Theorem 1.1 gives the speed of convergence for finding λ_N, but gives no information about how fast the algorithm locates the other λ_i. When producing a histogram of the distribution of $\hat{k}$, as it has been done in [6], one finds that $\hat{k} = N-1$ occurs with the largest probability. Hence, indeed, the first eigenvalue the algorithm provides is the smallest one λ_N and that is precisely the one for which the time of convergence is known. It is important to note that the complexity of the QR algorithm, which is described by a *Law of Large Numbers*-like statement for the time of convergence, is an intriguing topic which is very poorly understood. First, the time until, for any given $i \in \{1, ..., N\}$, the eigenvalue λ_i is located up to precision ϵ_i needs to be examined. Secondly, the time until the very first λ_j, j being dependent on the particular realization of the SCM H, is located up to ϵ_0 is a much more challenging object, corresponding to the deflation time T. And third, the time the algorithm needs to locate all the eigenvalues $\lambda_{k(i)}$, $k(i) \in \{1, ..., N\}$, one by one, deflation by deflation, up to $\epsilon_{k(i)}$ needs to be studied. However, the emergence of universality, which is a *Central Limit Theorem*-like statement, is a completely independent phenomenon which occurs on a much smaller scale.

4 Outlook

The main open question that remains is related to the asymptotics of the time of full deflation $T = \min_{1 \le k \le N-1} T^{(k)}_{\epsilon,A,E,N}$. Proving theorems about T is particularly difficult as one has to analyze the minimum of $N-1$ correlated random variables. P. Deift and T. Trogdon conjecture that T has a sub-Gaussian limiting distribution which is related to the largest gap in the spectrum of the stochastic Airy operator. Another open question is whether theorems similar to Theorem 1.1 can be extended to other algorithms, such as the *Jacobi algorithm*, the *Conjugate Gradient Algorithm* or the *GMRES Algorithm* (these are all defined in [5]). Even more interesting problems are genuinely infinite dimensional. For instance, one may want to theoretically understand an iterative algorithm to solve the Dirichlet

problem $\Delta u = 0$ in a random star-shaped region $\Omega \subset \mathbb{R}^2$ with random boundary data f on $\partial\Omega$. Though a numerical evidence for the latter is available (see [5]), no theoretical analysis exists, to the best of my knowledge.

Strictly speaking, Theorem 1.1 is not even a two-component universality result for the $(N-1)$-deflation time $T^{(N-1)}_{\epsilon,QR,SCM,N}$. For the latter would be equivalent to the following statement:

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{T^{(N-1)}_{\epsilon,QR,SCM,N} - \langle T^{(N-1)}_{\epsilon,QR,SCM,N} \rangle}{\sigma_{T^{(N-1)}_{\epsilon,QR,SCM,N}}} \leq t \right) = F^{\text{gap}}_{\beta}(\kappa t + \nu) \tag{24}$$

for some $\kappa \in \mathbb{R}$ and some $\nu \in \mathbb{R}$. Preferably, $\nu = \mathbb{E}[\xi]$ and $\kappa^2 = \text{Var}[\xi]$, with $\xi = \xi_\beta$ being the random variable with distribution $F^{\text{gap}}_{\beta}(t)$, $\beta \in \{1, 2\}$. This would be a true two-component universality theorem for the $(N-1)$-deflation time as the limiting distribution would not depend on the individual entries of the matrix ensemble, just whether it is real or complex. For the Toda algorithm, such a statement has been partially proven by P. Deift and T. Trogdon in [7] by relying on an extension of level repulsion estimates found in [4]. For the moment, validating Eq. (24) would be the next step towards the rigorous proof of genuine two-component universality for the QR algorithm. $lambda_N \ll N^{-2/3}$

Acknowledgement

First and foremost, I would like to thank Prof. Dr. László Erdős for his hospitality at IST Austria, his introduction to the field of random matrices, their connection to eigenvalue algorithms and, above all, for providing fruitful discussions on a wide range of topics including many more than those just mentioned.

References

[1] P. Arbenz. Lecture notes on large scale eigenvalue problems. *ETH Zürich Lecture: Numerical Methods for Solving Large Scale Eigenvalue Problems*, April 2016.

[2] Y. Bakhtin and J. Correll. A neural computation model for decision-making times. *Journal of Mathematical Psychology*, 56(5):333–340, October 2012.

[3] A. Bloemendal, , L. Erdös, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized wigner matrices. *Electron. J. Probab.*, 19:53 pp., June 2014.

[4] P. Bourgade, L. Erdös, and H.-T. Yau. Edge universality of beta ensembles. *Communications in Mathematical Physics*, 332(1):261–353, November 2014.

[5] P. Deift, G. Menon, S. Olver, and T. Trogdon. Universality in numerical computation with random data. Case studies. *Proceedings of the National Academy of Sciences*, 111(42):14973–14978, July 2014.

[6] P. Deift and T. Trogdon. Universality for eigenvalue algorithms on sample covariance matrices. *SIAM Journal on Numerical Analysis*, 55, January 2017.

[7] P. Deift and T. Trogdon. Universality for the toda algorithm to compute the largest eigenvalue of a random matrix. *Communications on Pure and Applied Mathematics*, 71(3):505–536, August 2017.

[8] P. Deift and T. Trogdon. Universality in numerical computation with random data. Case studies, analytic results and some speculations. *Proceedings of the National Academy of Sciences*, 55, March 2017.

[9] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou. Communication-optimal parallel and sequential qr and lu factorizations. *SIAM Journal on Scientific Computing*, 34(1):A206–A239, February 2012.

[10] P. Deuflhard and A. Hohmann. Numerische Mathematik. Eine algorithmisch orientierte Einführung. *de Gruyter*, June 1993.

[11] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[12] G. Menon. How long does it take to compute the eigenvalues of a random symmetric matrix? *Banff International Research Station for Mathematical Innovation and Discovery*, April 2016.

[13] Natesh S. Pillai and Jun Yin. Universality of covariance matrices. *Ann. Appl. Probab.*, 24(3):935–1001, April 2014.