Spring 2019

Universality in numerical computation. A case study: Eigenvalue computation

References: Insert

Lecture 1

About 8 or 9 years ago, Christian Pfrang, Govind Menon and P.D. ([PDM], published 2014) initiated a statistical study of the performance of various standard algorithms A to compute the eigenvalues of random real symmetric matrices H. Let Σ_N denote the set of real $N \times N$ symmetric matrices. Associated with each algorithm A, there is, in the discrete case such as QR, a map $\varphi = \varphi_A : \Sigma_N \to \Sigma_N$ with the properties

- (isospectral) $\mathrm{spec}(\varphi_A(H)) = \mathrm{spec}(H)$
- (convergence) The iterates $X_{k+1} = \varphi_A(X_k)$, $k \geq 0$, $X_0 = H$ given, converge to a diagonal matrix X_∞, $X_k \to X_\infty$ as $k \to \infty$.

and in the continuum case, such as Toda, there is a flow

$t \mapsto X(t) \in \Sigma_N$ with the properties

- (isospectral) $\operatorname{spec}(X(t))$ is constant
- (convergence) the flow $X(t)$, $t \geq 0$, $X(0) = H$ given, converges to a diagonal matrix X_∞, $X(t) \to X_\infty$ as $t \to \infty$

In both cases, necessarily, the (diagonal) entries of X_∞ are the eigenvalues of the given matrix H.

Given $\varepsilon > 0$, it follows, in the discrete case, that for some m the off-diagonal entries of X_m are $O(\varepsilon)$ and hence the diagonal entries of X_m give the eigenvalues of $X_0 = H$ to $O(\varepsilon)$. The situation is similar for continuous algorithms $t \mapsto X(t)$. Rather than running the algorithm until all the off-diagonal entries are $O(\varepsilon)$, it is customary to run the algorithm with <u>deflations</u> as follows, For an $N \times N$ matrix Y in block form

$$Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

with Y_{11} of size $k \times k$ and Y_{22} of size $N-k \times N-k$ for some $k \in \{1, 2, \ldots, N-1\}$, the process of projecting $Y \mapsto \mathrm{diag}(Y_{11}, Y_{22})$ is called deflation. For a given ε, algorithm A and matrix $H \in \Sigma_N$, define the <u>k-deflation</u> time $T^{(k)}(H) = T^{(k)}_{\varepsilon, A}(H)$, $1 \le k \le N-1$, to be the smallest value of m such that X_m, the m^{th} iterate of algorithm A with $X_0 = H$, has block form

$$X_m = \begin{bmatrix} X_{11}^{(k)} & X_{12}^{(k)} \\ X_{21}^{(k)} & X_{22}^{(k)} \end{bmatrix}$$

with $X_{11}^{(k)}$ of size $k \times k$ and $X_{22}^{(k)}$ of size $N-k \times N-k$ and $\| X_{12}^{(k)} \| = \| X_{21}^{(k)} \| \le \varepsilon$. The deflation time $T(H)$ is then defined as

$$T(H) = T_{\varepsilon, A}(H) = \min_{1 \le k \le N-1} T^{(k)}_{\varepsilon, A}(H) .$$

If $\hat{k} \in \{1, \dots, N-1\}$ is such that $T(H) = T^{(k)}_{\varepsilon, A}(H)$, it follows that the eigenvalues of $H = X_0$ are given by the eigenvalues of the block-diagonal matrix $\mathrm{diag}(X^{(k)}_{11}, X^{(k)}_{22})$ to $O(\varepsilon)$. After running the algorithm to time $T_{\varepsilon, A}(H)$, the algorithm restarts by applying the basic algorithm A separately to the smaller matrices $X^{(k)}_{11}$ and $X^{(k)}_{22}$ until the next deflation time, and so on. There are again similar considerations for continuous algorithms.

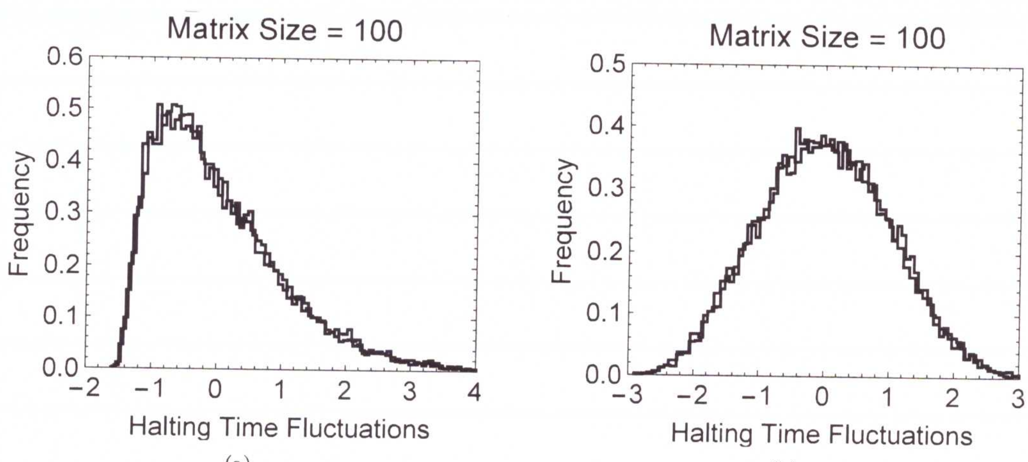
As the algorithm proceeds, the number of matrices after each deflation doubles. This is ~~counter~~ counterbalanced by the fact that the matrices are smaller and smaller in size, and the calculations are clearly parallelizable. Allowing for parallel computation, the number of deflations to compute all the eigenvalues of a given matrix H to a given accuracy ε, will vary from

$O(\log N)$ to $O(N)$.

In the work of Pfrang et al the authors considered the deflation time $T = T_{\varepsilon, A}$ for $N \times N$ matrices chosen from a given ensemble $\mathcal{E}$. Henceforth in these lectures we suppress the dependence on ε, N, A and $\mathcal{E}$, and simply write T with these variables understood. For a given algorithm A and ensemble $\mathcal{E}$, the authors computed $T(H)$ for 5,000 - 15,000 samples of matrices H chosen from $\mathcal{E}$, and recorded the normalized deflation time

$$\tilde{T}(H) \equiv \frac{T(H) - \langle T \rangle}{\sigma}$$

where $\langle T \rangle$ and $\sigma^2 = \langle (T - \langle T \rangle)^2 \rangle$ are the sample average and sample variance of $T(H)$, respectively. Surprisingly, the author found that for the given algorithm A, and ε and N in a suitable scaling range with $N \to \infty$, the histogram was of $\tilde{T}$ was <u>universal</u>, independent of the ensemble $\mathcal{E}$. In other words, the fluctuations in the deflation time $\tilde{T}$, suitably scaled, were universal, independent of $\mathcal{E}$. The following figure displays, in a slightly different form, some of the numerical results from Pfrang et al. Figure 1 (a) displays data for the QR algorithm, which is discrete, and Figure 1 (b) displays data for the Toda algorithm, which is continuous.



(a) (b)

Figure 1: Universality for $\tilde{T}$ when (a) $\mathcal{A}$ is the QR eigenvalue algorithm and when (b) $\mathcal{A}$ is the Toda algorithm. Panel (a) displays the overlay of two histograms for $\tilde{T}$ in the case of QR, one for each of the two ensembles $\mathcal{E}$ = BE, consisting of iid mean-zero Bernoulli random variables (see Definition A.1) and $\mathcal{E}$ = GOE, consisting of iid mean-zero normal random variables. Here $\epsilon = 10^{-10}$ and $N = 100$. Panel (b) displays the overlay of two histograms for $\tilde{T}$ in the case of the Toda algorithm, and again $\mathcal{E}$ = BE or GOE. And here $\epsilon = 10^{-8}$ and $N = 100$.

Subsequently, Govind Menon, Sheehan Olver, Tom Trogdon and I [DMOT] raised the question of whether the universality results of Pfrang et al. were limited to eigenvalue algorithms for real symmetric matrices, or whether they were present more generally in numerical computation. And, indeed, they found similar universality results for a wide variety of numerical algorithms, including

(a) other algorithms such as the QR algorithm with shifts, the Jacobi eigenvalue algorithm, and also algorithms applied to complex Hermitian ensembles

(b) the conjugate gradient and GMRES algorithms to solve linear system $Hx = b$ with H and b random

(c) an iterative algorithm to solve the Dirichlet problem $\Delta u = 0$ on a random star-shaped region $\Omega \subset \mathbb{R}^2$ with random boundary data f on $\partial\Omega$, and

(d) a genetic algorithm to compute the equilibrium measure for orthogonal polynomials on the line.

In [DMOT] the authors also discussed similar universality results obtained by Bakhtin and Correll in a series of experiments with live participants recording

(e) decision making times for a specific task.

Whereas (a) and (b) concern finite dimensional problems, (c) shows that universality is also present in problems that are genuinely infinite dimensional. And whereas (a), (b) and (c) concern, in effect, deterministic dynamical systems acting on random initial data, problem (d) shows that universality is also present in genuinely stochastic algorithms.

The demonstration of universality in problems (a)–(d) raised the following issue: Given the common view of neuroscientists that the brain is just a big computer with hardware and software, one should be able to find evidence of universality in some neural computations. It is this issue that the the authors in [DMOT] to the work of Bakhtin and Correll.

All of the above results are numerical/experimental. In order to establish universality as a bona fide phenomenon in numerical analysis, and not just an artifact, suggested, however strongly, by certain computations as above, D. and Trogdon [DT 2017(1)] sought out, and proved, universality for an algorithm of interest. The algorithm they analyzed was the Toda eigenvalue algorithm to compute the largest eigenvalue of a random real symmetric, (alternatively, Hermitian), matrix. The goal of these lectures is to describe this work in detail. In subsequent work D. and Trogdon [DT2017(2)] also proved universality for other eigenvalue algorithms, including QR on sample covariance matrices. We will also describe this work, but in less detail.

We now describe the Toda and QR algorithms. Here we focus on the case where the matrices H are real symmetric, i.e. $H \in \Sigma_N$. The Hermitian case will be considered in later lectures.

The Toda equations have the form

$$\frac{dX}{dt} = [X, B(X)] = X\,B(X) - B(X)\,X, \quad X(0) = H \tag{10.1}$$

where $B(X) = X_- - X_-^T = -B(X)^T$, where X_- is the strictly lower triangular part of X. These equations are clearly Lischitz and hence have a unique local solution, $X(t)$, $0 \le t < t^*$ for some $0 < t^* \le \infty$.

Now for $0 \le t < t^*$

$$\frac{d}{dt} X^2 = \frac{dX}{dt} X + X \frac{dX}{dt}$$

$$= (XB - BX)X + X(XB - BX)$$
$$= [X^2, B]$$

ie

(11.1) $$\frac{d}{dt} X^2 = [X^2, B] \quad , \quad 0 \le t < t^*$$

Also as $B(X) = -B(X)^T$, we have

$$\frac{d X^T}{dt} = [X, B]^T = B^T X^T - X^T B^T$$
$$= X^T B - B X^T$$
$$= [X^T, B(X)] \quad , \quad 0 \le t < t^*$$

ie $$\frac{dX^T}{dt} = [X^T, B(X)]$$

But for $X(0) = H \in \Sigma_N$

$$X^T(0) = H^T = H$$

Thus as $X^T(t)$ and $X(t)$ solve the same ode and $X^T(0)$ $= X(0)$, we must have for $0 \le t < t^*$

(11.2) $$X(t) = X(t)^T$$

ie the Toda equations preserve Σ_N. From (11.1), as the trace of a commutator is always zero, we have

$$\frac{d}{dt} X^2 = 0$$

(12.1) Thus $\quad \mathrm{tr}\, X(t)^2 = \mathrm{tr}\, H^2$.

But as $X = X^T$, this implies

(12.2) $$\sum_{i,j=1}^{N} X_{ij}^2(t) = \sum_{i,j=1}^{N} H_{ij}^2$$

In particular

(12.3) $$|X_{ij}(t)| \le \sqrt{\mathrm{tr}\, H^2} < \infty \qquad \forall i,j, \quad 0 \le t < t^*$$

which gives an a priori bound on the entries of $X(t)$.

It follows by standard ode techniques that in fact

$t^* = \infty$, i.e. (10.1) has a unique global solution $X(t) = X(t)^T$.

By induction, one sees as in (11.1) that

$$\frac{d}{dt} X^k = [X^k, B(X)], \qquad k = 1, 2, \ldots$$

and so

(12.4) $$\mathrm{tr}\, X^k(t) = \mathrm{tr}\, X^k(0),$$

which implies that the eigenvalues of $X(t)$ are <u>constants of the motion</u> i.e.

(12.5) $$t \mapsto X(t) \text{ is isospectral.}$$

But more is true.

Let $Q(t)$, $t \geq 0$, $Q(0) = I$ be the solution

of the equation

(13.1) $$\frac{dQ}{dt} = Q\,B(X(t)), \qquad Q(0) = I$$

where $X(t)$ solves (10.1). As (13.1) is linear, the solution

of (13.1) is unique and global.

Then
$$\frac{d}{dt} QQ^T = \dot{Q}Q^T + Q\dot{Q}^T$$
$$= QBQ^T + Q(-B)Q^T = 0$$

and so
$$Q(t)\,Q(t)^T = \text{const} = I$$

i.e.

(13.2) $$Q(t) \text{ is orthogonal for all } t \geq 0$$

Now set

(13.3) $$\tilde{X}(t) \equiv Q(t)^T\, H\, Q(t)$$

Then
$$\frac{d}{dt}\tilde{X} = \dot{Q}^T H Q + Q^T H \dot{Q} = -BQ^T HQ + Q^T HQB$$
$$= [\tilde{X}, B(X)]$$

so that $\tilde{X}(t)$ solves the same equation as $X(t)$,

and $\tilde{X}(0) = H = X(0)$ and so $\tilde{X}(t) = X(t)$

Thus

(14.1) $$X(t) = Q(t)^T H \; Q(t)$$

So we see that no only is $t \mapsto X(t)$ iso-spectral

as in (12.5), but in fact

(14.2) $$X(t) \text{ is orthogonally equivalent to } X(0) = H, \quad t \geq 0.$$

In a later lecture we will use an argument of Moser

[Mos] to show that indeed

(14.3) $$X(t) \to X_\infty \quad \text{as } t \to \infty$$

where

(14.4) $$X_\infty = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$$

As noted earlier, the λ_i's are the eigenvalues of $X(0) = H$.

We will also show that, generically, the flow (10.1) is

(14.5) sorting i.e. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$.

As concluded in Deift–Nanda–Tomei [DNT 1983], the Toda flow (10.1) gives rise to an eigenvalue algorithm, viz., in order to compute the eigenvalues of a given matrix $H \in \Sigma_N$, solve (10.1) for $X(t)$ with $X(0) = H$. Then as $t \to \infty$

$$X(t) \to X_\infty = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$$

where the λ_i's are the eigenvalues of H. This is the Toda algorithm.

We will also show that (10.1) is a completely integrable Hamiltonian flow which can be integrated explicitly.

The history of the Toda system, or Toda lattice, is as follows. The lattice was introduced by M. Toda in 1967 and describes the motion of N particles x_i, $i = 1, 2, \ldots, N$, on the line under the Hamiltonian

$$H_{\mathrm{Toda}}(x, y) = \frac{1}{2} \sum_{i=1}^{N} y_i^2 + \frac{1}{2} \sum_{i=1}^{N-1} e^{x_i - x_{i+1}}.$$

In 1974 Flaschka (and independently Manakov) showed that Hamilton's equations

$$\dot{x} = \frac{\partial H_{Toda}}{\partial y}, \quad \dot{y} = -\frac{\partial H_{Toda}}{\partial x}$$

can be written in so-called Lax-pair form (10.1)

where $X = X(t)$ is tridiagonal and

$$(16.1) \quad \begin{cases} X_{ii} = -y_i/2, & i = 1, \dots, N \\ X_{i\,i+1} = X_{i+1,\,i} = \frac{1}{2} e^{\frac{1}{2}(x_i - x_{i+1})}, & 1 \le i \le N-1 \end{cases}$$

and $B(X)$ is the (tridiagonal) skew-symmetric matrix $B(X) = X_- - X_-^T$ as above. Not only, as noted by Flaschka, is the flow $t \mapsto X(t)$ isospectral, so that the eigenvalues $\lambda_i(t) = \lambda_i(t=0)$, $i = 1, \dots, N$, give N constants of motion for the Toda flow, but the λ_i's are independent and Poisson commute in the underlying symplectic structure,

$(\mathbb{R}^{2n}, \omega = \sum_{i=1}^{N} dx_i \wedge dy_i)$: Thus the Toda lattice is integrable in the sense of Liouville. In later work, Moser showed how to solve the Toda lattice explicitly and he also showed how to evaluate the long-time behavior of the system. The Toda system (10.1) is the natural extension of the original tridiagonal Toda lattice to full $N \times N$ matrices. Although it is not a priori clear, this extended system is also Hamiltonian and completely integrable in the sense of Liouville. We will derive, and explain, all the above properties of the Toda system in later chapters

We note here that it is a basic mantra in the modern theory of integrable systems, that if a dynamical system can be written in Lax-pair form

$$\frac{dS}{dt} = [S, U] \quad ,$$

for some $U = U(S)$, then the flow $t \mapsto S(t)$ is isospectral so that the system has (at least) N integrals of the motion, where $N = \dim S$. This follows in general, by the argument - due to Lax - following (13.1).

The QR algorithm works in the following way. Let X_0 be an invertible matrix in Σ_N. Then X_0 has a <u>QR factorization</u>

(18.1) $$X_0 = Q_0 R_0$$

where Q_0 is orthogonal, $Q_0 Q_0^T = Q_0^T Q_0 = I$, and R_0 is upper triangular with $(R_0)_{ii} > 0$, $i = 1, \ldots, N$. The factorization is unique.

Set

(19.1) $$X_1 = R_0 Q_0$$

Substituting $R_0 = Q_0^T X_0$ from (18.1) we see that

(19.2) $$X_1 = Q_0^T X_0 Q_0$$

Denote the mapping

$$X \to QR \to RQ = X'$$

by φ_{QR}: ie $X' = \varphi_{QR}(X)$. From (19.2) we see that

(19.3) $$\varphi_{QR} : \Sigma_N \to \Sigma_N \quad \text{is isospectral}$$

Now X_1 has its own QR factorization,

$$X_1 = Q_1 R_1$$

Set

$$X_2 = R_1 Q_1$$
$$= Q_1^T X_1 Q_1$$

etc.

In this way we obtain an isospectral sequence

$$X_0, X_1, \ldots, X_k, \ldots$$

of matrices. Generically X_k converges as $k \to \infty$ to a diagonal matrix $X_\infty = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$. Again, necessarily, the λ_i's are the eigenvalues of X_0. This construction is at the heart of the so-called QR algorithm that plays an outsize role in numerical analysis and occupies a prime position in software (see LINPACK) for eigenvalue computation. It turns out that the QR algorithm is also Hamiltonian and completely integrable, in the following way. There is a

(20.1) <u>Stroboscope Theorem</u> ([DNT 1983])

For $X \in \Sigma_N$ set

$$H_{QR}(X) = \mathrm{tr}(X \log X - X)$$

Then H_{QR} generates a Hamiltonian flow

$$t \mapsto X_{QR}(t), \qquad X(0) = H \text{ given}$$

with the property that

(20.2) $$X_{QR}(k) = X_k, \quad k = 0, 1, 2, \ldots$$

where X_k are the QR iterates generated by $X_{QR}(0) = X_0 = H$. Moreover the flow $t \mapsto X_{QR}(t)$ is completely integrable and commutes with the Toda flow generated by (10.1). We will describe, and explain, all the above properties of the QR algorithm in later chapters.

The goal of these lectures is to prove the result of D and Trogdon [DT 2017 (1)] on universality for the Toda eigenvalue algorithm to compute the largest eigenvalue of a matrix $H \in \Sigma_N$.

The result is the following: Let $X(t)$, $t \geq 0$, solve (10.1) with $X(0) = H \in \Sigma_N$. Set

(21.1) $$E(t) = \sum_{n=2}^{N} (X_{1n}(t))^2$$

so that $E(t) = 0$ implies $X_{11}(t)$ is an eigenvalue of H. Thus with $E(t)$ as in (21.1), the halting time (or the 1-deflation time) for the Toda algorithm is given by

$$T^{(1)}(H) = \inf\{t \geq 0 : E(t) \leq \varepsilon^2\}$$

Note that by the min-max principle, if $E(t) < \varepsilon^2$, $|X_{11}(t) - \lambda_j| < \varepsilon$ for some eigenvalue λ_j of $X(0) = H$.

As the Toda algorithm is generically sorting in the sense that as $X(t) \to X_\infty = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_N)$, $t \to \infty$, the eigenvalues appear ordered, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$. It follows that for generic $X(0) = H$, λ_j above is λ_1, the top eigenvalue of H.

Order the eigenvalues of a matrix $H \in \Sigma_N$ $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$.

For invariant and Wigner random matrix ensembles (see later lectures for appropriate definitions) there is a constant $c_V > 0$, which depends on the ensemble, such that the following limit exists

(23.1) $$F^{gap}(t) = \lim_{t\to\infty} \text{Prob}\left(\frac{1}{c_V^{2/3} 2^{-2/3} N^{2/3} (\lambda_1 - \lambda_2))} \le t \right), \quad t \ge 0$$

The basic universality result in [DT 2017 (1)] is the following.

(23.2) Theorem

Let $0 < \sigma < 1$ be fixed and let (ε, N) be in the scaling region

(23.3) $$\frac{\log \varepsilon^{-1}}{\log N} \geq \frac{5}{3} + \frac{\sigma}{2}$$

Then if H is distributed according to any invariant or Wigner ensemble we have for $t > 0$

(23.4) $$\lim_{N\to\infty} \text{Prob}\left(\frac{T^{(1)}}{c_V^{2/3} 2^{-2/3} N^{2/3} (\log \varepsilon^{-1} - \frac{2}{3}\log N)} \le t \right) = F^{gap}(t).$$

Thus the halting time for the Toda algorithm to compute the top eigenvalue of a random matrix ensemble is universal and behaves statistically like the inverse of the gap $\lambda_1 - \lambda_2$ of the two top eigenvalues of a random matrix. We will prove this result, and more, including the case where H is Hermitian, in the lectures that follow. As noted above, we will also describe related results for the QR algorithm and other eigenvalue algorithms, but in less detail.

Note finally that the scaling regime (23.3) in which random matrix behavior is guaranteed to appear, includes a common arena for numerical computing. Indeed for $\varepsilon = 10^{-16}$ and $N < 10^9$, we have

$$\frac{\log \varepsilon^{-1}}{\log N} > \frac{16}{9} > \frac{5}{3}.$$

The outline for these lectures is as follows:

- introduction to Hamiltonian mechanics and integrable systems
- the Toda lattice and its properties and its generalizations
- properties of random matrix ensembles
- proof of universality for the Toda algorithm to compute the top eigenvalue of a random matrix.