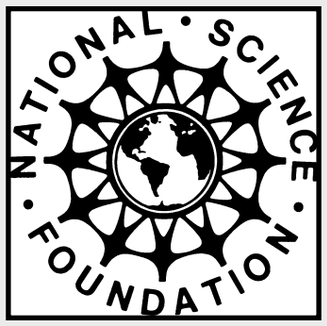
Numerical linear algebra and random matrix theory

Tom Trogdon
ttrogdon@math.uci.edu
UC Irvine

Acknowledgements

This is joint work with:

- Percy Deift (Courant)
- Aukosh Jagannath (Harvard)
- Yann LeCun (Courant)
- Govind Menon (Brown)
- Sheehan Olver (Imperial College)
- Raj Rao (U. of Mich.)
- Levent Sagun (ENS-Paris)



NSF Funding via:

CAREER: Numerical linear algebra, random matrix theory and applications

A short (incomplete) history of random matrix theory

1928: The first (?) appearance of random matrices

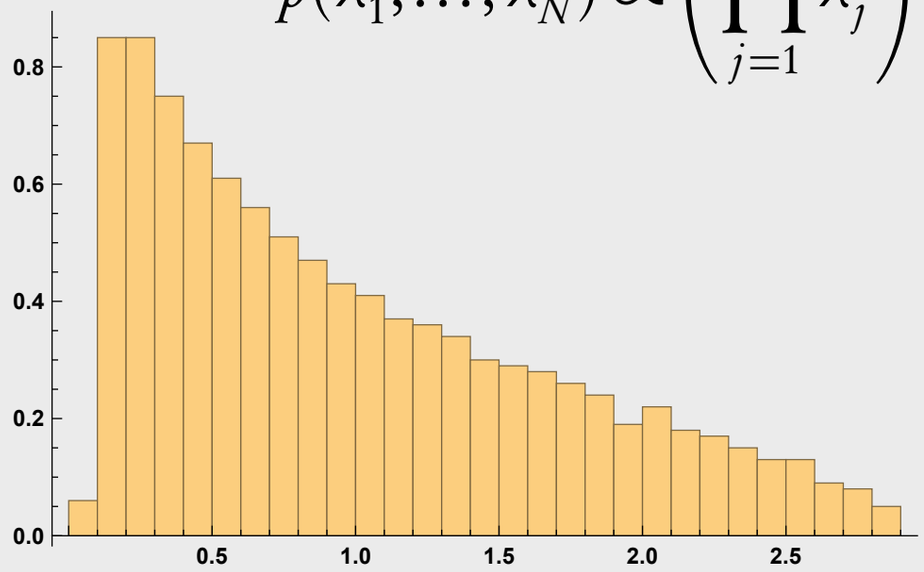
John Wishart

THE GENERALISED PRODUCT MOMENT DISTRIBUTION IN SAMPLES FROM A NORMAL MULTIVARIATE POPULATION.

By JOHN WISHART, M.A., B.Sc. Statistical Department, Rothamsted Experimental Station.

$$W = XX^* \qquad X \in \mathbb{C}^{N\times M}$$

$$p(\lambda_1,\ldots,\lambda_N) \propto \left(\prod_{j=1}^{N}\lambda_j\right)^{\alpha}\left(\prod_{j\neq k}|\lambda_j-\lambda_k|^2\right)\mathrm{e}^{-\frac{1}{2}\sum_j \lambda_j}$$



J Wishart. The generalised product moment distribution in samples from a normal multivariate population. Biometrika, 20A(1-2):32–52, 1928

Photo courtesy of Wikimedia Commons

1951: Random matrix theory and numerical linear algebra

John von Neumann

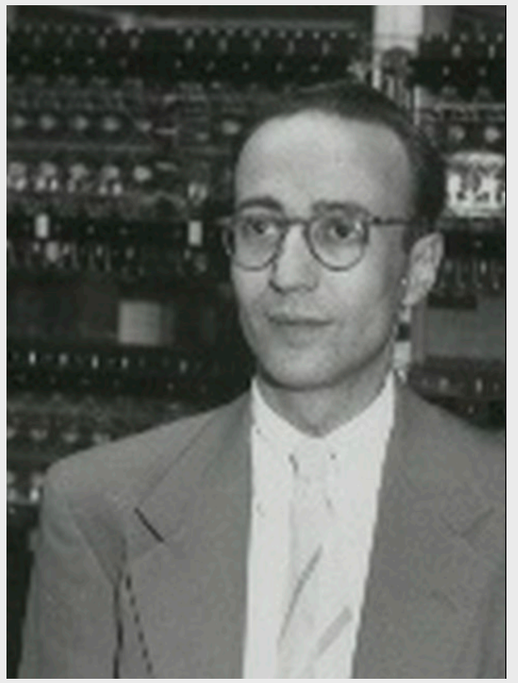
NUMERICAL INVERTING OF MATRICES OF HIGH ORDER. II

HERMAN H. GOLDSTINE AND JOHN VON NEUMANN

TABLE OF CONTENTS

PREFACE 188
CHAPTER VIII. Probabilistic estimates for bounds of matrices
8.1 A result of Bargmann, Montgomery and von Neumann 188
8.2 An estimate for the length of a vector 191
8.3 The fundamental lemma 192
8.4 Some discrete distributions 194
8.5 Continuation 196
8.6 Two applications of (8.16) 198
CHAPTER IX. The error estimates
9.1 Reconsideration of the estimates (6.42)–(6.44) and their consequences .. 199
9.2 The general A_I 200
9.3 Concluding evaluation 200

(8.9) The probability that the upper bound $|A|$ of the matrix A of (8.1) exceeds $2.72\sigma n^{1/2}$ is less than $.027 \times 2^{-n} n^{-1/2}$, that is, with probability greater than 99% the upper bound of A is less than $2.72\sigma n^{1/2}$ for $n = 2, 3, \cdots$.



Herman Goldstine

H H Goldstine and J von Neumann. Numerical inverting of matrices of high order. II. Proc. AMS, 2(2):188–202, feb 1951

Photos courtesy of Wikimedia Commons

1955: Random symmetric matrices

Eugene Wigner

ANNALS OF MATHEMATICS
Vol. 62, No. 3, November, 1955
Printed in U.S.A.

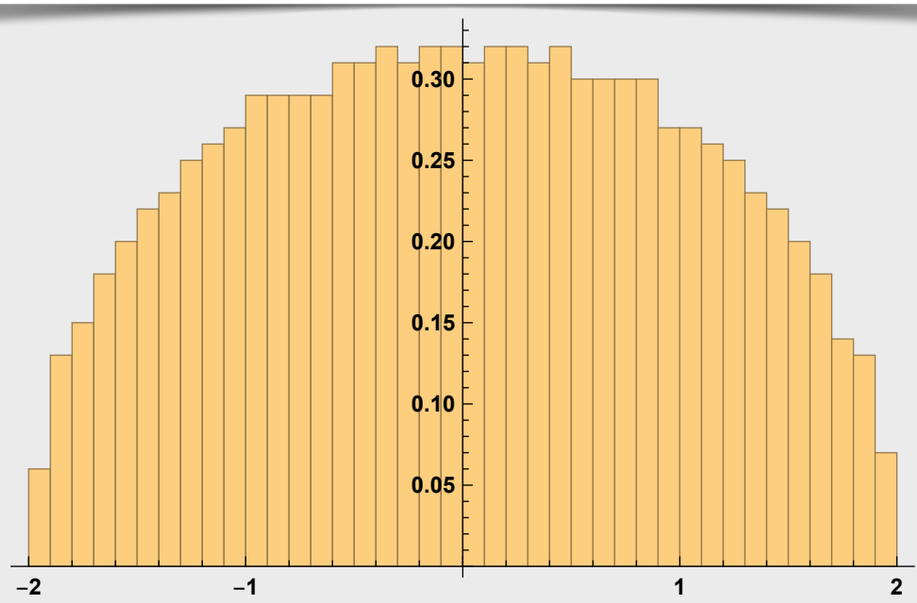
CHARACTERISTIC VECTORS OF BORDERED MATRICES WITH INFINITE DIMENSIONS

BY EUGENE P. WIGNER

(Received April 18, 1955)

Random sign symmetric matrix

The matrices to be considered are $2N + 1$ dimensional real symmetric matrices; N is a very large number. The diagonal elements of these matrices are zero, the non diagonal elements $v_{ik} = v_{ki} = \pm v$ have all the same absolute value but random signs. There are $\mathfrak{N} = 2^{N(2N+1)}$ such matrices. We shall calculate, after



E P Wigner. Characteristic Vectors of Bordered Matrices With Infinite Dimensions. Ann. Math., 62(3):548, nov 1955

Photo courtesy of Wikimedia Commons

Basics of random matrix theory

```
gin = @(N,M) randn(N,M)/sqrt(M);
```

The real Ginibre Ensemble, $\mathrm{Gin}_{\mathbb{R}}(N,M)$, is the matrix

$$Y = \left(\frac{Y_{ij}}{\sqrt{M}}\right)_{1\le i\le N, 1\le j\le M}, \quad Y_{ij} \text{ iid standard normal random variables}$$

```
Y = gin(N,M);
```

The complex Ginibre Ensemble, $\mathrm{Gin}_{\mathbb{C}}(N,M)$, is the matrix

$$X = \frac{1}{\sqrt{2}}(Y_1 + \mathrm{i}Y_2), \quad Y_1, Y_2 \text{ independent } \mathrm{Gin}_{\mathbb{R}}(N,M).$$

```
X = (gin(N,M) + i*gin(N,M))/sqrt(2);
```

If $N = M$ we use $\mathrm{Gin}_{\mathbb{F}}(N)$.

The Gaussian Unitary Ensemble, $\mathrm{GUE}(N)$, is the matrix

$$H = \frac{1}{\sqrt{2}}(X + X^*), \quad X \sim \mathrm{Gin}_{\mathbb{C}}(N).$$

```
X = (gin(N,N) + i*gin(N,N))/sqrt(2);
H = (X + X')/sqrt(2);
```

Fundamental distributions

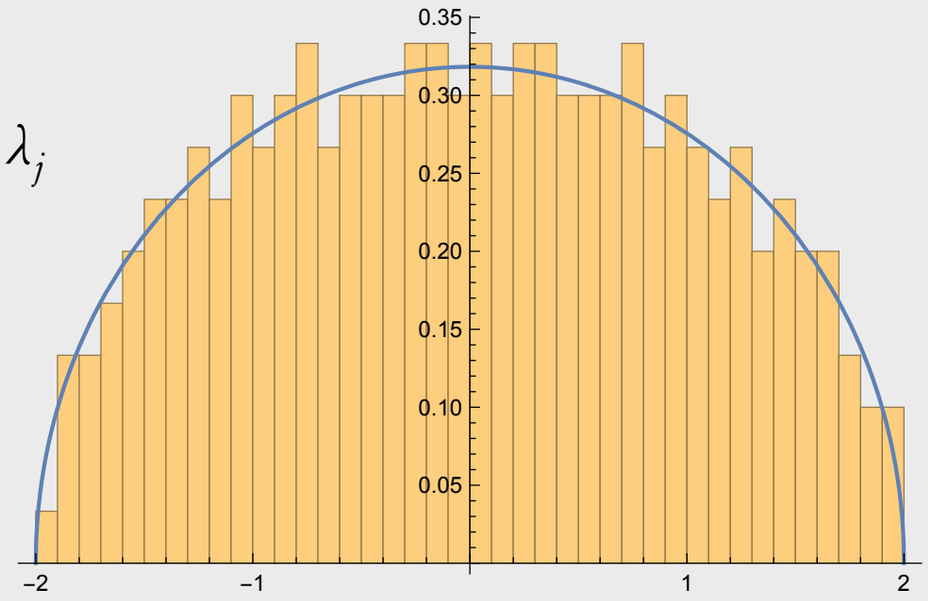
The global eigenvalue distribution is given by the famed semicircle law.

Let $H \sim \mathrm{GUE}(N)$ and let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ be its eigenvalues.

Let $\mathcal{N}(a,b)$ be the number of eigenvalues that lie in the interval (a,b). Then

$$\lim_{N\to\infty} \frac{1}{N}\mathcal{N}(a,b) = \int_a^b p_{\mathrm{sc}}(x)\mathrm{d}x \;\; a.s., \qquad p_{\mathrm{sc}}(x) = \frac{1}{2\pi}\sqrt{|4-x^2|_+}.$$

Histogram of the eigenvalues λ_j



Fundamental distributions

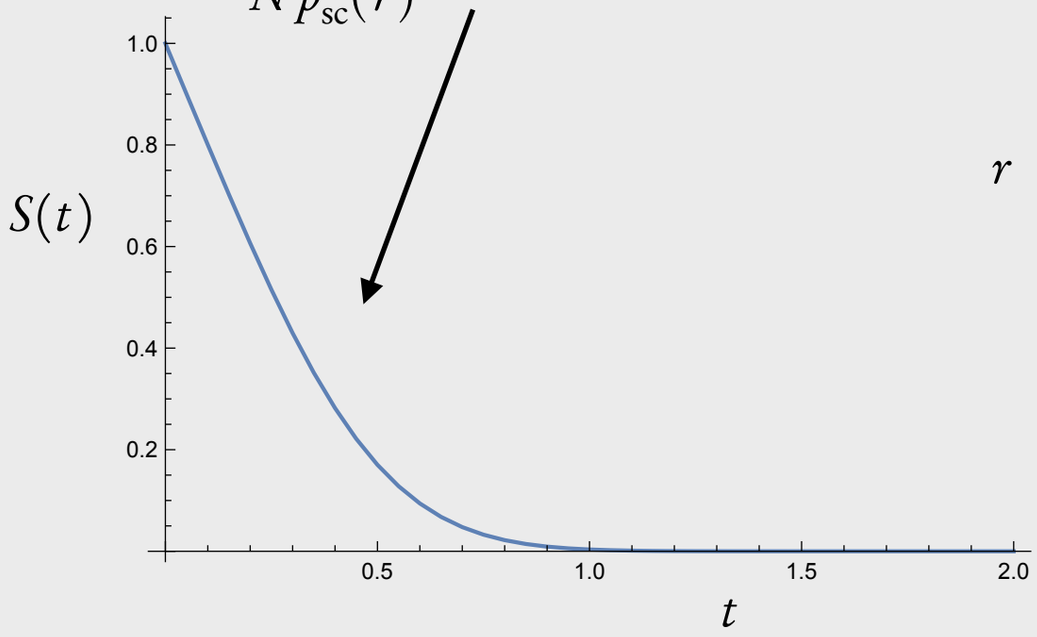
The other fundamental distributions from random matrix theory have Fredholm determinant representations.

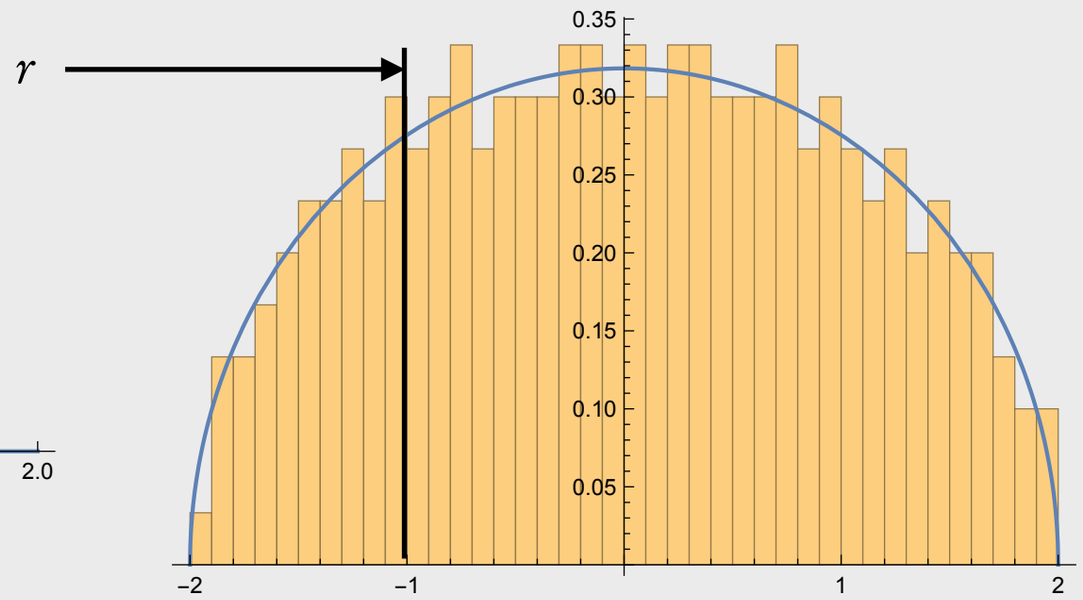
For $r \in (-2,2)$ and $t > 0$

$$\lim_{N\to\infty} \mathbb{P}\Big(N p_{\rm sc}(r)\Big(\lambda_j - r\Big) \notin (-t,t),\ 1 \le j \le N\Big) = \det(I - K_{\rm sine}|_{L^2((-t,t))}) =: S(t).$$

$$K_{\rm sine}(x,y) = \frac{\sin \pi(x-y)}{\pi(x-y)}$$

Probability of a gap of size at least $\dfrac{2t}{N p_{\rm sc}(r)}$, centered at r





P Deift. Orthogonal Polynomials and Random Matrices: a Riemann-Hilbert Approach. Amer. Math. Soc., Providence, RI, 2000

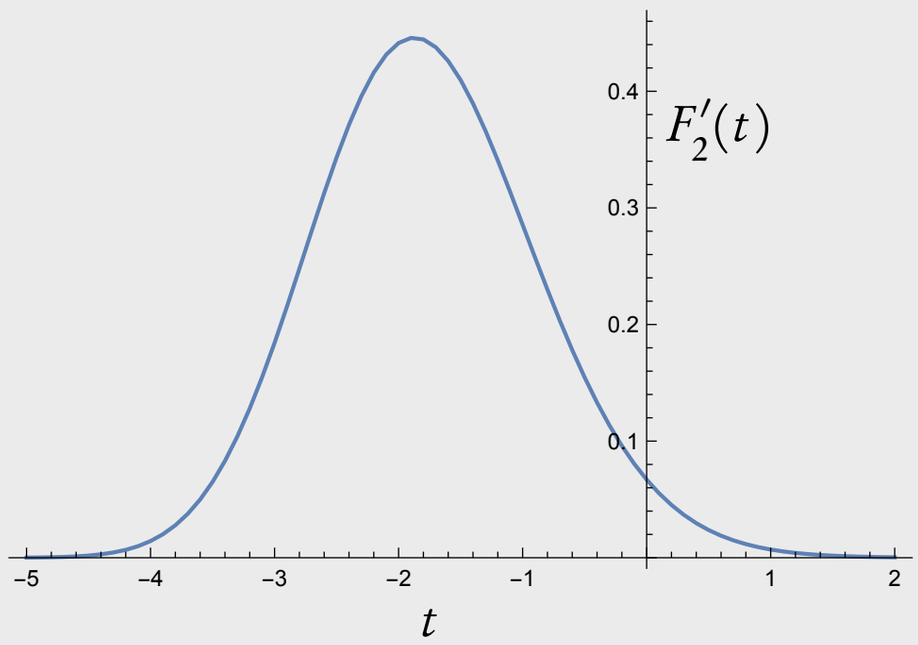
Fundamental distributions

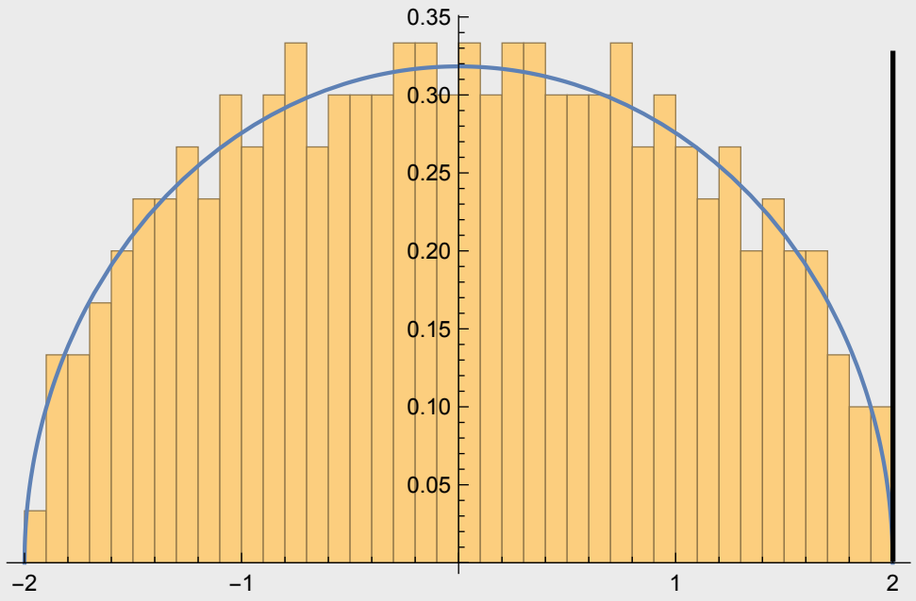
For the largest eigenvalue λ_N we have

$$\lim_{N\to\infty} \mathbb{P}\left(N^{2/3}(\lambda_N - 2) \le t\right) = \det(I - K_{\text{Airy}}|_{L^2((t,\infty))}) =: F_2(t).$$

$$K_{\text{Airy}}(x,y) = \frac{\text{Ai}(x)\,\text{Ai}'(y) - \text{Ai}(y)\,\text{Ai}'(x)}{x-y}.$$

$$\lambda_N \approx 2 + \xi N^{-2/3}, \quad \xi \sim F_2$$





C A Tracy and H Widom. Level-spacing distributions and the Airy kernel. Comm. Math. Phys., 159:151–174, 1994

Wigner's surmise and gaps

One can also ask about the distance between two eigenvalues, say, for simplicity,

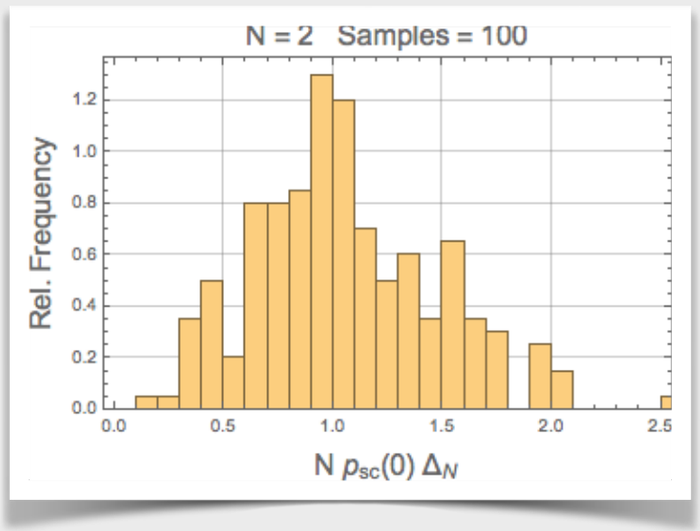
$$\Delta_N = \lambda_{\lfloor \frac{N}{2} \rfloor + 1} - \lambda_{\lfloor \frac{N}{2} \rfloor}.$$

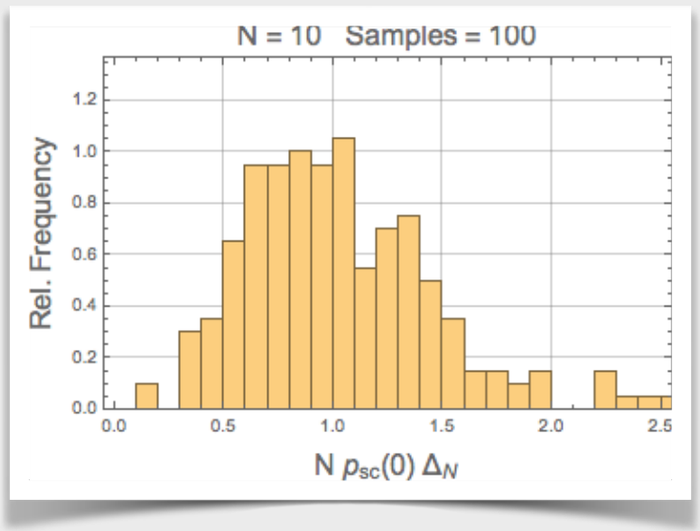
From the gap calculation, it follows that $\Delta_N = O(N^{-1})$.

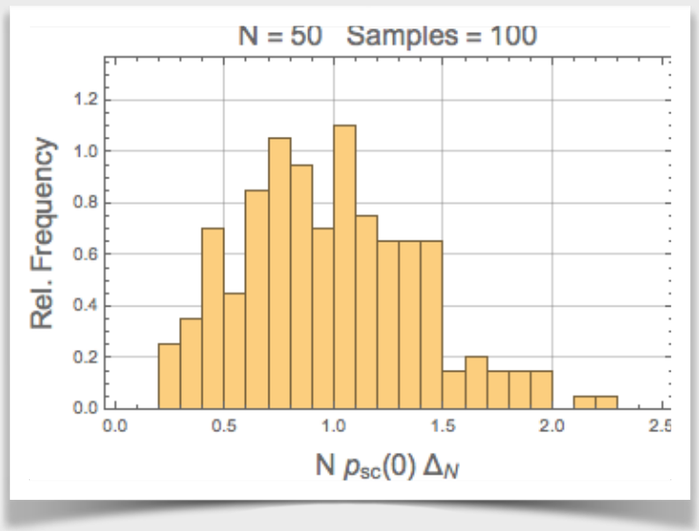
More specifically, we expect

$$N p_{\mathrm{sc}}(0) \Delta_N$$

to converge in distribution to something with mean one.







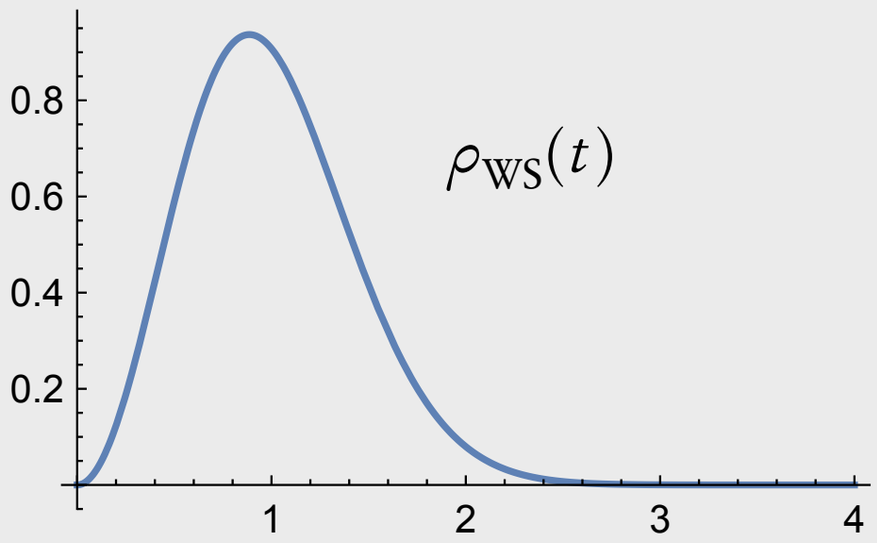
Wigner's surmise and gaps

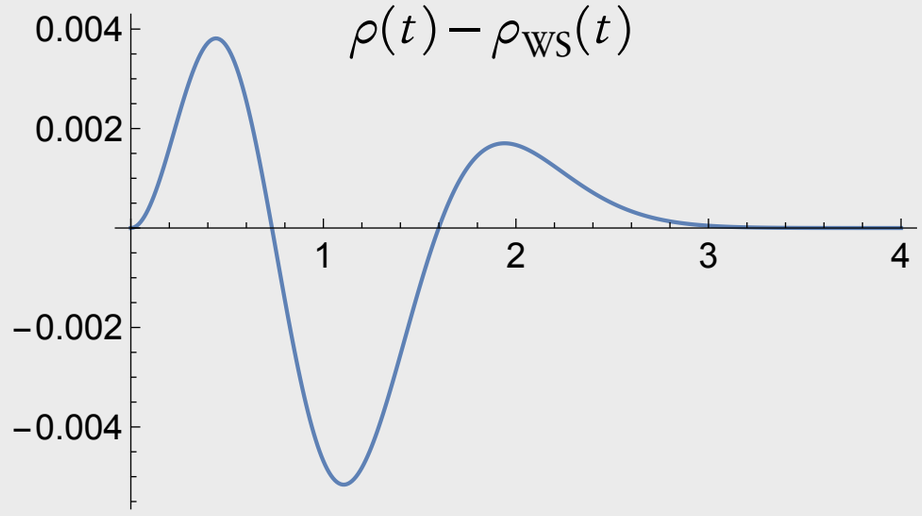
The "Wigner surmise" is an estimate of the limiting density $\rho(t)$ of $N p_{\mathrm{sc}}(0)\Delta_N$:

$$\rho_{\mathrm{WS}}(t) \approx \rho(t), \quad \rho_{\mathrm{WS}}(t) := \frac{32}{\pi^2} t^2 \mathrm{e}^{-\frac{4}{\pi}t^2}.$$

This is exact for the 2×2 case, normalized to mean one. The true limit is,

$$\rho(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left[S\left(\frac{t}{2}\right)\right], \quad S(t) = \det(I - K_{\mathrm{sine}}|_{L^2((-t,t))}).$$





T Tao. The asymptotic distribution of a single eigenvalue gap of a Wigner matrix. Probab. Theory Relat. Fields, 157(1-2):81–106, 2013

P Deift. Orthogonal Polynomials and Random Matrices: a Riemann-Hilbert Approach. Amer. Math. Soc., Providence, RI, 2000

1962: Dyson Brownian Motion

Freeman Dyson

A Brownian-Motion Model for the Eigenvalues of a Random Matrix

FREEMAN J. DYSON
Institute for Advanced Study, Princeton, New Jersey
(Received June 22, 1962)

A new type of Coulomb gas is defined, consisting of n point charges executing Brownian motions under the influence of their mutual electrostatic repulsions. It is proved that this gas gives an exact mathematical description of the behavior of the eigenvalues of an $(n \times n)$ Hermitian matrix, when the elements of the matrix execute independent Brownian motions without mutual interaction. By a suitable choice of initial conditions, the Brownian motion leads to an ensemble of random matrices which is a good statistical model for the Hamiltonian of a complex system possessing approximate conservation laws. The development with time of the Coulomb gas represents the statistical behavior of the eigenvalues of a complex system as the strength of conservation-destroying interactions is gradually increased. A "virial theorem" is proved for the Brownian-motion gas, and various properties of the stationary Coulomb gas are deduced as corollaries.

The particle at x_i feels an external electric force

$$E(x_i) = -\frac{\partial W}{\partial x_i} = \sum_{i \neq j} \left[\frac{1}{x_i - x_j}\right] - \frac{x_i}{a^2}, \qquad (16)$$

in addition to the local frictional force and the constantly fluctuating forces which give rise to the Brownian motion.

F J Dyson. A Brownian-Motion Model for the Eigenvalues of a Random Matrix. J. Math. Phys., 3(6):1191–1198, nov 1962

Photo courtesy of sigmapisigma.org

Dyson Brownian Motion and random eigenvalues

The equations of motion for Dyson Brownian motion are

$$\mathrm{d}\lambda_j = \frac{1}{N}\sum_{k:k\neq j}\frac{\mathrm{d}t}{\lambda_j-\lambda_k} - \frac{1}{2}V'(\lambda_j)\mathrm{d}t + \sqrt{\frac{1}{N}}\mathrm{d}W_j, \quad j=1,\ldots,N.$$

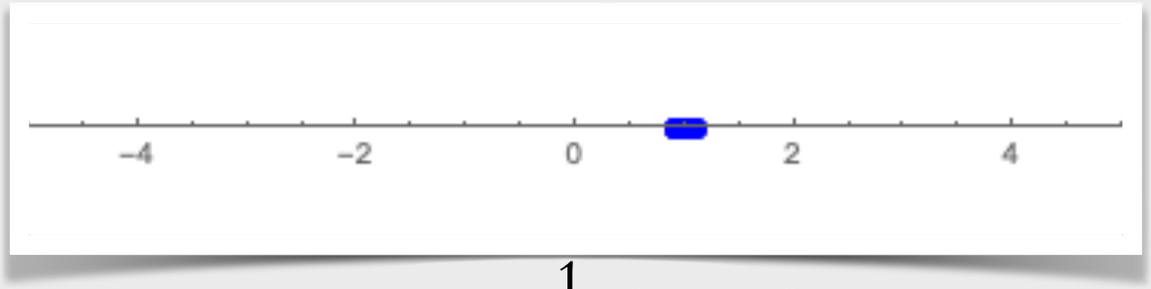
As T becomes large, the joint distribution of $(\lambda_1(T),\ldots,\lambda_N(T))$ tends to

$$\frac{1}{Z_N}\prod_{j\neq k}|\lambda_j-\lambda_k|^2 \mathrm{e}^{-N\sum_j V(\lambda_j)}.$$

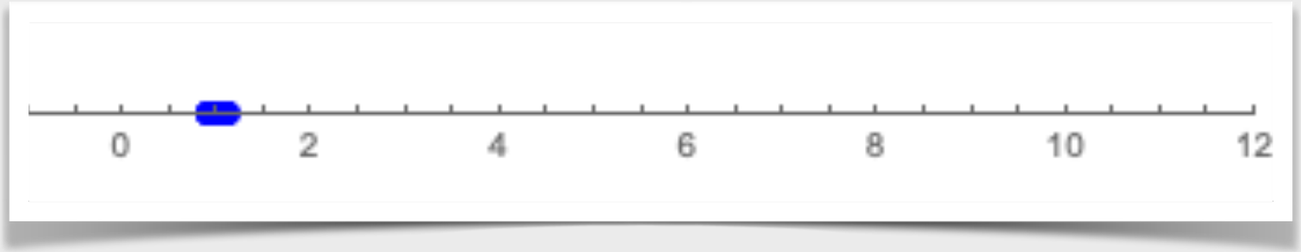
This turns out to be exactly the joint distribution of a random matrix ensemble

$$\frac{1}{Z_N}\mathrm{e}^{-N\mathrm{tr}V(M)}\mathrm{d}M.$$

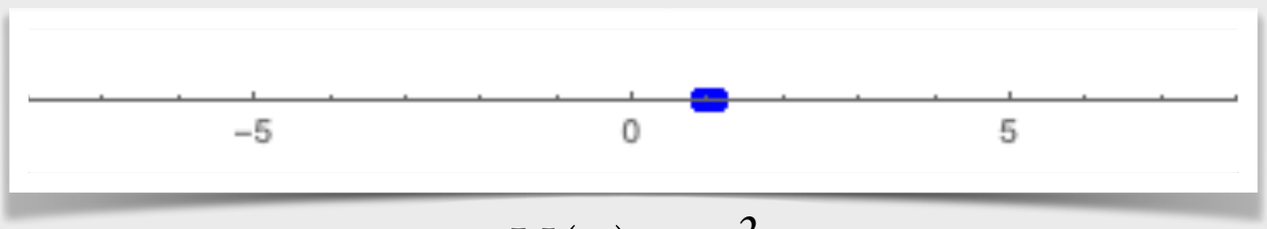
Dyson Brownian motion with different geometries and potentials



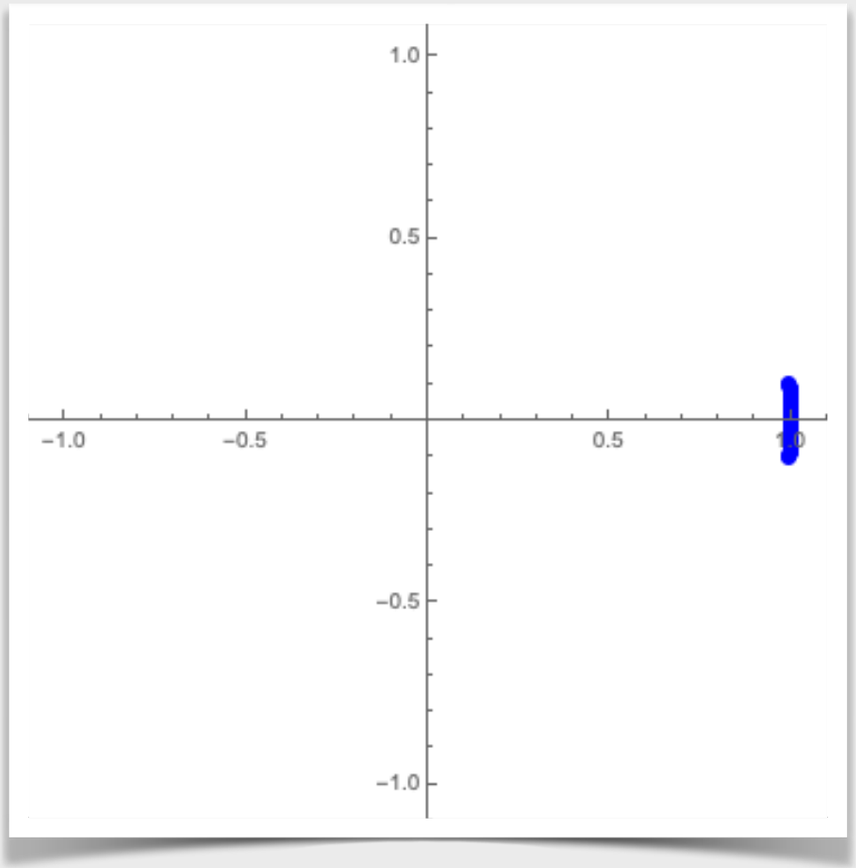
$$V(x) = \frac{1}{4}x^4 - 5x^2$$



$$V(x) = x - \log x^4, \quad x > 0$$



$$V(x) = x^2$$



$$V(x) = 1, \quad |x| = 1$$

X. H. Li and G. Menon. Numerical solution of Dyson Brownian motion and a sampling scheme for invariant matrix ensembles. J. Stat. Phys., 153(5):801–812, oct 2013

Universality

In short, "universality" is the statement that the above limits persist for non-normal matrix entries. It is arguably the most important aspect of random matrix theory

For example, if one replaces $\mathrm{Gin}_{\mathbb{R}}(N)$ with a matrix of iid Bernoulli $\pm 1/\sqrt{N}$ random variables, the above limits remain. Various tail/moment assumptions are required.

Entries can even be independent but not identically distributed. For the so-called invariant ensembles the entries are not even independent.

T Tao and V Vu. Random Matrices: Universality of Local Eigenvalue Statistics up to the Edge. Commun. Math. Phys., 298(2):549–572, apr 2010

P Deift. Orthogonal Polynomials and Random Matrices: a Riemann-Hilbert Approach. Amer. Math. Soc., Providence, RI, 2000

L Erdős, H-T Yau, and J Yin. Rigidity of eigenvalues of generalized Wigner matrices. Adv. Math. (N. Y)., 229(3):1435–1515, feb 2012

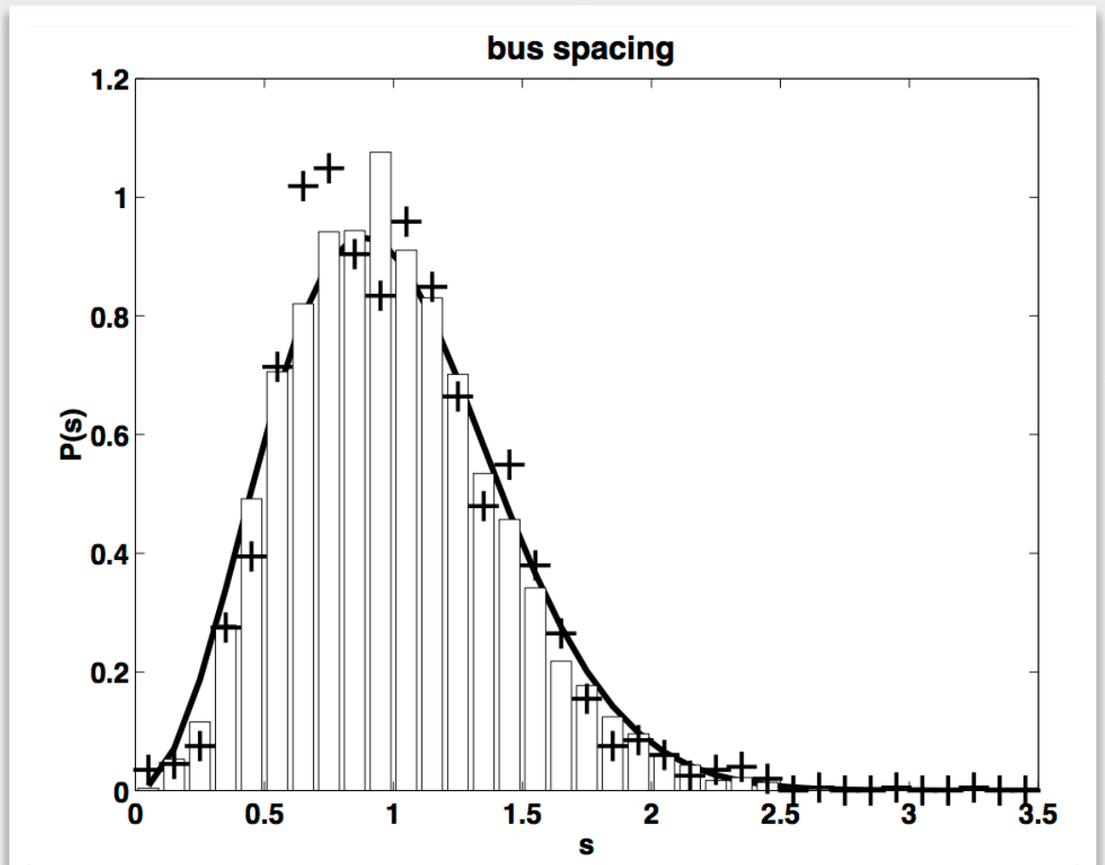
P Bourgade, L Erdős, and H-T Yau. Edge Universality of Beta Ensembles. Commun. Math. Phys., 332(1):261–353, nov 2014

Transportation statistics

The bus system in Cuernavaca, Mexico in the late 1990's has become a canonical example of a system that is well-modeled by RMT.

The busses are independent agents that space themselves out to maximize profits.

The statistics of the gap between busses were compared with the Wigner surmise.



A comparison to $\rho_{\mathrm{WS}}(t)$

M Krbálek and P Seba. The statistical properties of the city transport in Cuernavaca (Mexico) and random matrix ensembles. *J. Phys. A. Math. Gen.*, 33(26):L229–L234, jul 2000

J Baik, A Borodin, P Deift, and T Suidan. A model for the bus system in Cuernavaca (Mexico). *J. Phys. A. Math. Gen.*, 39(28):8965–8975, jul 2006

Transportation statistics

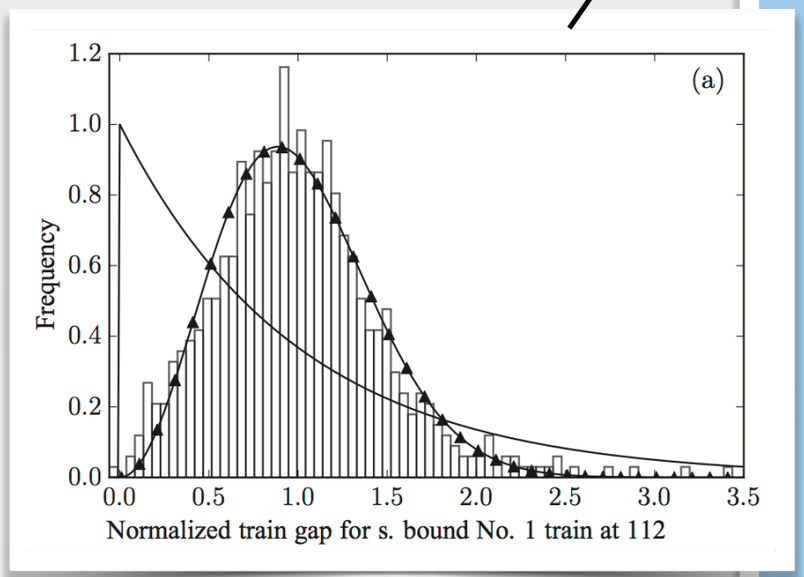
Recently, we (with A. Jagannath) showed that a similar phenomenon exists within the New York City subway system (MTA).

NYC gives access to real-time data feeds for train arrivals. We analyzed data for the #1 and #6 trains during rush hour for 20+ days.

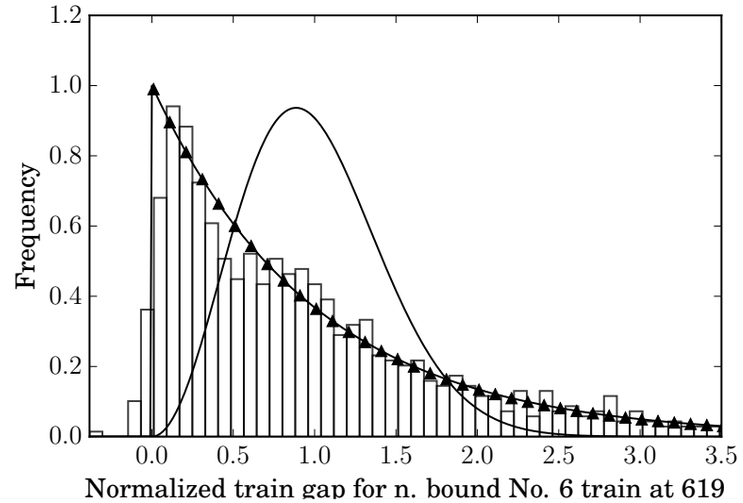
There is no global control system for the MTA. Trains are individually operatated but the drivers respond to signals. The MTA is currently undergoing a $37 billion renovation that includes implementing global control.

A Jagannath and T T. Random matrices and the New York City subway system. *Phys. Rev. E*, 96(3):030101, sep 2017

RMT in NYC

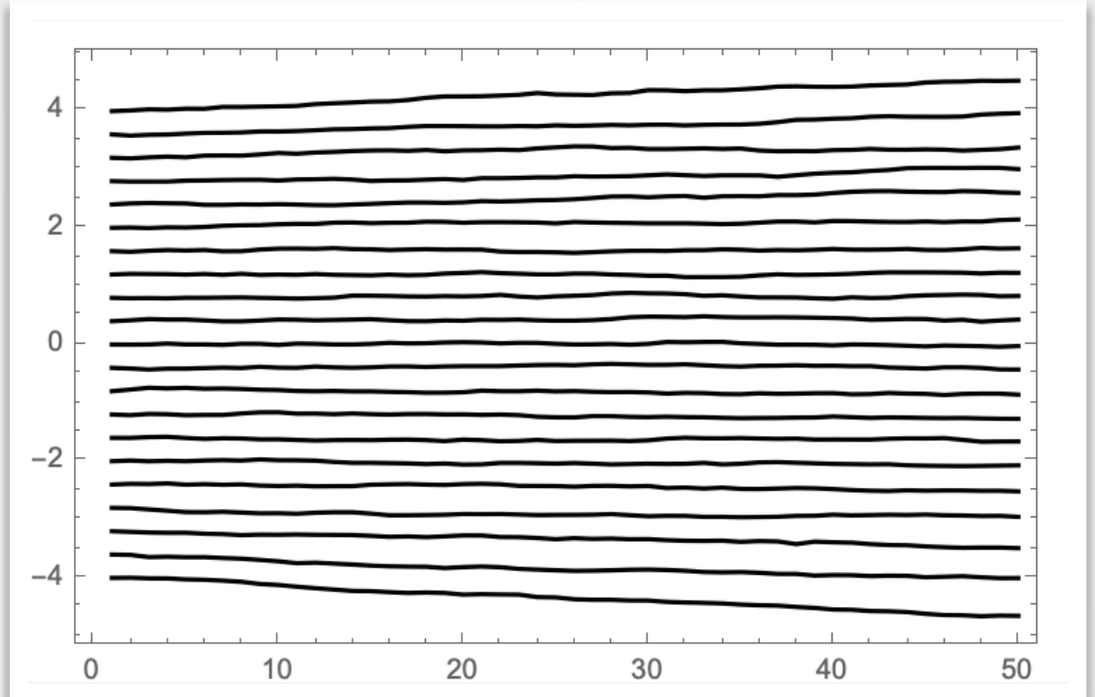


A comparison to $\rho_{\mathrm{WS}}(t)$

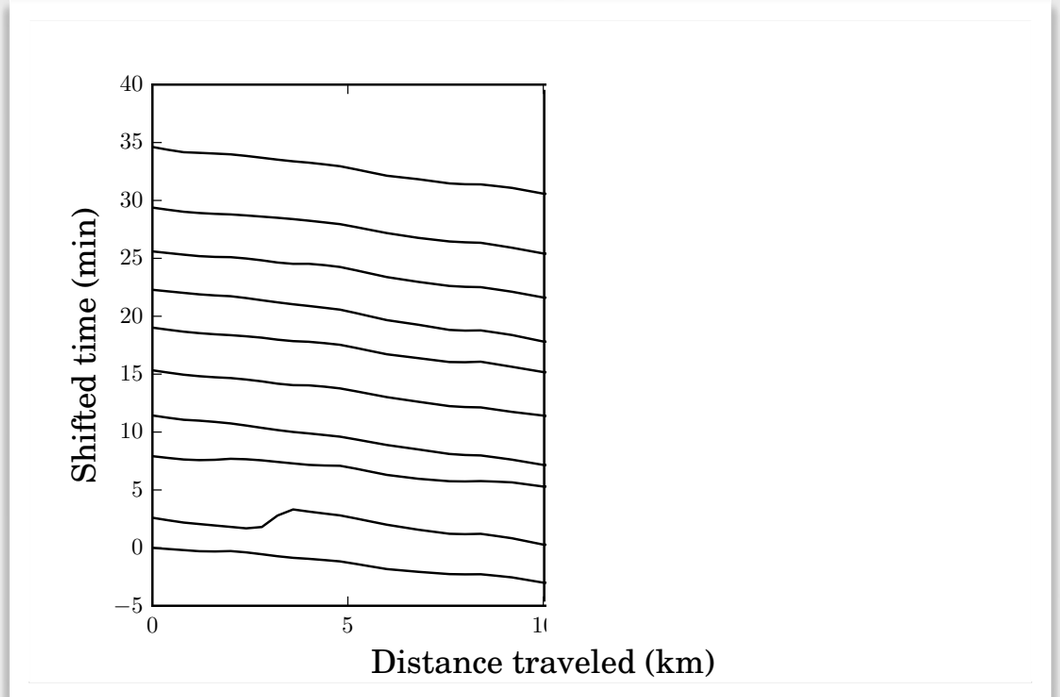


A comparison to e^{-t}

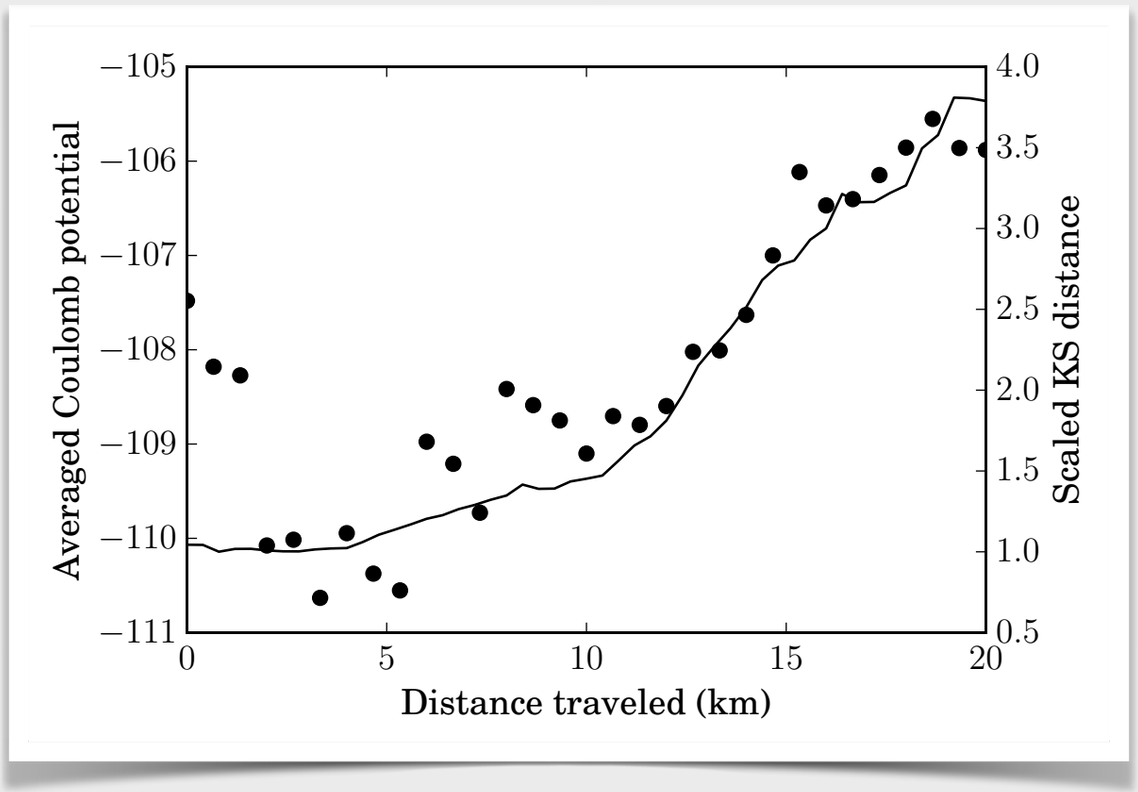
Noisy trajectories



Dyson Brownian motion trajectories



Rescaled train trajectories



Universality in numerical linear algebra

Numerical analysis and random matrix theory

- Using techniques from numerical analysis to analyze random matrices

  Trotter (1984), Silverstein (1985), Edelman (1989), Dumitriu and Edelman (2002)

- Computing distributions from random matrix theory

  Bornemann (2008), Witte, Bornemann and Forrester (2012), Olver and T (2014)

- Generating samples from random matrix distributions

  Mezzadri (2006), Edelman and Rao (2005), Menon and Li (2014), Olver, Rao and T (2015)

- Using random matrices to analyze algorithms, statistically

  Spielman and Teng (2009), Borgwardt (2012), Smale (1983, 1985), Deift and T (2016, 2017, 2019), Menon and T (2016), Feldheim, Paquette, and Zeitouni (2014)

- Randomized linear algebra

  Strohmer and Vershynin (2009), Halko, Martinsson and Tropp (2011), T (2018)

Statistical analysis of algorithms
Example 1

- Smoothed analysis: Spielman and Teng (2009)
- Average-case analysis: Borgwardt (2012), Smale (1985)

$$\mathbb{E}\left[\begin{array}{c}\text{Number of iterations to}\\ \text{solve a problem of dimension } N\\ \text{to accuracy } \epsilon\end{array}\right] \leq C_\epsilon N^\alpha$$

Statistical analysis of algorithms
Example 2

- Universality: Pfrang, Deift and Menon (2014)

$$\mathbb{P}\left(\begin{array}{c}\text{Number of iterations to}\\ \text{solve a problem of dimension } N\\ \text{to accuracy } \epsilon\end{array} \leq C_\epsilon N^\alpha + t D_\epsilon N^\gamma\right) \overset{N\to\infty}{\longrightarrow} F(t)$$

Statistical analysis of algorithms
Example 3

- Error concentration: Deift and TT (2019)

$$\mathbb{P}\left(\left|\begin{array}{c}\text{Error at iteration } k \\ \text{for a problem} \\ \text{of dimension } N\end{array} - E_k\right| \geq \epsilon\right) \leq C\mathrm{e}^{-cN}$$

Sample covariance matrices (SCMs)

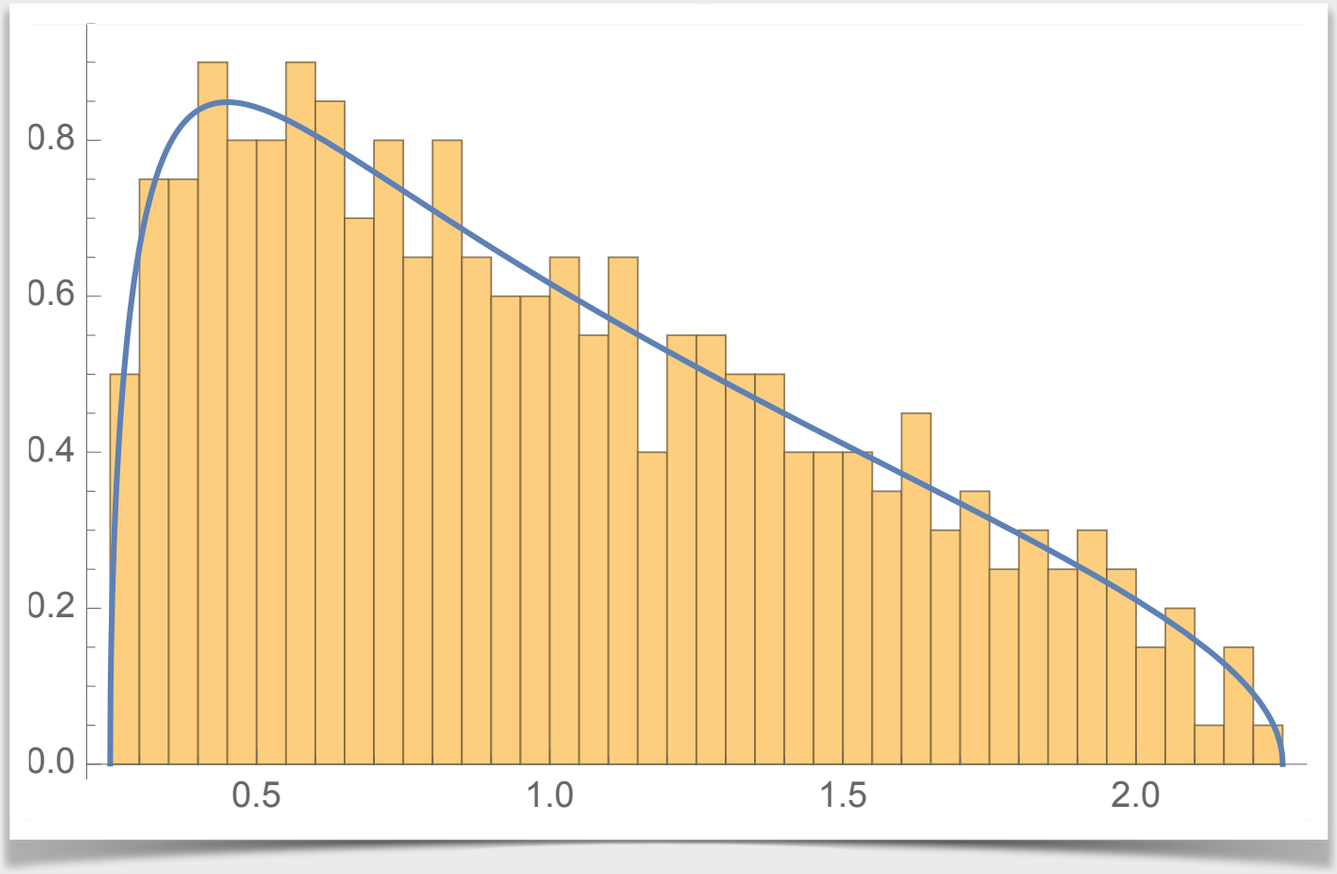
The classical examples o

The Laguerre Orthogon

```
M = floor(N/d); X = gin(N,M);
W = X*X';

% or

M = floor(N/d);
X = (gin(N,M) + i*gin(N,M))/sqrt(2);
W =X*X';
```



The semicircle distribution is replaced with the Marchenko–Pastur distribution:

$$p_d(x) = \frac{1}{2\pi d}\frac{\sqrt{[(\lambda_+ - x)(x-\lambda_-)]_+}}{|x|}$$

$$\lambda_\pm = (1 \pm \sqrt{d})^2$$

$$M \sim N/d$$

The power method on sample covariance matrices (SCMs)

The ensembles: Choose an SCM that is real ($\beta = 1$) or complex ($\beta = 2$), with

$$W = XX^*, \quad X \text{ iid and } N \times M, \quad M \sim N/d, \quad \mathbb{E}X_{ij} = 0, \quad \mathbb{E}|X_{ij}|^2 = 1/M.$$

(Ignoring technical assumptions...)

The method: The power method is given by ($\mathbf{x}_0$ specified)

$$\mu_k = \frac{\mathbf{x}_0^* W^{2k-1} \mathbf{x}_0}{\mathbf{x}_0^* W^{2k-2} \mathbf{x}_0} \to \lambda_N = \lambda_{\max}, \ \ k \to \infty.$$

The halting time: The halting time (or runtime, or iteration count) is given by

$$T(W, \mathbf{x}_0, \epsilon) = \min\left\{n : |\mu_n - \mu_{n-1}| < \epsilon^2\right\}.$$

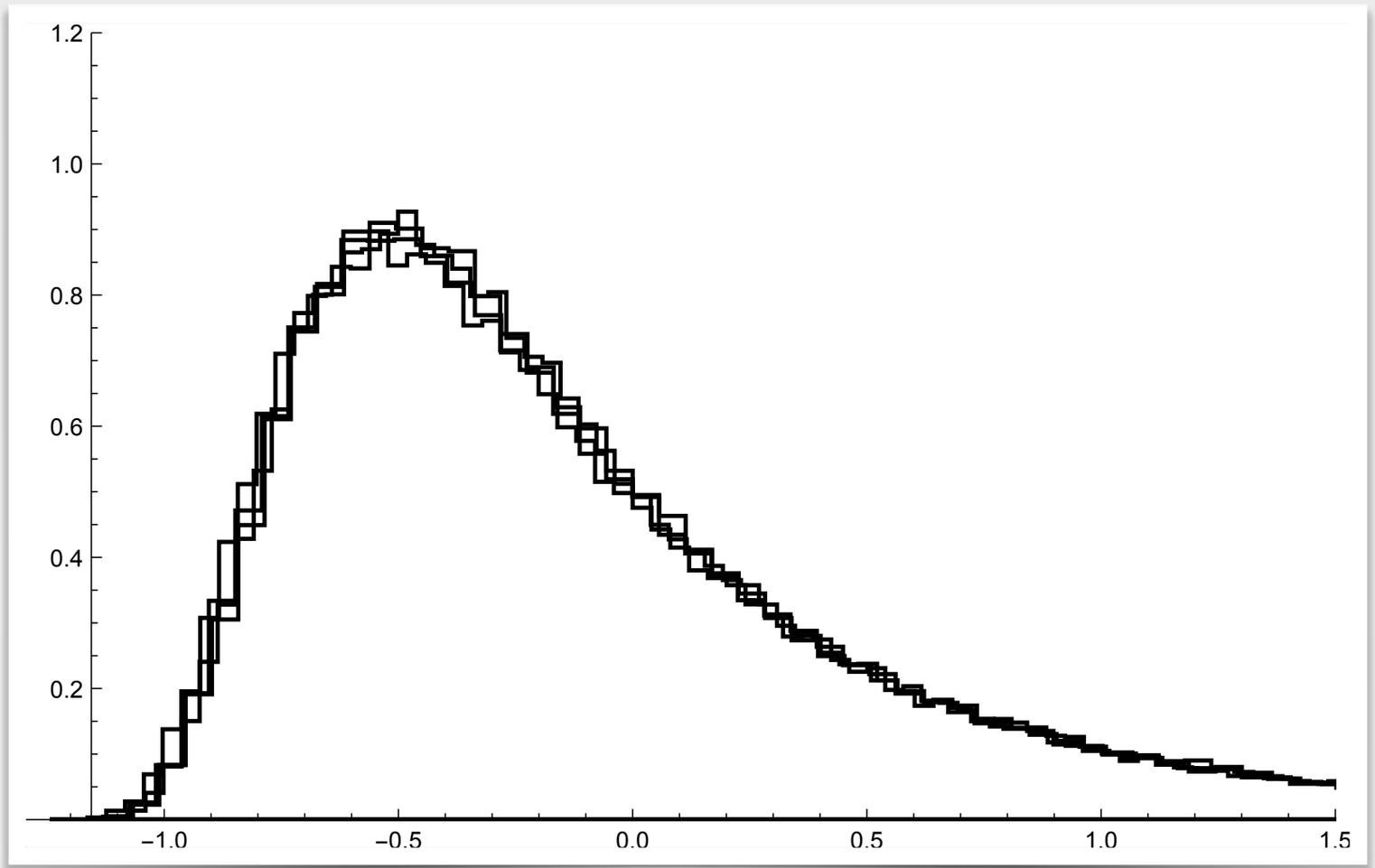
The question: What is the distribution of T? Is it universal as $N \to \infty$?

Observing universality

Using 4 complex SCMs, the fluctuations

$$\tau(H,\epsilon) = \frac{T(H,\epsilon) - \langle T \rangle}{\sigma_T},$$

appear universal.



A formula to estimate the error

Let $W = U\Lambda U^*$ and $U^*\mathbf{x}_0 = [\beta_1, \beta_2, \ldots, \beta_N]^T$,

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N), \quad 0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N.$$

To estmate the halting time for the power method, we must find the large N and large k asymptotics for

$$\mu_k = \frac{\sum_{j=1}^{N} |\beta_j|^2 \lambda_j^{2k-1}}{\sum_{j=1}^{N} |\beta_j|^2 \lambda_j^{2k-2}} = \lambda_N \left(\frac{1 + \sum_{j=1}^{N-1} \left| \frac{\beta_j}{\beta_N} \right|^2 \left(\frac{\lambda_j}{\lambda_N} \right)^{2k-1}}{1 + \sum_{j=1}^{N-1} \left| \frac{\beta_j}{\beta_N} \right|^2 \left(\frac{\lambda_j}{\lambda_N} \right)^{2k-2}} \right).$$

A historical interlude

For eigenvalues:

- The seminal work of Geman (1980) showed that the largest eigenvalue of an SCM converges a.s.
- Silverstein (1985) established that the smallest eigenvalue converges a.s. to λ_- for iid standard normal random variables.
- Johnstone (2001); Johansson (2000); Forrester (1993) gave the first results on the fluctuations of the largest and smallest eigenvalues for (real or complex) standard normal distributions.
- Universality was obtained by Soshnikov (2001) and Ben Arous and Péché (2005) (see also Tao and Vu (2012)).
- We reference Pillai and Yin (2014) and Bloemendal, Knowles, Yau and Yin (2016) for the most comprehensive results.

For eigenvectors:

- The limits of the eigenvectors have also been considered in various ways, see Silverstein (1986); Bai, Miao and Pan (2007).
- We reference Bloemendal, Knowles, Yau and Yin (2016) for the generality nee to prove our theorems.

We require exponential tails which is stronger than the assumptions in Yin (1986); Geman (1980).

Universality for key statistics

Theorem. For SCMs

$$N^{1/2}(|\beta_N|, |\beta_{N-1}|, |\beta_{N-2}|)$$

converge jointly in distribution to $(|X_1|, |X_2|, |X_3|)$ where $\{X_1, X_2, X_3\}$ are iid real ($\beta = 1$) or complex ($\beta = 2$) standard normal random variables. Additionally,

$$N^{2/3}\lambda_+^{-2/3} d^{1/2}(\lambda_+ - \lambda_N, \lambda_+ - \lambda_{N-1}, \lambda_+ - \lambda_{N-2})$$

converge jointly in distribution to random variables $(\Lambda_{1,\beta}, \Lambda_{2,\beta}, \Lambda_{3,\beta})$ which are the smallest three eigenvalues of the so-called stochastic Airy operator.

This follows from Ramírez, Rider and Virág (2011); Pillai and Yin (2014); Bloemendal, Knowles, Yau and Yin (2016).

J A Ramírez, B Rider, and B Virág. Beta ensembles, stochastic Airy spectrum, and a diffusion. J. Am. Math. Soc., 24(4):919–944, jan 2011

N S Pillai and J Yin. Universality of covariance matrices. Ann. Appl. Probab., 24(3):935–1001, jun 2014

A Bloemendal, A Knowles, H-T Yau, and J Yin. On the principal components of sample covariance matrices. Probab. Theory Relat. Fields, 164(1-2):459–552, feb 2016

Estimates for the rest

Theorem (Pillai and Yin (2014)). For any $s > 0$

$$\lim_{N\to\infty} \mathbb{P}\Big(|\lambda_n - \gamma_n| \le N^{-2/3+s}(\max\{n, N-n+1\})^{-1/3} \text{ for all } n\Big) = 1.$$

Theorem (Bloemendal, Knowles, Yau and Yin (2016)). For any $s > 0$

$$\lim_{N\to\infty} \mathbb{P}\Big(|\beta_n| \le N^{-1/2+s} \text{ for all } n\Big) = 1.$$

The first theorem is known as rigidity and it was first shown for Wigner ensembles by Erdős, Yau and Yin (2012) (see also Bourgade, Erdős and Yau (2014)).

The second theorem is known as delocalization.

Universality for the power method

The distribution function $F_\beta^{\mathrm{gap}}(t)$, supported on $t \geq 0$ for $\beta = 1, 2$ is defined by

$$F_\beta^{\mathrm{gap}}(t) := \lim_{N\to\infty} \mathbb{P}\left(\frac{1}{2^{-7/6} N^{2/3} \lambda_+^{-2/3} d^{-1/2} (\lambda_N - \lambda_{N-1})} \leq t \right).$$

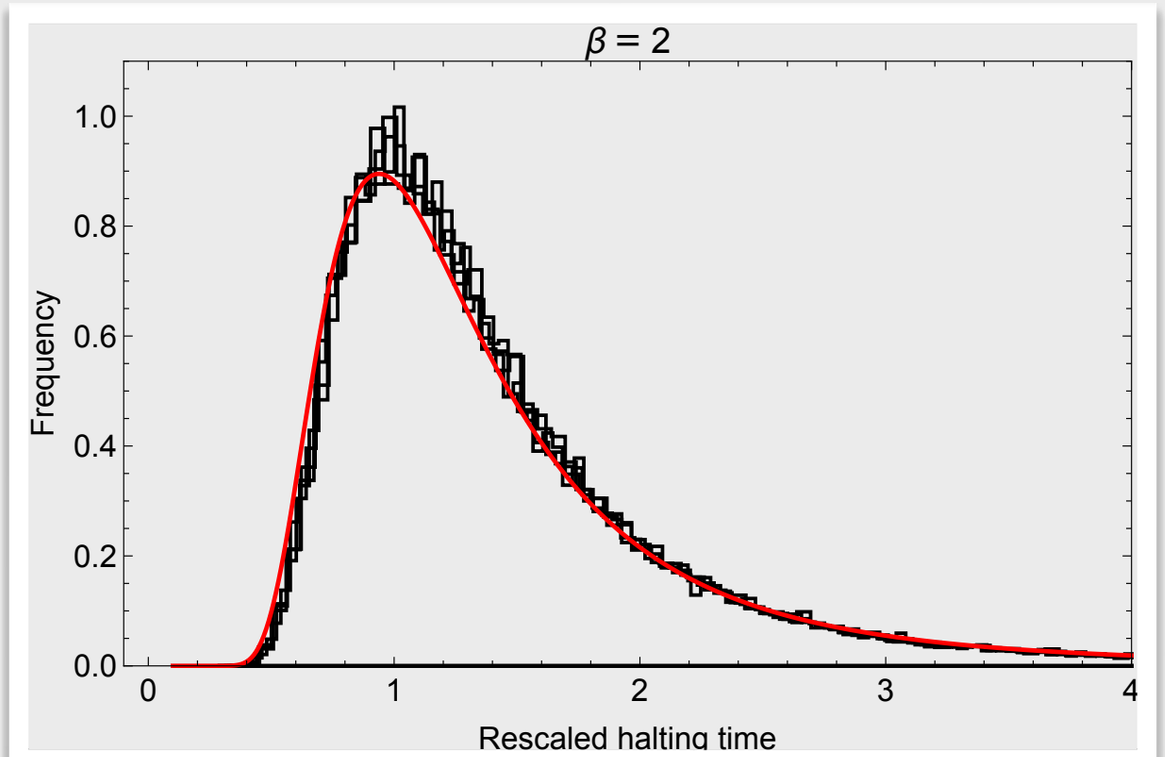
Theorem (Deift and T). Let W be a real ($\beta = 1$) or complex ($\beta = 2$) SCM and let $\mathbf{x}_0$ be a random unit vector independent of W. Assuming $\epsilon \leq N^{-5/3-\sigma}$, $\sigma > 0$,

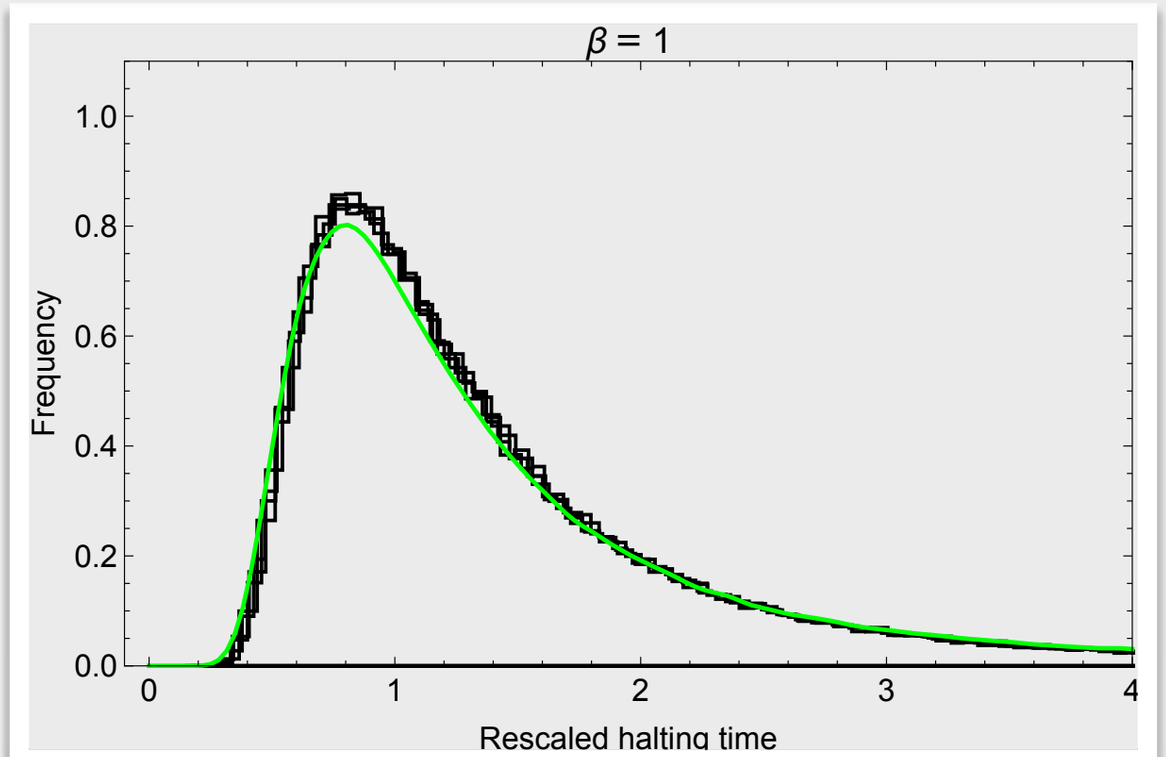
$$\lim_{N\to\infty} \left(\frac{T(W, \mathbf{x}_0, \epsilon)}{2^{-7/6} \lambda_+^{1/3} d^{-1/2} N^{2/3} (\log \epsilon^{-1} - 2/3 \log N)} \leq t \right) = F_\beta^{\mathrm{gap}}(t).$$

P Deift and T T. Universality for Eigenvalue Algorithms on Sample Covariance Matrices. SIAM J. Numer. Anal., 55(6):2835–2862, jan 2017

A demonstration

$N = 500$





Thanks to Folkmar Bornemann (TUM)

$$\lim_{N\to\infty}\left(\frac{T(W,\mathbf{x}_0,\epsilon)}{2^{-7/6}\lambda_+^{1/3}d^{-1/2}N^{2/3}(\log\epsilon^{-1}-2/3\log N)}\le t\right)=F_\beta^{\mathrm{gap}}(t)$$

Universal Universality

Other results

Similar techniques give universality theorems for:

- The inverse power method
- The QR eigenvalue algorithm
- The Toda algorithm
- Gradient descent

P Deift and T T. Universality for the Toda Algorithm to Compute the Largest Eigenvalue of a Random Matrix. Commun. Pure Appl. Math., 71(3):505–536, mar 2018

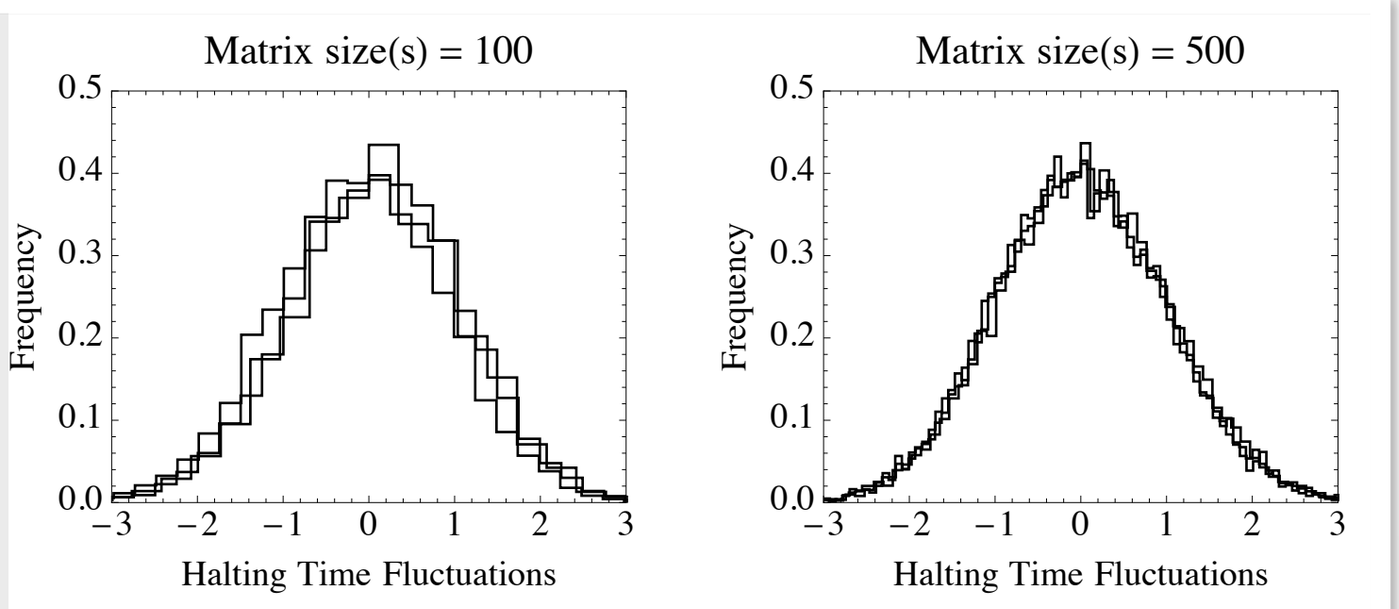
P Deift and T T. Universality for Eigenvalue Algorithms on Sample Covariance Matrices. SIAM J. Numer. Anal., 55(6):2835–2862, jan 2017

L Sagun, T T, and Y LeCun. Universal halting times in optimization and machine learning. Q. Appl. Math., 76(2), 2015

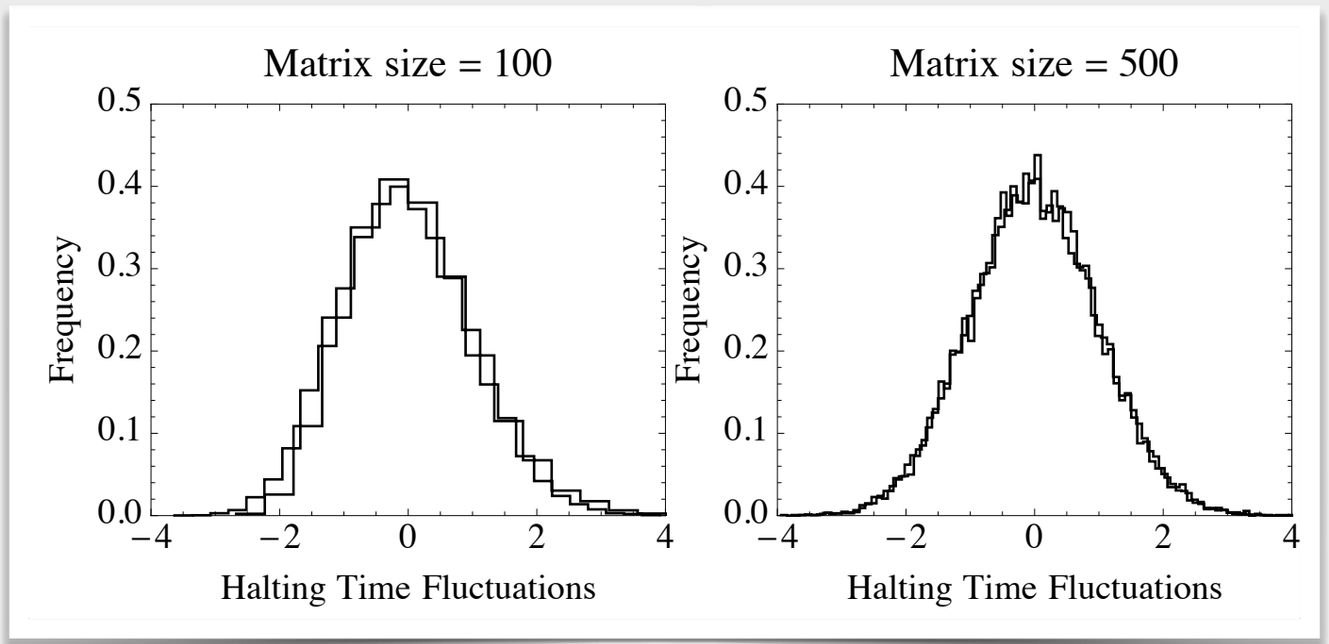
Other algorithms

w/ Percy Deift, Govind Menon and Sheehan Olver

Conjugate gradient algorithm

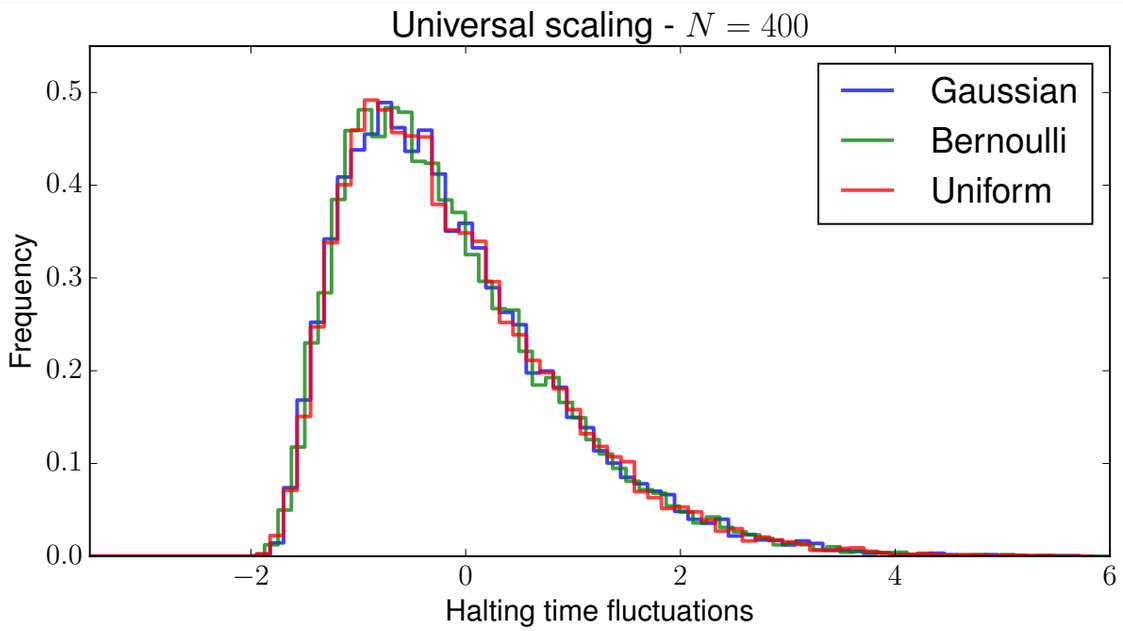


GMRES algorithm

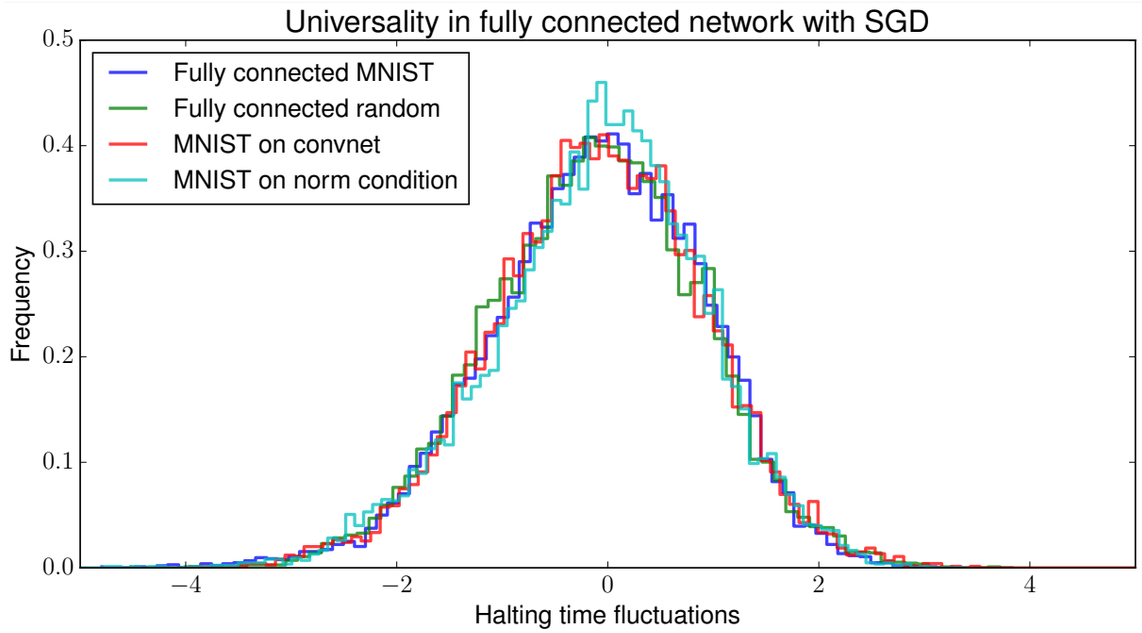


Other algorithms

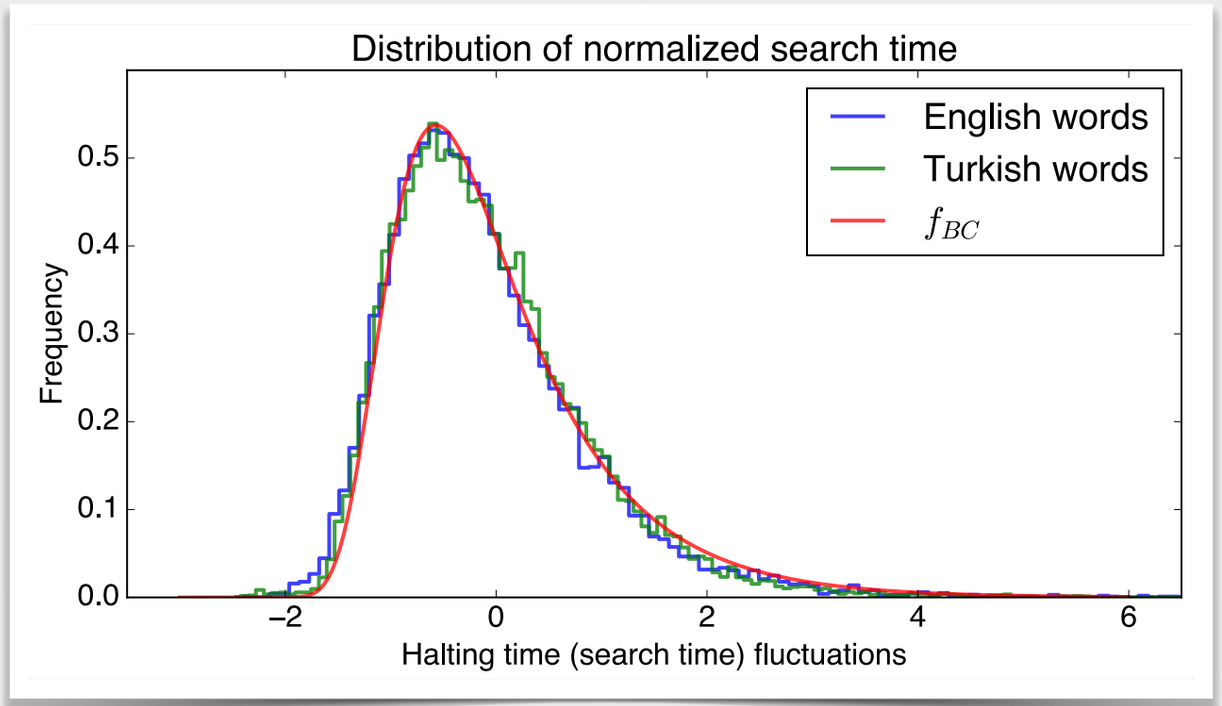
w/ Levent Sagun and Yann LeCun



Gradient descent for spin glasses



Stochastic gradient descent in machine learning



Google search times

The conjugate gradient algorithm (smoothed analysis)

The setup for conjugate gradient

The ensemble: Let $W \sim \mathrm{LUE}(N, M)$, $M \sim N/d$, $d = 1 - cN^{-\alpha}$, $0 \leq \alpha, c < 1$,

The method: The conjugate gradient algorithm is an iterative method to solve $W\mathbf{x} = \mathbf{b}$:

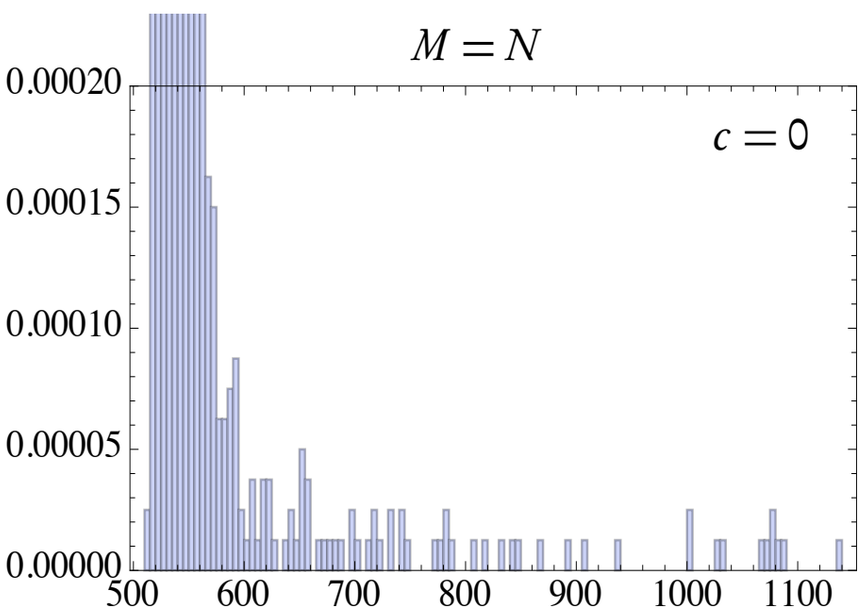
$$0 = \mathbf{x}_0 \to \mathbf{x}_1 \to \cdots \to \mathbf{x}_k \to \cdots \to \mathbf{x}_N = \mathbf{x}.$$

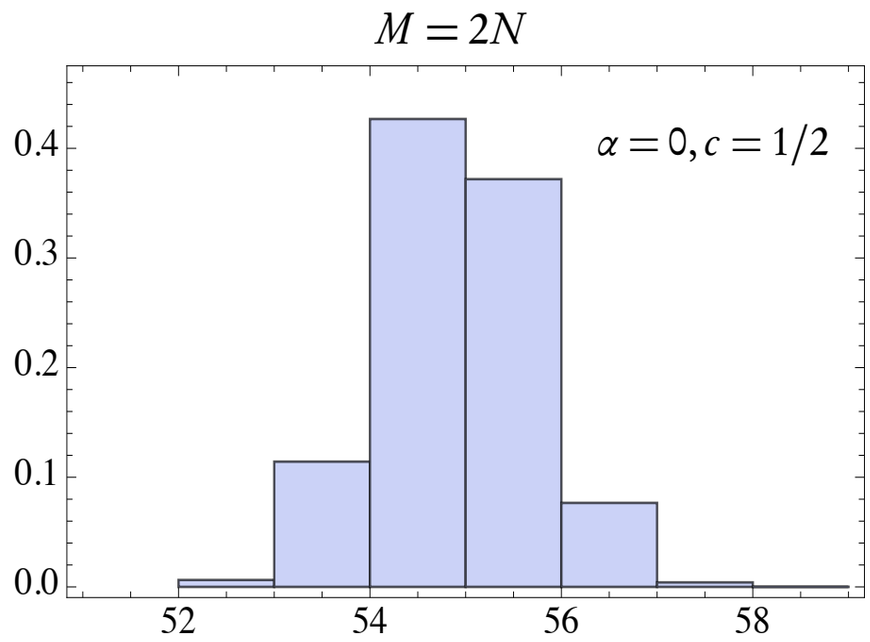
The halting time: The halting time is given by

$$T(W, \mathbf{b}, \epsilon) = \min\{k : ||W\mathbf{x}_k - \mathbf{b}||_2 < \epsilon\}.$$

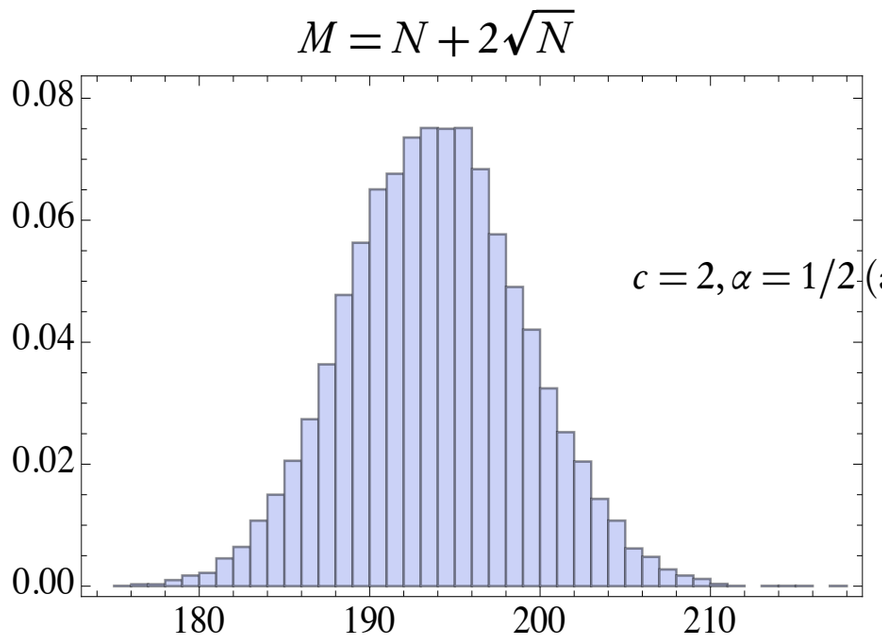
The general question: What is the behavior of T? Can we determine either its (asymptotic) distribution or estimate its moments?

The performance of the conjugate gradient algorithm





$N = 500$



$c = 2, \alpha = 1/2$ (asymptotically)

The scaling

Here we have chosen $d = 1 - cN^{-\alpha}$.

The rationale for this is the following:

- If $d < 1$ is fixed ($\alpha = 0, 0 < c < 1$), then X has a fixed aspect ratio (well conditioned, concentrated & bounded condition number).
- If $d = 1$ is fixed ($c = 0$) then W has a heavy–tailed condition number (ill conditioned, unbounded & unconcentrated).

By varying α one can describe the transition from the ill-conditioned to the well-conditioned case (from heavy-tailed to sub-exponential distributions for the condition number).

The scaling

If $d = 1 - (4c)^{-1/2} N^{-1/2}$ then (see Deift, Menon & T. (2017))

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{\kappa - 4Nc^{-1}}{4c^{-4/3}N^{2/3}} \leq t \right) = F_2(t) \quad \Leftarrow \quad \text{Tracy–Widom dist.}$$

The condition number grows with N but the (limit) distribution has sub-exponential tails, giving a finite N matrix that gives the Borodin–Forrester transition.

P A Deift, G Menon, and T T. On the condition number of the critically-scaled Laguerre Unitary Ensemble. Discret. Contin. Dyn. Syst., 36(8):4287–4347, mar 2016

Smoothed analysis for the CG algorithm

Theorem (Menon and T (2016)). As $N \to \infty$, for $1/2 \leq \alpha < 1$,

$$\sup_{||A|| \leq 1, A \geq 0} \mathbb{E}\left[T^j(A + \sigma^2 W, \mathbf{b}, \epsilon) \right] = O\left(N^{\alpha j} \left(1 + \frac{1}{\sigma^2}\right)^j \log^j \left[N^\alpha \left(1 + \frac{1}{\sigma^2}\right) \epsilon^{-1} \right] \right).$$

In particular,

$$\mathbb{E}\left[T(W, \mathbf{b}, \epsilon) \right] = O\left(N^\alpha \log\left[N^\alpha \epsilon^{-1} \right] \right),$$

and the right-hand side is less than N (for suff. large N).

The proof uses tail estimates on the condition number derived from the Riemann–Hilbert analysis in Deift, Menon, and T (2016).

G Menon and T T. Smoothed analysis for the conjugate gradient algorithm. Symmetry, Integr. Geom. Methods Appl., 12, 2016

P A Deift, G Menon, and T T. On the condition number of the critically-scaled Laguerre Unitary Ensemble. Discret. Contin. Dyn. Syst., 36(8):4287–4347, mar 2016

Smoothed analysis for the CG algorithm

Let $\mathbf{x}_k$ be the approximation of $\mathbf{x}$, $W\mathbf{x} = \mathbf{b}$ at iteration k. In exact arithmetic for GMRES, it is known from the work of Greenbaum, Pták & Strakŏs that any non-increasing sequence for

$$\frac{||\mathbf{x}_k - \mathbf{x}||}{||\mathbf{x}_0 - \mathbf{x}||}, \quad k = 1, 2, \ldots, N,$$

is attainable by choosing W and $\mathbf{b}$ appropriately (W may not be normal).

A pathological case: For $\epsilon > 0$,

$$\frac{||\mathbf{x}_k - \mathbf{x}||_W}{||\mathbf{x}_0 - \mathbf{x}||_W} > 1 - \epsilon, \quad k = 1, 2, \ldots, N-1,$$

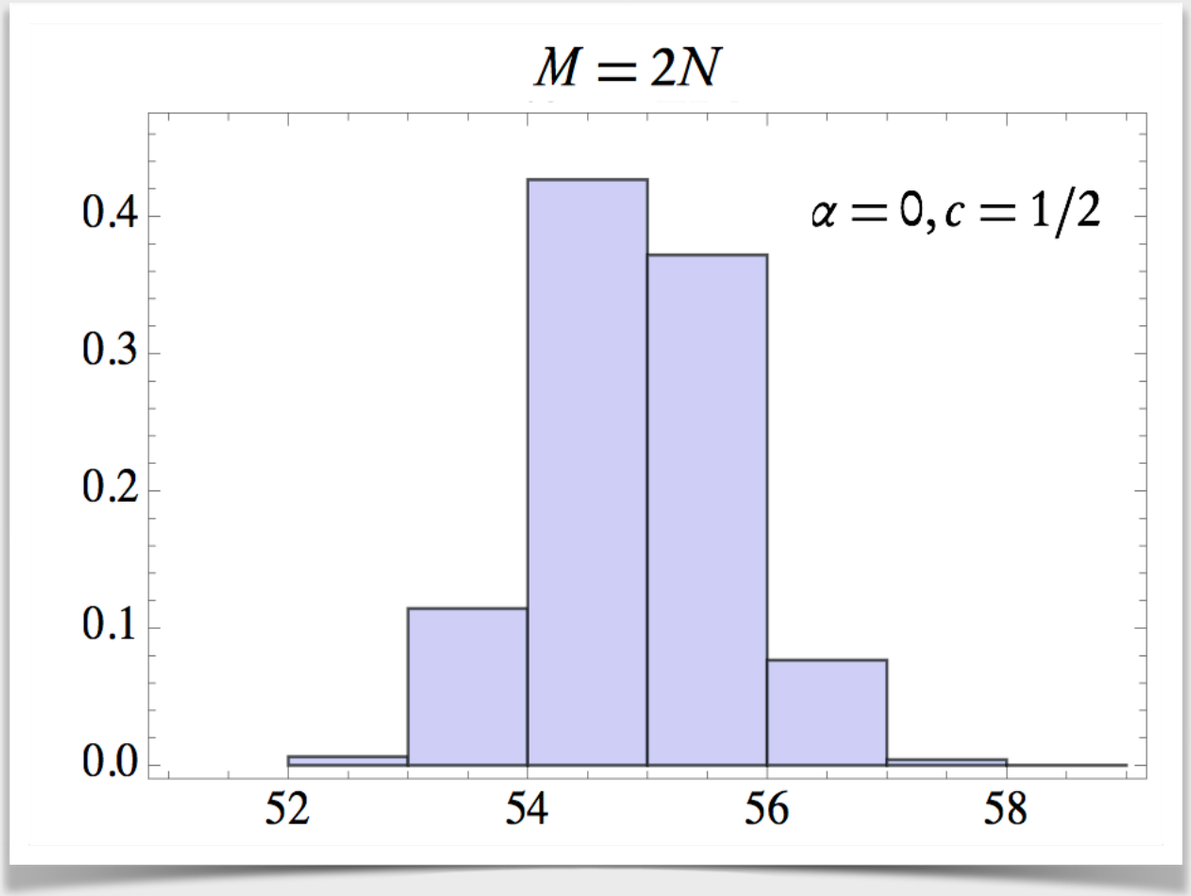
and, of course, $||\mathbf{x}_N - \mathbf{x}||_W = 0$.

For CG: The probability of such a pathalogical case decays faster than N^{-k} for any $k > 0$.

A Greenbaum, V Pták, and Z Strakoš. Any Nonincreasing Convergence Curve is Possible for GMRES. SIAM J. Matrix Anal. Appl., 17(3):465–469, 1996

The conjugate gradient algorithm (error concentration)

The well-conditioned case



$N = 500$

Despite the fact that we are using the conjugate gradient algorithm on a random linear system, the number of iterations seems almost deterministic.

The conjugate gradient algorithm on well-conditioned SCMs is almost deterministic

Theorem (Deift and T (2019)). Suppose $0 < d < 1$ ($\alpha = 0, 0 < c < 1$) and the conjugate gradient algorithm is applied to solve $W\mathbf{x} = \mathbf{b}$ where $W \sim \mathrm{LUE}(N, M)$, $\mathrm{LOE}(N, M)$, $M = N/d$. Suppose that $\mathbf{b}$ is a (possibly) random unit vector, independent of W. Then as $N \to \infty$

$$\mathbb{P}\left(\left|\|\mathbf{x}-\mathbf{x}_k\|_W - \frac{d^{k/2}}{\|\mathbf{b}-W\mathbf{x}_k\|_2}\right| \geq \epsilon\right) \leq Ce^{-cN} \qquad \mathbb{P}\left(\left|\|\mathbf{b}-W\mathbf{x}_k\|_2 - d^{k/2}\right| \geq \epsilon\right) \leq Ce^{-cN}$$

$$\|\mathbf{x}-\mathbf{x}_k\|_W^2 \to \frac{d^k}{1-d} \text{ in probability, and } T(W,\mathbf{b},\epsilon) \sim \left\lceil \frac{2\log\epsilon}{\log d} \right\rceil \text{ effectively (in a probabilistic sense).}$$

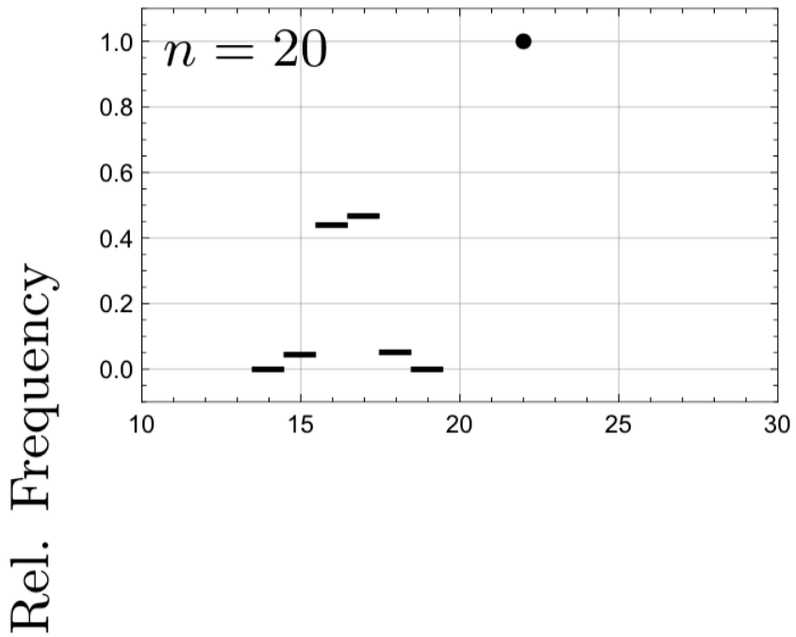
Consequently, for fixed ϵ, $T(W, \mathbf{b}, \epsilon)$ converges in probability to an explicit constant.

With high probability, one can predict the number of iterations that the conjugate gradient algorithm will require!

The proof uses connections between Householder bidiagonalization, the Lanczos iteration, the invariance of random matrix distributions and the classical error formula for the conjugate gradient algorithm.

P Deift and T T. The conjugate gradient algorithm on well-conditioned Wishart matrices is almost deteriministic. arXiv Prepr. arXiv1901.09007, 2019

A demonstration



Statistics for $T(W, \mathbf{b}, \epsilon)$

$\mathrm{LOE}(n, M)$, $d = 0.2$, $\epsilon \approx 6 \times 10^{-8}$
20,000 samples

The proof uses connections between Householder bidiagonalization, the Lanczos iteration, the invariance of random matrix distributions and the classical error formula for the conjugate gradient algorithm.

Fact 1. Let $\mathbf{x} = W^{-1}\mathbf{b}$. Then

$$\mathbf{x}_k = \operatorname{argmin}_{y \in \mathcal{K}_k} \|\mathbf{x} - \mathbf{y}\|_W, \quad \|\mathbf{y}\|_W^2 = \mathbf{y}^* W \mathbf{y},$$
$$\mathcal{K}_k = \operatorname{span}\left\{\mathbf{b}, W\mathbf{b}, \ldots, W^{k-1}\mathbf{b}\right\}$$

Fact 2. Let $\mathbf{x} = W^{-1}\mathbf{b}$. Then

$$\|\mathbf{x} - \mathbf{x}_k\|_W = \operatorname{argmin}_{p \in \mathbb{P}_k^{(0)}} \|p(W)\mathbf{x}\|_W,$$
$$\mathbb{P}_k^{(0)} = \{\text{polynomials } p \text{ of degree} \le k, p(0) = 1\}.$$

Let $p_k^\dagger$ be the minimizer.

Fact 3. Let T_k be the matrix obtained after the k step of the Lanczos iteration applied to W with starting vector $\mathbf{b}$. Then

$$p_k^\dagger(\lambda) = \frac{\det(T_k - \lambda I)}{\det T_k}.$$

Fact 4. In exact arithmetic, the Lanczos iteration run to completion coincides with Householder tridiagionalization if $\mathbf{b} = [1, 0, \ldots, 0]^T$. For LOE and LUE, the distribution of T_k is known explicitly.

Fact 5. Let u_j denote the first component of the jth normalized eigenvector of W. Also, let

$$\mu_N = \sum_{j=1}^{N} \delta_{\lambda_j} u_j^2$$

be the empirical spectral measure. Then

$$\|\mathbf{x} - \mathbf{x}_k\|_W^2 = \int_0^\infty \frac{\det(T_k - \lambda I)^2}{\lambda \det T_k^2} \mu_N(\mathrm{d}\lambda),$$

$$\|\mathbf{b} - W\mathbf{x}_k\|_2^2 = \int_0^\infty \frac{\det(T_k - \lambda I)^2}{\det T_k^2} \mu_N(\mathrm{d}\lambda).$$

Casual random matrix theory

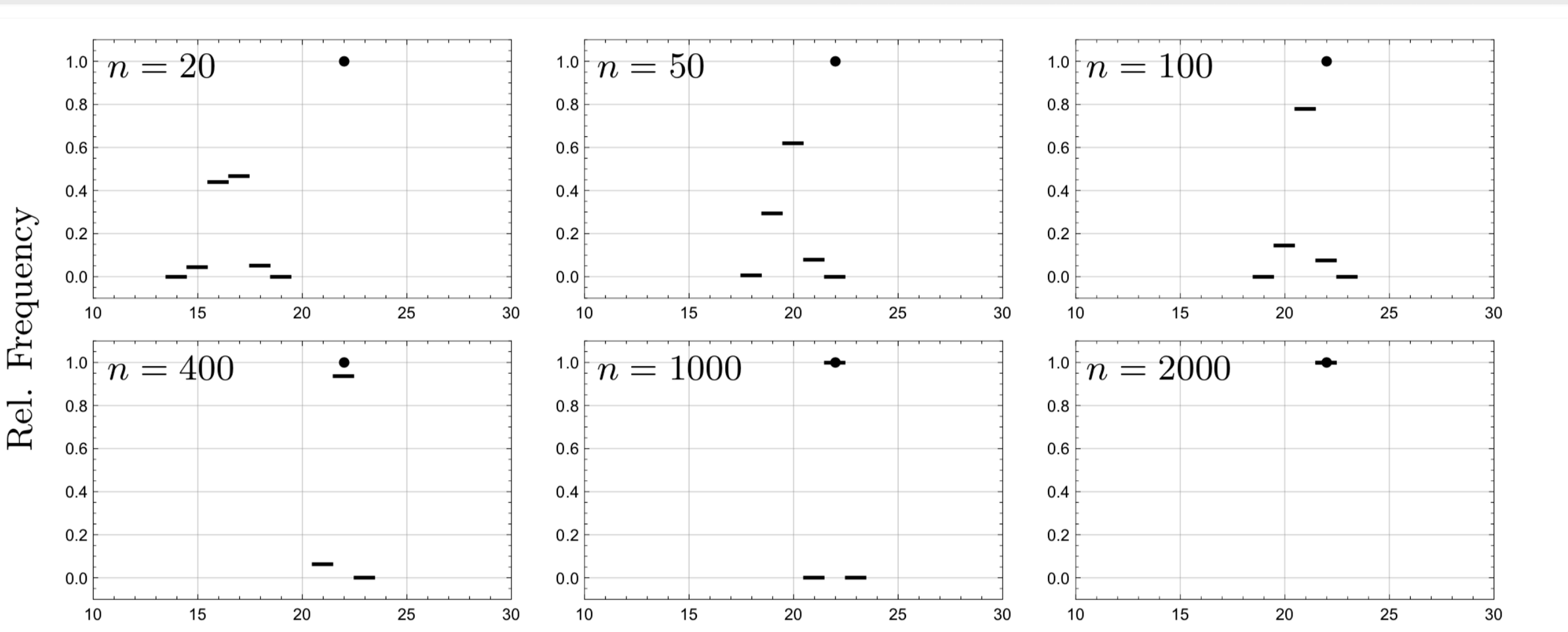
For LOE(N,M) and LUE(N,M) (and many others...)

$$
\begin{aligned}
T_k &\to \mathbb{T}_k, \\
\det T_k &\to 1, \\
\mu_N(\mathrm{d}\lambda) &\to \frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{2\pi d\,\lambda}\mathbb{1}_{[\lambda_-,\lambda_+]}(\lambda)\mathrm{d}\lambda, \quad \lambda_\pm = (1\pm\sqrt{d})^2.
\end{aligned}
$$

$$
\|\mathbf{x}-\mathbf{x}_k\|_W^2 = \int_0^\infty \frac{\det(T_k-\lambda I)^2}{\lambda \det T_k^2}\mu_N(\mathrm{d}\lambda) \to \frac{1}{2\pi d}\int_{\lambda_-}^{\lambda_+} \det(\mathbb{T}_k-\lambda I)^2 \frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda^2}\mathrm{d}\lambda = \frac{d^k}{1-d}
$$

$$
\|\mathbf{b}-W\mathbf{x}_k\|_2^2 = \int_0^\infty \frac{\det(T_k-\lambda I)^2}{\det T_k^2}\mu_N(\mathrm{d}\lambda) \to \frac{1}{2\pi d}\int_{\lambda_-}^{\lambda_+} \det(\mathbb{T}_k-\lambda I)^2 \frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda}\mathrm{d}\lambda = d^k
$$

A demonstration



Statistics for $T(W, \mathbf{b}, \epsilon)$

$\mathrm{LOE}(n, M)$, $d = 0.2$, $\epsilon \approx 6 \times 10^{-8}$
20,000 samples