[1]

HOW WELL–CONDITIONED CAN THE EIGENVECTOR PROBLEM BE?

CARLOS BELTRÁN, LAURENT BÉTERMIN, PETER GRABNER, AND STEFAN STEINERBERGER

Abstract. The condition number for eigenvector computations is a well–studied quantity. But how small can it possibly be?: Specifically, what matrices are perfectly conditioned with respect to eigenvector computations? In this note we answer this question for $n \times n$ matrices, giving a solution that is exact to first-order as $n \to \infty$.

1. Introduction and Result

1.1. **The condition number.** The sensitivity of the solution to a numerical problem under perturbations in the input can often be quantified by the *condition number*. Following [4, Overture], we first briefly recall the general definition of condition number that has its origins in the work of Turing, Von Neumann and Goldstein [10, 12]. Let φ be any mapping codifying the output of a numerical problem. For example, φ could take a square nonsingular matrix A and send it to $x = \varphi(A) = A^{-1}b$, the solution of $Ax = b$ where b is fixed a priori. Consider now the perturbed output $\hat{x} = \varphi(\hat{A})$ for a perturbed input $\hat{A} \approx A$. The condition number of φ measures $\|x - \hat{x}\|$ compared to $\|A - \hat{A}\|$, for appropriate choices of norms (or distance functions) in the input and output spaces. Formally, the standard definition for the condition number is (see for example [4, Def. (O.1)]:

$$\kappa(\varphi, A) = \lim_{\varepsilon \to 0} \sup_{\|A - \hat{A}\| \leq \varepsilon} \frac{\|x - \hat{x}\|}{\|A - \hat{A}\|}, \tag{1}$$

[1]SubjClass: Primary 65F15, Secondary 31C20

Facultad de Ciencias, Universidad of Cantabria, 39005 Santander, Spain. CB is Partially supported by Ministerio de Economía y Competitividad, Gobierno de España, through grant PID2020-113887GB-I00; and by Banco de Santander and Universidad de Cantabria, through grant 21.SI01.64658. beltranc@unican.es .

Institut Camille Jordan, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France. LB was supported by the Austrian Science Fund (FWF) and the German Research Foundation (DFG) through the joint project FR 4083/3-1/I 4354 during his stay at the University of Vienna. betermin@math.univ-lyon1.fr .

Institute of Analysis and Number Theory, Graz University of Technology, Kopernikusgasse 24, 8010 Graz, Austria. PG is supported by the Austrian Science Fund FWF project F5503 (part of the Special Research Program (SFB) "Quasi-Monte Carlo Methods: Theory and Applications"). peter.grabner@tugraz.at .

Department of Mathematics, University of Washington, Seattle, USA. SS is partially supported by the NSF (DMS-2123224) and the Alfred P. Sloan Foundation (FG-2021-14114). steiner@uw.edu .

or, more generally, for distance functions (for example, metrics of relative error) instead of norms,

$$\kappa(\varphi, A) = \lim_{\varepsilon \to 0} \sup_{\text{distance}(A,\hat{A}) \leq \varepsilon} \frac{\text{distance}(x, \hat{x})}{\text{distance}(A, \hat{A})}.$$

This definition is obviously equivalent to: $\kappa(\varphi, A)$ is the smallest number with the property that, for $\|A - \hat{A}\|$ very small,

$$\|x - \hat{x}\| \leq \kappa(\varphi, A)\|A - \hat{A}\| + o(\|A - \hat{A}\|).$$

When φ is fixed to be $\varphi(A) = A^{-1}b$, we usually drop it for the notation, writing only $\kappa(A)$. When solving for $Ax = b$ with b fixed, it is customary to choose the operator norm ($\|\cdot\|_{\text{op}}$) or the Frobenius norm ($\|\cdot\|_F$) in the space of matrices and to measure distances in relative error, which yields respectively Turing's [10] and Demmel's [6] condition numbers

$$\kappa(A) = \|A\|_{\text{op}} \|A^{-1}\|_{\text{op}} \qquad \text{and} \qquad \kappa_D(A) = \|A\|_F \|A^{-1}\|_{\text{op}}.$$

Sometimes φ is only defined locally. For example if φ sends a matrix A to some eigenvalue λ, then perturbing A will produce another matrix $\hat{A}$ with (in principle) several eigenvalues, but we must denote by $\hat{\lambda}$ only the eigenvalue of $\hat{A}$ that is obtained by continuation from λ (so, if λ has algebraic multiplicity 2 or higher, in general we cannot expect $\hat{\lambda}$ to be well–defined and the condition number is set to ∞ in this case). It is usual to make this fact explicit in the notation, so that $\kappa_\lambda(A)$ will measure the sensitivity of the eigenvalue λ under perturbations in A, but A may well have another eigenvalue λ' with a different value for $\kappa_{\lambda'}(A)$. A similar reasoning applies to eigenvectors.

1.2. **The condition number for the eigenproblem.** The sensitivity of eigenvalue and eigenvector computations is a well–studied topic in the Linear Algebra literature, see for example [1, 5, 13]. Throughout this paper, we measure the stability of an eigenvalue problem as follows: given $Ax = \lambda x$ and a nearby solution $\hat{A}\hat{x} = \hat{\lambda}\hat{x}$, the difference between the eigenvalues is given by $|\lambda - \hat{\lambda}|$. Eigenvectors really describe subspaces which makes the angle between subspaces a natural notion of distance: we will be interested in matrices where all eigenvalues are simple which naturally leads to the angle $\angle(x, \hat{x})$ defined by

$$\cos\left(\angle(x, \hat{x})\right) = \frac{|\langle x, \hat{x}\rangle|}{\|x\| \cdot \|\hat{x}\|}.$$

Following [11], we recall the explicit expression that arises from (1) in this setting: for a matrix $A \in \mathbb{C}^{n\times n}$ and an eigenpair $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$, the classical Schur decomposition yields a unitary matrix Q such that

$$Q^H A Q = \begin{pmatrix} \lambda & w^H \\ 0 & B \end{pmatrix}, \tag{2}$$

where $\cdot^H$ denotes the Hermitian conjugate and $w \in \mathbb{C}^{n-1}$ is a vector. Denoting by $y \in \mathbb{C}^n$ the corresponding left eigenvector, the condition numbers for the eigenvalue λ and eigenvector x, defined by the general approach (1), satisfy

$$\kappa_\lambda(A) = \frac{\|y\|\,\|x\|}{|y^H x|} \quad \text{and} \quad \kappa_x(A) = \frac{1}{\sigma_{\min}(B - \lambda \cdot \text{Id}_{(n-1)\times(n-1)})},$$

where $\sigma_{\min}$ denotes the least singular value and Id is the identity matrix. In other words, if we allow for an ε–size perturbation of A, the eigenpair $(\hat{\lambda}, \hat{x})$ of the perturbed matrix will satisfy

$$|\lambda - \hat{\lambda}| \leq \varepsilon \kappa_\lambda(A) + O(\varepsilon^2), \quad \angle(x, \hat{x}) \leq \varepsilon \kappa_x(A) + O(\varepsilon^2),$$

and $\kappa_\lambda(A)$ (resp. $\kappa_x(A)$) is the smallest number satisfying this property. We stress that these numbers contain information on how *one* particular pair (eigenvalue, eigenvector) changes under perturbations in the input. One can of course consider the maximum of all these quantities to deduce how the whole eigenproblem behaves under these perturbations. If the matrix $A \in \mathbb{C}^{n\times n}$ is diagonal with pairwise different entries $z_1, \ldots, z_n$, we have $\kappa_{z_i}(A) = 1$ for all i and the eigenvector condition number admits a simpler expression

$$\kappa_{e_i}(A) = \frac{1}{\min\limits_{j\neq i} |z_i - z_j|},$$

where $e_1, \ldots, e_n$ are the standard coordinate vectors.

1.3. **Main result.** In this note we investigate the following natural question: *how good can the eigenvector conditioning of an $n \times n$ matrix be?* The answer has a concrete application in the search for good starting points for homotopy methods for the eigenvector problem, see [2], but it is just such a natural and basic question that it deserves an answer on its own right! The answer is not trivial. For example, if $A = \mathrm{Id}_{n\times n}$ we clearly have $\kappa_x(A) = \infty$ (and it is a rather well–known fact that one can perturb the identity matrix $\mathrm{Id}_{n\times n}$ with very small changes to get any desired collection of eigenvectors). In general, one is interested in perturbations which are relative to the size of A, that is, perturbations of size $\varepsilon\|A\|_*$ where $\|A\|_*$ is either the operator or the Frobenius norm. In other words, since the stability of eigenvectors is given by

$$\sup_{\hat{A}:\, \frac{\|A-\hat{A}\|_*}{\|A\|_*} \leq \varepsilon} \angle(x, \hat{x}) \leq \varepsilon \kappa_x(A)\|A\|_* + o(\varepsilon),$$

the quantity of interest is precisely $\kappa_x(A)\|A\|_*$.

Problem. *What is the optimal value for the relative–error perturbation eigenvector conditioning of $A \in \mathbb{C}^{n\times n}$, that is, which matrix minimizes the quantity*

$$\kappa_{\max,*}(A) = \max_x \left(\kappa_x(A)\|A\|_*\right),$$

where x runs over all eigenvectors of A and $\|A\|_* = \|A\|_F$ or $\|A\|_* = \|A\|_{\mathrm{op}}$ is the Frobenius or the operator norm?

We recall the unit–side triangular (sometimes called hexagonal) lattice in $\mathbb{C} \equiv \mathbb{R}^2$ which is the set of points of the form

$$\begin{pmatrix} 1 & 1/2 \\ 0 & \sqrt{3}/2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}, \quad a, b \in \mathbb{Z}.$$

Our main result states that the diagonal matrix whose entries range over the triangular lattice, solves our problem to leading order.

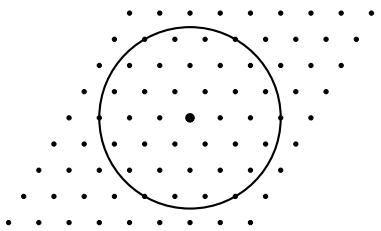

FIGURE 1. An extremal configuration (to leading order): a circle at the origin and the points of a triangular lattice inside the circle.

Theorem (Main Result)**.** *Let $\{z_1, \dots, z_n\} \subset \mathbb{C}^n$ be first n points in the unit–side triangular lattice, in increasing order of modulus (if two points have the same modulus, we take any of them). Then, as $n \to +\infty$,*

$$\kappa_{\max,F}(\mathrm{Diag}(z_1, \dots, z_n)) = \frac{3^{1/4}}{2\sqrt{\pi}} n + o(n),$$
$$\kappa_{\max,\mathrm{op}}(\mathrm{Diag}(z_1, \dots, z_n)) = \frac{3^{1/4}}{\sqrt{2\pi}} \sqrt{n} + o(\sqrt{n}).$$

Moreover, this diagonal matrix is asymptotically optimal in the sense that for any matrix $A \in \mathbb{C}^{n\times n}$ the equalities above give lower bounds for $\kappa_{\max,F}$ and $\kappa_{\max,\mathrm{op}}$.

It would naturally be interesting to have a better understanding of the error terms and, in particular, to have a better understanding of the extremal configurations. We believe it to be conceivable that our construction may perhaps be quite close to optimal even with respect to lower order error terms.

1.4. **Another type of stability.** Another standard choice for quantifying the stability of eigenvector computations is given by the condition number of the eigenvector matrix divided by the eigenvalue gap, that is

$$\frac{\inf_X \kappa(X)}{\min_{j\neq i} |z_i - z_j|},$$

where the infimum is taken over all X such that $A = XDX^{-1}$ for some diagonal D. This number has been recently exploited by Banks, Garza–Vargas, Kulkarni & Srivastava [3] with great success in the search for algorithms for eigendecompositions of matrices of norm at most 1. If the matrix A is normal and $\|A\|_* = 1$, the value of this new condition number is

$$\frac{\|A\|_*}{\min_{j\neq i} |z_i - z_j|},$$

that is exactly equal to ours! All in one, for unit norm normal matrices, these two notions yield the same result, and since the optimally conditioned matrix is diagonal in both cases, our result actually applies to this other standard.

1.5. **A General Inequality.** Our Theorem will follow relatively quickly from the following asymptotic inequality which may be of interest in its own right.

Proposition. *Let $p > 0$ be fixed. We have, for any $z_1, \dots, z_n \in \mathbb{C}$, as $n \to \infty$,*

$$\frac{1}{\min_{i \neq j} |z_i - z_j|} \left(\sum_{i=1}^{n} |z_i|^p \right)^{1/p} \geq \left(\frac{2}{p+2} \right)^{1/p} \frac{3^{1/4}}{\sqrt{2\pi}} n^{\frac{1}{2}+\frac{1}{p}} + o\left(n^{\frac{1}{2}+\frac{1}{p}} \right),$$

and this bound is matched by $z_1, \dots, z_n$ as in the Theorem.

Just as in the Theorem, it might be interesting to obtain a better understanding of the lower-order terms. If optimal configurations are indeed close to a subset of the hexagonal lattice, then this is strongly related to the Gauss Circle Problem and these techniques might apply (we observe that the function $z \to |z|^p$ is also smoother than the cut-off function used in the Gauss circle problem, so Fourier-based techniques might be useful).

2. Proofs

Section 2.1. gives the relatively short proof of the Theorem (assuming the Proposition). Section 2.2 contains a simple geometric lemma. Proving the Proposition will be the core of the argument, the proof is given in Section 2.3.

2.1. **Proof of the Theorem.**

Proof. From (2) it is clear that the eigenvector conditioning of a matrix is invariant under conjugation by unitary matrices. From the Schur decomposition, we can thus assume that A is upper–triangular. Now, for any eigenvector x the definition of $\kappa_x(A)$ does not involve w in (2), but w contributes to the norm $\|A\|_*$, so the value of $\kappa_{\max,*}$ does not grow by setting $w = 0$ for all eigenvectors. It follows that the matrix with optimal value of $\kappa_{\max,*}$ can be chosen diagonal (of course, conjugating it by any unitary matrix we obtain a normal matrix with identical conditioning). Thus, to prove the last claim of the Theorem we can assume that A is diagonal, but in this case we note that

$$\kappa_{\max,F}(\mathrm{Diag}(z_1, \dots, z_n)) = \max_{i \neq j} \frac{\|(z_1, \dots, z_n)\|_2}{|z_i - z_j|},$$

$$\kappa_{\max,\mathrm{op}}(\mathrm{Diag}(z_1, \dots, z_n)) = \max_{i \neq j} \frac{\|(z_1, \dots, z_n)\|_\infty}{|z_i - z_j|},$$

and the result is immediate from Proposition 1.5 for the cases $p = 2$ and $p = \infty$. □

2.2. **A Lemma.**

We denote a disk of radius r centered in the origin by

$$B_r = \{z \in \mathbb{C} : |z| < r\}.$$

We say that a set $\{z_1, \dots, z_n\} \subset \mathbb{C}$ is 1-separated if, for all $i \neq j$

$$|z_i - z_j| \geq 1.$$

We also introduce the counting function $N : [0, \infty] \to \mathbb{N}$ as follows: $N(r)$ is the cardinality of the largest possible 1-separated set contained in B_r. This quantity was already studied by L. Fejes Tóth in 1940 who determined its growth.

Lemma (Fejes Tóth [7])**.** *We have, as $r \to +\infty$,*

$$N(r) = \frac{2\pi}{\sqrt{3}} r^2 + o(r^2).$$

Equality to leading order is attained for the unit–side triangular lattice in B_r.

We give a simple sketch why this would be the case – the simple geometric argument makes use of Apollonian Circle Packings and the well–understood fact that the asymptotically densest packing of circles in the plane is given by the hexagonal lattice (which, not entirely coincidentally, is also a result of Fejes Tóth [8]).

Sketch of Proof. The claim on the triangular lattice (which implies the lower bound for the equality in the lemma) follows from [9, Eq. (1.6)]: for any lattice $L \subset \mathbb{R}^2$ (given by the set of points of the form $T\binom{a}{b}$ with $T \in \mathbb{R}^{2\times 2}$ and $a, b \in \mathbb{Z}$,

$$\#\{L \cap B_r\} = \frac{\pi r^2}{\det(T)} + O(r^{\frac{1}{3}}\sqrt{\log r}), \quad r \to +\infty,$$

where we use $\#S$ to denote the cardinality of a set S. In the case of the triangular lattice, $\det(T) = \sqrt{3}/2$ and we obtain the desired result. As for the upper bound, we argue by contradiction. Suppose that there exists $\varepsilon > 0$ such that there exists a divergent sequence $(r_n) \to \infty$ such that

$$N(r_n) \geq \left(\frac{2\pi}{\sqrt{3}} + \varepsilon\right) r_n^2.$$

Let us pick a sufficiently large r_n and let us tile $\mathbb{R}^2$ using disks of radius r_n in the fashion of a square or hexagonal lattice (it does not really matter). This allows us to cover a large amount of space with very efficient 1-separated point sets.

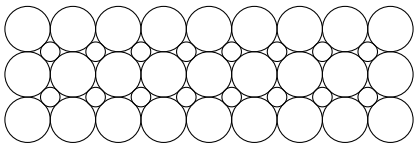

FIGURE 2. Using more disks with smaller radii allows for more precise approximation.

We then refine the disk packing by packing disks of smaller radius (see Fig. 2) between the big disks and use a standard hexagonal lattice to fill those disks. We can iterate this construction until we capture $\sim 1-$ of the entire plane. Note that, since r_n can be chosen to be arbitrarily large, this can always be done with a finite number of steps only depending on ε. However, this then allows us to generate a period packing of disks whose asymptotic density exceeds $\pi/\sqrt{12}$ which is a contradiction to the fact that the asymptotically densest packing is given by the hexagonal lattice (see Fejes Tóth [8]).

□

2.3. **Proof of the Proposition.**

Proof. Let $p > 0$ and let $\{z_1, \dots, z_n\} \subset \mathbb{R}^2$ be a 1-separated set. Our goal is to prove a lower bound on

$$\frac{1}{\min\limits_{i\neq j} |z_i - z_j|} \left(\sum_{i=1}^{n} |z_i|^p\right)^{1/p}.$$

We use invariance under dilation to assume without loss of generality that

$$\min_{i\neq j}|z_i - z_j| = 1$$

and this will be assumed in all subsequent arguments. We abbreviate

$$M := \max_{1\leq i\leq n}|z_i|.$$

Note that the Lemma implies that M cannot be too small: there cannot be too many $1-$separated points close to the origin. Indeed, the Lemma implies that for every $\varepsilon > 0$ there is an n_0 such that for every $n \geq n_0$

$$M > \frac{3^{1/4} - \varepsilon}{\sqrt{2\pi}}\sqrt{n}. \tag{3}$$

We now write

$$\begin{aligned}
\sum_{i=1}^{n}|z_i|^p &= p\sum_{i=1}^{n}\int_0^{|z_i|} y^{p-1}\,dy \\
&= p\int_0^M y^{p-1}\cdot\#\left\{1\leq i\leq n : |z_i| > y\right\}\,dy \\
&= p\int_0^M y^{p-1}\cdot\left(n - \#\left\{1\leq i\leq n : |z_i|\leq y\right\}\right)\,dy \\
&\geq p\int_0^M y^{p-1}\max(n - N(y), 0)\,dy.
\end{aligned}$$

Using our lower bound on M, we have

$$\begin{aligned}
\sum_{i=1}^{n}|z_i|^p &\geq p\int_0^{\frac{3^{1/4}-\varepsilon}{\sqrt{2\pi}}\sqrt{n}} y^{p-1}\max\left(n - \frac{2\pi}{\sqrt{3}}y^2 + o(y^2), 0\right)\,dy \\
&= \left(\frac{3^{1/4}-\varepsilon}{\sqrt{2\pi}}\sqrt{n}\right)^p\left(1 - \frac{p}{p+2}\frac{(3^{1/4}-\varepsilon)^2}{\sqrt{3}}\right)n + o\left(n^{\frac{p+2}{2}}\right),
\end{aligned}$$

which implies the Proposition. For $p = \infty$, the assertion is just equation (3). By observing that the bounds used for $N(r)$ are sharp to leading order for the hexagonal lattice, we see that all the inequalities are asymptotically sharp which establishes the second half of the statement. However, the upper bound could also be established using a direct computation, valid for $p > 1$, using the construction in [2, Lemma 2.27]: For any given n, we take the first n points in the unit-side triangular lattice, with increasing modulus (if two points have the same modulus, we take any of them). We first note that the function $z \to |z|^p$ is convex, hence by Jensen's inequality, for any regular hexagon H centered at z_i we have

$$\mathbb{E}_{z\in H}(|z|^p) \geq |\mathbb{E}_{z\in H}(z)|^p = |z_i|^p,$$

where $\mathbb{E}$ has to be understood as the expected value with respect to one Voronoi cell in the hexagon. Since by construction the points are in the unit-side triangular lattice, the Voronoi cells surrounding each z_i are hexagons H_i and we thus have

$$\sum_{i=1}^{n}|z_i|^p \leq \sum_{i=1}^{n}\mathbb{E}_{z\in H_i}(|z|^p) = \frac{1}{\mathrm{vol}(H)}\int_{\bigcup_i H_i}|z|^p\,dz.$$

(note that all the hexagons have the same area, which we denote by $\mathrm{vol}(H) = \sqrt{3}/2$). Now, from the Lemma it follows that in the disk of radius r there are at

least $2\pi r^2/\sqrt{3}+O(r^{1/3}\sqrt{\log r})$ points of the unit–side triangular lattice. Therefore, it follows that all the H_i are contained in a disk of radius $r_n := 3^{1/4}\sqrt{n}/\sqrt{2\pi}+o(n)$, which yields

$$\sum_{i=1}^{n} |z_i|^p \leq \frac{1}{\mathrm{vol}(H)} \int_{|z|\leq r_n} |z|^p \, dz$$
$$= \frac{2\pi r_n^{p+2}}{(p+2)\mathrm{vol}(H)} = \frac{4\pi 3^{\frac{p+2}{4}}}{\sqrt{3}(p+2)(2\pi)^{1+p/2}} n^{1+p/2} + o(n^{1+p/2}),$$

which proves the upper bound. □

Acknowledgments

This work was initiated during the workshop "Minimal energy problems with Riesz potentials" held at the American Institute of Mathematics in May 2021. The authors are very grateful to two anonymous referees for their valuable remarks and suggestions.

References

[1] R. Alam and S. Bora, *On sensitivity of eigenvalues and eigendecompositions of matrices.* Linear Algebra Appl. **396** (2005) 273–301.
[2] D. Armentano, C. Beltrán, P. Bürgisser, F. Cucker, and M. Shub, *A stable, polynomial-time algorithm for the eigenpair problem*, J. Eur. Math. Soc. (JEMS) **20** (2018), 1375–1437.
[3] J. Banks, J. Garza–Vargas, A. Kulkarni and N. Srivastava. Pseudospectral Shattering, the Sign Function, and Diagonalization in Nearly Matrix Multiplication Time. *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS), Vol 1, pp. 529–540*
[4] P. Bürgisser, F. Cucker. Condition: The Geometry of Numerical Algorithms. Springer, 2013.
[5] J. W. Demmel, *On Condition Numbers and the Distance to the Nearest Ill-posed Problem*, Numer. Math., **51** (1987), 251–289.
[6] J. W. Demmel, *The probability that a numerical analysis problem is difficult*, Math. Comp., **50** (1988), 449–480
[7] L. Fejes, *Über einen geometrischen Satz*, Math. Z. **46** (1940), 83–85.
[8] L. Fejes, *Über die dichteste Kugellagerung*, Math. Z. **48** (1943), 676–684.
[9] P. D. Lax and R. S. Phillips, *The asymptotic distribution of lattice points in Euclidean and non-Euclidean spaces*, J. Functional Analysis **46** (1982), 280–350.
[10] A. M. Turing, *Rounding-off errors in matrix processes.* Quarterly Journal of Mechanics and Applied Mathematics, **1** (1948), 287–308.
[11] C. Van Loan, *On estimating the condition of eigenvalues and eigenvectors*, Linear Algebra Appl. **88/89** (1987), 715–732.
[12] J. von Neumann and H. H. Goldstine, *Numerical inverting matrices of high order.* Bull. Amer. Math. Soc. **53** (1947), 1021–1099.
[13] J. H. Wilkinson, *Note on Matrices with a Very Ill-Conditioned Eigenproblem.* Nume. Math. **19** (1972), 176–178.