The Asymptotics of Wilkinson's Shift: Loss of Cubic Convergence

Ricardo S. Leite, Nicolau C. Saldanha and Carlos Tomei

November 28, 2008

Abstract

One of the most widely used methods for eigenvalue computation is the QR iteration with Wilkinson's shift: here the shift s is the eigenvalue of the bottom 2×2 principal minor closest to the corner entry. It has been a long-standing conjecture that the rate of convergence of the algorithm is cubic. In contrast, we show that there exist matrices for which the rate of convergence is strictly quadratic. More precisely, let $T_{\mathcal{X}}$ be the 3×3 matrix having only two nonzero entries $(T_{\mathcal{X}})_{12} = (T_{\mathcal{X}})_{21} = 1$ and let $\mathcal{T}_\Lambda$ be the set of real, symmetric tridiagonal matrices with the same spectrum as $T_{\mathcal{X}}$. There exists a neighborhood $\boldsymbol{\mathcal{U}} \subset \mathcal{T}_\Lambda$ of $T_{\mathcal{X}}$ which is invariant under Wilkinson's shift strategy with the following properties. For $T_0 \in \boldsymbol{\mathcal{U}}$, the sequence of iterates (T_k) exhibits either strictly quadratic or strictly cubic convergence to zero of the entry $(T_k)_{23}$. In fact, quadratic convergence occurs exactly when $\lim T_k = T_{\mathcal{X}}$. Let $\boldsymbol{\mathcal{X}}$ be the union of such quadratically convergent sequences (T_k): the set $\boldsymbol{\mathcal{X}}$ has Hausdorff dimension 1 and is a union of disjoint arcs $\boldsymbol{\mathcal{X}}^\sigma$ meeting at $T_{\mathcal{X}}$, where σ ranges over a Cantor set.

Keywords: Wilkinson's shift, asymptotic convergence rates, symbolic dynamics.

MSC-class: 65F15; 37E99; 37N30.

1 Introduction

The QR iteration is a standard algorithm to compute eigenvalues of matrices in $\mathcal{T}$, the vector space of $n \times n$ real symmetric tridiagonal matrices ([15], [4], [13]). More precisely, consider $T \in \mathcal{T}$ and a shift $s \in \mathbb{R}$ so that $T - sI$ is an invertible matrix. Given the QR decomposition $T - sI = QR$, where Q is orthogonal and R is upper triangular with positive diagonal, the *shifted step* obtains a new matrix $\mathbf{F}(s, T) = Q^*TQ = RTR^{-1}$. Let $\boldsymbol{\omega}_-(T) \le \boldsymbol{\omega}_+(T)$ be the eigenvalues of the bottom 2×2 principal minor of T and let $\boldsymbol{\omega}(T)$ be the eigenvalue closer to the bottom entry $(T)_{nn}$. *Wilkinson's shift* is the choice $s = \boldsymbol{\omega}(T)$ and *Wilkinson's step* is $\mathbf{W}(T) = \mathbf{F}(\boldsymbol{\omega}(T), T)$, provided $\boldsymbol{\omega}(T)$ is well defined and is not an eigenvalue of T.

A matrix $T \in \mathcal{T}$ is *unreduced* if $(T)_{i,i+1} = (T)_{i+1,i} \neq 0$ for $1 \le i < n$. Recall that if T_0 is unreduced and the iterates $T_k = \mathbf{W}^k(T_0)$, $k \in \mathbb{N}$, are well defined then T_k is also unreduced and the bottom off-diagonal entry $(T_k)_{n-1,n}$ tends to 0. The quick convergence of Wilkinson's algorithm is a well known fact, as discussed in Section 8-11 of [13]: we are interested in the precise rate of convergence of this sequence. It has been conjectured ([13], [4]) that it should be *cubic*, i.e., $|(T_{k+1})_{n,n-1}| = O(|(T_k)_{n,n-1}|^3)$. Here, instead, we show that cubic convergence does not hold in general: there are unreduced matrices T_0 for which the rate of convergence of the sequence $(T_k)_{23}$ to 0 is, in the words of Parlett, merely quadratic.

Given $T \in \mathcal{T}$, the above formulae imply that $\mathbf{W}(T)$, if well defined, is symmetric and upper Hessenberg and therefore T and $\mathbf{W}(T)$ are matrices in $\mathcal{T}$ with the same

spectrum. For $\Lambda = \mathrm{diag}(1, -1, 0)$, denote by $\mathcal{T}_\Lambda$ the set of matrices in $\mathcal{T}$ similar to Λ and consider $\mathbf{W}$ as a map from $\mathcal{T}_\Lambda$ to $\mathcal{T}_\Lambda$. There are some technical aspects to consider. First, there are matrices $T \in \mathcal{T}_\Lambda$ with bottom entry $T_{3,3}$ equidistant from $\boldsymbol{\omega}_-$ and $\boldsymbol{\omega}_+$. This introduces a step discontinuity in the map $\mathbf{W}$: when T tends to such a matrix T_0, $\mathbf{W}(T)$ may approach either $\mathbf{F}(\boldsymbol{\omega}_-(T_0), T_0)$ or $\mathbf{F}(\boldsymbol{\omega}_+(T_0), T_0)$. Also, strictly speaking, the definition of $\mathbf{W}$ does not apply to matrices T for which $T_{23} = 0$ because then $\boldsymbol{\omega}(T)$ is an eigenvalue of T. It turns out, however, that $\mathbf{W}$ can be continuously extended to such matrices.

We introduce the notation required to state the main result of this paper. Let

$$T_{\mathcal{X}} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathcal{T}_\Lambda.$$

An *infinite sign sequence* is a function $\sigma : \mathbb{N} \to \{+, -\}$. Let Σ be the set of infinite sign sequences: Σ admits the natural metric $d(\sigma_1, \sigma_2) = 2^{-n}$ where n is the smallest number for which $\sigma_1(n) \neq \sigma_2(n)$. The metric space Σ is homeomorphic to the middle-third Cantor set contained in $[0, 1]$. For $\sigma \in \Sigma$, let $\sigma^\sharp \in \Sigma$ be obtained by deleting the first sign, i.e., $\sigma^\sharp(n) = \sigma(n + 1)$.

Theorem 1.1 *There is an open neighborhood $\boldsymbol{\mathcal{U}} \subset \mathcal{T}_\Lambda$ of $T_{\mathcal{X}}$ with the following properties.*

(a) If $T \in \boldsymbol{\mathcal{U}}$ then $\mathbf{W}(T) \in \boldsymbol{\mathcal{U}}$.

(b) If $T_0 \in \boldsymbol{\mathcal{U}}$ then the sequence $T_k = \mathbf{W}^k(T_0)$ converges to T_∞ with $(T_\infty)_{23} = 0$. If $T_\infty = T_{\mathcal{X}}$ then the convergence rate of $(T_k)_{23}$ to 0 is strictly quadratic; otherwise it is strictly cubic.

(c) Let $\boldsymbol{\mathcal{X}} \subset \boldsymbol{\mathcal{U}}$ be the set of initial conditions T_0 for which $T_\infty = T_{\mathcal{X}}$. Then $\boldsymbol{\mathcal{X}}$ is the union of arcs $\boldsymbol{\mathcal{X}}^\sigma$, $\sigma \in \Sigma$, which are disjoint except for the common point $T_{\mathcal{X}}$. Also, $\mathbf{W}(\boldsymbol{\mathcal{X}}^\sigma) \subseteq \boldsymbol{\mathcal{X}}^{\sigma^\sharp}$.

(d) The set $\boldsymbol{\mathcal{X}} \subset \boldsymbol{\mathcal{U}}$ has Hausdorff dimension 1.

We opted to make this paper as self-contained as reasonably possible at the expense of rendering several sections more technical. The inductive arguments employed to control the iteration make use of an assortment of explicit estimates. The parameters obtained are loose and might be replaced by a chain of existential arguments: we hope that our choice led to a clearer presentation.

In Section 2 we introduce an explicit chart $\phi : \mathbb{R}^2 \to \mathcal{T}_\Lambda$, satisfying $\phi(p_{\mathcal{X}}) = T_{\mathcal{X}}$ where $p_{\mathcal{X}} = (2, 0)$. Shifted steps in these (x, y) coordinates, i.e., the functions $F(s, x, y) = \phi^{-1}(\mathbf{F}(s, \phi(x, y)))$, admit a simple formula:

$$F(s, x, y) = \left(\frac{1 + s}{1 - s} x, \frac{|s|}{1 + s} y \right).$$

One may think of (x, y) coordinates as a variation of spectral data in the sense of [2], [3] and [7]: eigenvalues and the absolute values of the first coordinates of the normalized eigenvectors. The map ϕ is an instance of *bidiagonal coordinates* ([8]), which form an atlas of $\mathcal{T}_\Lambda$: ϕ is a chart whose image is an open dense set containing $T_{\mathcal{X}}$. In this chart, convergence issues reduce to local theory. The present text might serve as an illustration of the general construction: the downside is that the formula for ϕ is presented without a natural process leading to it.

The rather technical Section 3 is dedicated to the study of the sub-eigenvalues $\boldsymbol{\omega}_\pm(T)$ and their counterparts in (x, y) coordinates $\omega_\pm = \boldsymbol{\omega}_\pm \circ \phi$. We also introduce

the rectangle $\mathcal{R}_a = [2-a, 2+a] \times [-a/10, a/10]$, $0 < a \leq 1/10$, whose image $\boldsymbol{\mathcal{R}}_a = \phi(\mathcal{R}_a)$ is the closure of the invariant neighborhood $\boldsymbol{\mathcal{U}}$ in Theorem 1.1. The discontinuities of ω lie on the vertical line $x = 2$: if $x < 2$ (resp. $x > 2$), $\omega(x, y) = \omega_+(x, y)$ (resp. $\omega(x, y) = \omega_-(x, y)$).

In Section 4 we show that $\mathcal{R}_a$ is invariant under $W = \phi^{-1} \circ \mathbf{W} \circ \phi$ (for sufficiently small a). The discontinuous map W has smooth restrictions to the interior of the rectangles $\mathcal{R}_{a,+} = [2-a, 2] \times [-a/10, a/10]$ and $\mathcal{R}_{a,-} = [2, 2+a] \times [-a/10, a/10]$ which extend continuously to $W_\pm : \mathcal{R}_{a,\pm} \to \mathcal{R}_a$.

The rest of the proof of Theorem 1.1 proceeds by studying the iterations of W. At this point, the vocabulary and techniques from dynamical systems are natural (another application of dynamical systems to numerical spectral theory is the work of Batterson and Smillie [1] on the Rayleigh quotient iteration). In Section 5 we provide different characterizations of the set $\mathcal{X} = \phi^{-1}(\boldsymbol{\mathcal{X}})$ and show its nontriviality.

In Section 6 we construct a continuous map from $[-a/10, a/10] \times \Sigma \to \mathcal{X}$ taking a pair (y, σ) to the only point $p_0 = (g^\sigma(y), y) \in \mathcal{X}$ for which $p_k = W^k(p_0) \in \mathcal{R}_{a,\sigma(k)}$. Informally, the sign sequence σ specifies the side of the rectangle $\mathcal{R}_a$ in which p_k lies. This yields the decomposition of $\mathcal{X}$ as a union of the Lipschitz arcs $\mathcal{X}^\sigma$, $\sigma \in \Sigma$. Sharper estimates concerning $g^\sigma(y)$ are needed to prove that $\mathcal{X}$ is very thin, i.e., that its Hausdorff dimension is 1. The concept of Hausdorff dimension is only used at the very end of the argument in order to obtain a more concise formulation of an estimate on the number of balls of radius r required to cover $\mathcal{X}$.

2 Local coordinates and s-steps

The *QR decomposition* of an invertible matrix M is $M = \mathbf{Q}(M)\,\mathbf{R}(M)$, where $\mathbf{Q}(M)$ is orthogonal and $\mathbf{R}(M)$ is upper triangular with nonnegative diagonal. The *s-step* with shift s is the map

$$\mathbf{F}(s, M) = (\mathbf{Q}(M - sI))^*\, M\, \mathbf{Q}(M - sI) \tag{1}$$

$$= \mathbf{R}(M - sI)\, M\, (\mathbf{R}(M - sI))^{-1}. \tag{2}$$

This is well defined provided $M - sI$ is invertible.

As an example, to be used in the sequel, consider

$$T_{\mathcal{X}} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Clearly, the matrix $T_{\mathcal{X}} - sI$ is invertible if $s \neq 0, \pm 1$. For $s \neq 0$, $|s| < 1$, the QR decomposition of $T_{\mathcal{X}} - sI$ is

$$\mathbf{Q}(T_{\mathcal{X}} - sI)\ \mathbf{R}(T_{\mathcal{X}} - sI) = \begin{pmatrix} -ss_1 & s_1 & 0 \\ s_1 & ss_1 & 0 \\ 0 & 0 & -\mathrm{sign}(s) \end{pmatrix} \begin{pmatrix} s_1^{-1} & -2ss_1 & 0 \\ 0 & (1-s^2)s_1 & 0 \\ 0 & 0 & |s| \end{pmatrix}$$

where $s_1 = (1+s^2)^{-1/2}$. Notice that $\mathbf{Q}(T - sI)$ has a jump discontinuity at $s = 0$ while $\mathbf{R}(T_{\mathcal{X}} - sI)$ is continuously defined but ceases to be invertible at $s = 0$. Standard s-steps for $T_{\mathcal{X}}$ are given by

$$\mathbf{F}(s, T_{\mathcal{X}}) = \frac{1}{1+s^2} \begin{pmatrix} -2s & 1-s^2 & 0 \\ 1-s^2 & 2s & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Notice that this formula extends continuously to $s = 0$ with $\mathbf{F}(0, T_{\mathcal{X}}) = T_{\mathcal{X}}$. Such continuous extensions will be important throughout the paper.

Let $\mathcal{T}$ be the set of 3×3 real, symmetric, tridiagonal matrices. Let $\Lambda = \mathrm{diag}(1, -1, 0)$ and $\mathcal{T}_\Lambda = \{Q^*\Lambda Q, Q \in O(n)\}$ be the set of matrices in $\mathcal{T}$ similar to Λ. In fact, $\mathcal{T}_\Lambda$ is a bitorus ([14], [8]) but this kind of global information will not be used in this paper.

We now define the relevant parametrization of $\mathcal{T}_\Lambda$ near $T_\mathcal{X}$. Let $\mathcal{E}$ be the set of *signature matrices*, i.e., diagonal $n \times n$ matrices with diagonal entries equal to ± 1. An invertible matrix M is *LU-positive* if there exist lower and upper triangular matrices L and U with positive diagonals so that $M = LU$; equivalently, M is LU-positive if its leading principal minors have positive determinant. For instance, the only LU-positive orthogonal matrix Q with $T_\mathcal{X} = Q^*\Lambda Q$ is

$$Q = Q_\mathcal{X} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Proposition 2.1 *Set $p_\mathcal{X} = (2, 0)$ and $\phi : \mathbb{R}^2 \to \mathcal{T}$,*

$$\phi(x, y) = \frac{1}{r_1^2 r_2^2} \begin{pmatrix} (4 - x^2)r_2^2 & 2xr_2^3 & 0 \\ 2xr_2^3 & -4(4 - x^2 - 4x^2y^4 + x^4y^4) & 2yr_1^3 \\ 0 & 2yr_1^3 & y^2(x^2 - 4)r_1^2 \end{pmatrix},$$

where $r_1 = \sqrt{4 + x^2 + 4x^2y^2}$ and $r_2 = \sqrt{4 + 4y^2 + x^2y^2}$. Then $\phi(p_\mathcal{X}) = T_\mathcal{X}$ and there exist neighborhoods $\mathcal{U}_1 \subset \mathbb{R}^2$ of $p_\mathcal{X}$ and $\mathcal{B}_1 \subset \mathcal{T}$ of $T_\mathcal{X}$ such that $\phi|_{\mathcal{U}_1}$ is an immersion with image $\boldsymbol{\mathcal{U}}_1 = \mathcal{T}_\Lambda \cap \mathcal{B}_1$. Also, $\mathrm{sign}(x) = \mathrm{sign}((\phi(x, y))_{12})$ *and* $\mathrm{sign}(y) = \mathrm{sign}((\phi(x, y))_{23})$. *Finally, for matrices $T = \phi(x, y) \in \boldsymbol{\mathcal{U}}_1$, the only LU-positive orthogonal matrix for which $T = Q^*\Lambda Q$ is*

$$Q = \frac{1}{r_1 r_2} \begin{pmatrix} 2r_2 & 2x(1 + 2y^2) & xyr_1 \\ -xr_2 & 2(2 + x^2y^2) & -2yr_1 \\ -2xyr_2 & y(4 - x^2) & 2r_1 \end{pmatrix}.$$

Since we are interested in the behavior of Wilkinson's step near $T_\mathcal{X}$, we successively define nested neighborhoods $\boldsymbol{\mathcal{U}}_k$ of $T_\mathcal{X}$ and $\mathcal{U}_k$ of $p_\mathcal{X}$: the process stops with the contruction of the rectangle $\mathcal{R}_a$ at Proposition 3.3. Constructions on matrices use boldface symbols.

Proof: The example above obtains $T_\mathcal{X} = \phi(p_\mathcal{X})$. The remaining statements also follow from numerical checks, but instead we provide motivation for the construction of the map ϕ. Given the eigenvalue matrix Λ, any matrix $T \in \mathcal{T}_\Lambda$ admits an orthogonal diagonalization $T = Q^*\Lambda Q$, which is not unique: all other orthogonal diagonalizations are of the form $T = Q^*E\Lambda EQ$ for some signature matrix $E \in \mathcal{E}$.

Now, the rows of Q are normal eigenvectors of T, so its columns q_1, q_2 and q_3 are also orthonormal. For vectors q and w, we write $q \sim w$ to indicate that they are collinear. Say $q_1 \sim w_1 = (2, -x, -2xy)^*$. It is well known that simple eigenvalues and their normalized eigenvectors vary smoothly with the related matrix; thus, for $T \in \mathcal{T}_\Lambda$, T near $T_\mathcal{X}$, the first column of Q can indeed be written in the form above for an appropriate choice of x and y, (x, y) near $p_\mathcal{X}$. Since $0 = T_{31} = \langle q_3, \Lambda q_1 \rangle$, we must have q_3 orthogonal both to w_1 and Λw_1 and therefore $q_3 \sim w_1 \times \Lambda w_1 \sim w_3 = (xy, -2y, 2)^*$. Similarly, $q_2 \sim w_2 = w_3 \times w_1 = (2x(1+2y^2), 2(2+x^2y^2), y(4-x^2))^*$. Thus, we search for an LU-positive orthogonal matrix $Q = EWN$ where $E \in \mathcal{E}$, W has columns w_i and N is a positive diagonal matrix. It is now easy to verify that $N = \mathrm{diag}(1/r_1, 1/(r_1r_2), 1/r_2)$ where r_1 and r_2 are given in the statement; also, by computing signs of pricipal minors, WN is LU-positive and therefore $E = I$. Expansion of the product $Q^*\Lambda Q$ obtains the formula for ϕ.

To show that ϕ is an immersion near $T_\mathcal{X}$, it suffices to verify that the Jacobian $D\phi$ at $p_\mathcal{X}$ is injective. This is easily seen by computing partial derivatives of the entries T_{11} and T_{23} of $T = \phi(x, y)$ with respect to x and y. ■

Actually, ϕ is a diffeomorphism from $\mathbb{R}^2$ to its image [8], but this will not be used in this text. A reason for choosing such a parametrization for the first column of Q will be clarified by the next proposition: an s-step admits a simple representation in (x, y)-coordinates.

Proposition 2.2 *There are open neighborhoods $S_2 \subset (-1, 1)$ of $s = 0$, $\mathcal{U}_2 \subset \mathcal{U}_1$ of $p = p_{\mathcal{X}}$ and $\boldsymbol{\mathcal{U}}_2 = \phi(\mathcal{U}_2) \subset \boldsymbol{\mathcal{U}}_1$ with the following properties.*

(a) The function $\mathbf{F} : (S_2 \setminus \{0\}) \times \boldsymbol{\mathcal{U}}_2 \to \boldsymbol{\mathcal{U}}_1$ is smooth and extends continuously (but not smoothly) to $S_2 \times \boldsymbol{\mathcal{U}}_2$.

(b) Let $F : S_2 \times \mathcal{U}_2 \to \mathcal{U}_1$ be $\mathbf{F}$ expressed in (x, y)-coordinates, i.e., $F(s, p) = \phi^{-1}(\mathbf{F}(s, \phi(p)))$. Then, for $p = (x, y) \in \mathcal{U}_2$,

$$F(s, x, y) = \left(\frac{1+s}{1-s} x, \frac{|s|}{1+s} y \right).$$

Proof: By definition, $\mathbf{F}(s, T) = (\mathbf{Q}(T - sI))^* T \mathbf{Q}(T - sI)$. For $T \approx T_{\mathcal{X}}$ (i.e., T near $T_{\mathcal{X}}$) and $s \approx 0$, $s \neq 0$, we have $T - sI \approx T_{\mathcal{X}}$. The first two columns of $\mathbf{Q}(T - sI)$ can be obtained from the corresponding columns of $T - sI$ by the Gram-Schmidt algorithm and therefore

$$\mathbf{Q}(T - sI) \approx \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -\text{sign}(\det(T - sI)) \end{pmatrix}.$$

Whatever the sign of $\det(T - sI)$, $\mathbf{F}(s, T) \approx T_{\mathcal{X}}$ so that

$$\lim_{s \to 0, s \neq 0, T \to T_{\mathcal{X}}} \mathbf{F}(s, T) = T_{\mathcal{X}}.$$

This proves that there exist open neighborhoods S_2 and $\boldsymbol{\mathcal{U}}_2$ such that $\mathbf{F} : (S_2 \setminus \{0\}) \times \boldsymbol{\mathcal{U}}_2 \to \boldsymbol{\mathcal{U}}_1$ is well defined and extends continuously to the point $(0, T_{\mathcal{X}})$ with $\mathbf{F}(0, T_{\mathcal{X}}) = T_{\mathcal{X}}$. The smoothness of $\mathbf{F}$ in $(S_2 \setminus \{0\}) \times \boldsymbol{\mathcal{U}}_2$ follows from the smoothness of $\mathbf{Q}$ in the set of invertible matrices. The fact that $\mathbf{F}$ extends continuously to $S_2 \times \boldsymbol{\mathcal{U}}_2$ will follow from the explicit formula in item (b).

As in the proof of Proposition 2.1, for $(s, T_0) \in S_2 \times \boldsymbol{\mathcal{U}}_2$, write $T_0 - sI = Q_0^*(\Lambda - sI)Q_0$, where Q_0 is LU-positive. Notice that if $(a_0, b_0, c_0)^*$ is the first column of Q_0 and $T_0 = \phi(x_0, y_0)$ then $x_0 = -2b_0/a_0$ and $y_0 = c_0/(2b_0)$. For $s \neq 0$, let $(x_1, y_1) = G(s, x_0, y_0)$ and $T_1 = \mathbf{F}(s, T_0) = \phi(x_1, y_1) = Q_1^* \Lambda Q_1$ where Q_1 is LU-positive. Let (a_1, b_1, c_1) be the first column of Q_1: we have $x_1 = -2b_1/a_1$ and $y_1 = c_1/(2b_1)$. We must therefore compute a_1, b_1, c_1.

By definition,

$$T_1 = (\mathbf{Q}(T_0 - sI))^* T_0 \mathbf{Q}(T_0 - sI) = (\mathbf{Q}(T_0 - sI))^* Q_0^* \Lambda Q_0 \mathbf{Q}(T_0 - sI),$$

so that $Q_1 = E Q_0 \mathbf{Q}(T_0 - sI)$ for some signature matrix $E = \text{diag}(e_1, e_2, e_3) \in \mathcal{E}$. Set $Q_1 = E\mathbf{Q}(Q_0(T_0 - sI)) = E\mathbf{Q}((\Lambda - sI)Q_0)$. Assuming $|s| < 1$, we have the positive collinearity $(a_1, b_1, c_1)^* \sim (e_1(1 - s)a_0, e_2(-1 - s)b_0, e_3(-s)c_0)^*$ and thus

$$x_1 = -\frac{e_2}{e_1} \frac{1+s}{1-s} x_0, \quad y_1 = \frac{e_3}{e_2} \frac{s}{1+s} y_0.$$

Now, $\text{sign}(x_i) = \text{sign}((T_i)_{12})$, $i = 0, 1$, and $\text{sign}((T_1)_{12}) = \text{sign}((T_0)_{12})$ (from Proposition 2.1 and equation (2)) and therefore $\text{sign}(x_1) = \text{sign}(x_0)$; similarly $\text{sign}(y_1) = \text{sign}(y_0)$, completing the proof. ■

The computations above fit into a larger context, which we now outline. As with s-steps, many eigenvalue algorithms act on $n \times n$ *Jacobi matrices* (as in [2], [7] and [12], real symmetric tridiagonal matrices T with $T_{i+1,i} > 0$, $i = 1, 2, \ldots, n-1$) with non-Jacobi limit points. Recall that Jacobi matrices have simple (real) eigenvalues λ_i and that its normalized eigenvectors v_i can be chosen so that the first coordinates c_i are positive; notice that $\sum c_i^2 = 1$. The eigenvalues λ_i and the *norming constants* c_i form a standard set of coordinates for Jacobi matrices. In these standard coordinates ([3]), the s-step acting on a Jacobi matrix T is given by

$$(\lambda_i, c_i), \quad \mapsto \quad (\lambda_i, \frac{|\lambda_i - s|c_i}{\sum_i (|\lambda_i - s|c_i)^2}), \quad i = 1, \ldots, n$$

provided $T - sI$ is invertible. Such coordinates require some modification in order to extend beyond the set of Jacobi matrices. The *bidiagonal coordinates* in [8] are, up to multiplicative constants, quotients c_{i+1}/c_i which admit a simpler evolution under s-steps. The map $\phi(x, y)$ retrieves a matrix from one among 3! possible choices of bidiagonal coordinates. In general, each permutation π of the eigenvalues obtains a map $\phi_\pi : \mathbb{R}^n \to \mathcal{T}_\Lambda$ and this family of maps is an atlas for $\mathcal{T}_\Lambda$.

3 Sub-eigenvalues

For a matrix $T \in \mathcal{T}$, let $\hat{T}$ be its bottom 2×2 diagonal block. Given $T \in \mathcal{T}_\Lambda$, the eigenvalues $\boldsymbol{\omega}_-(T) \leq \boldsymbol{\omega}_+(T)$ of $\hat{T}$ are the *sub-eigenvalues* of T. Notice that $\boldsymbol{\omega}_-(T_\mathcal{X}) = \boldsymbol{\omega}_+(T_\mathcal{X}) = 0$.

Consider the circles $\mathbf{c}_h, \mathbf{c}_v \subset \mathcal{T}_\Lambda$ through $T_\mathcal{X}$ parametrized by

$$\begin{pmatrix} \sin\theta & \cos\theta & 0 \\ \cos\theta & -\sin\theta & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & \cos\theta & 0 \\ \cos\theta & 0 & \sin\theta \\ 0 & \sin\theta & 0 \end{pmatrix},$$

respectively. Clearly, $T \in \mathbf{c}_h$ if and only if $T_{23} = T_{33} = 0$; also, $T \in \mathbf{c}_v$ if and only if $T_{11} = T_{22} = T_{33} = 0$.

Proposition 3.1 *There is an open connected neighborhood $\boldsymbol{\mathcal{U}}_3 \subset \boldsymbol{\mathcal{U}}_2$ of $T_\mathcal{X}$ with the following properties.*

(a) For all $T \in \boldsymbol{\mathcal{U}}_3$ the following interlacing inequalities hold:

$$-1 \leq \boldsymbol{\omega}_-(T) \leq 0 \leq \boldsymbol{\omega}_+(T) \leq 1, \quad \boldsymbol{\omega}_-(T) \leq T_{33} \leq \boldsymbol{\omega}_+(T).$$

(b) The only matrix $T \in \boldsymbol{\mathcal{U}}_3$ for which $\boldsymbol{\omega}_-(T) = \boldsymbol{\omega}_+(T)$ is $T = T_\mathcal{X}$. The functions $\boldsymbol{\omega}_\pm : \boldsymbol{\mathcal{U}}_3 \to \mathbb{R}$ are continuous and smooth on $\boldsymbol{\mathcal{U}}_3 \smallsetminus \{T_\mathcal{X}\}$.

(c) A matrix $T \in \boldsymbol{\mathcal{U}}_3$ belongs to $\mathbf{c}_h$ if and only if at least one sub-eigenvalue of T coincides with an eigenvalue of T. In particular, $T_{12} > 0$ for all $T \in \boldsymbol{\mathcal{U}}_3$.

(d) A matrix $T \in \boldsymbol{\mathcal{U}}_3$ belongs to $\mathbf{c}_v$ if and only if T_{33} is equidistant from the sub-eigenvalues of T.

(e) For all $T \in \boldsymbol{\mathcal{U}}_3$, $\boldsymbol{\omega}_\pm(T) \in S_2$.

The set S_2 mentioned in item (e) was defined in Proposition 2.2.

Proof: The first inequality is the interlacing of the eigenvalues of T and $\hat{T}$, the second is the interlacing of those of $\hat{T}$ and T_{33}.

From item (a), if the sub-eigenvalues are equal then $\boldsymbol{\omega}_-(T) = \boldsymbol{\omega}_+(T) = T_{33} = 0$. Thus, $\hat{T} = 0$ and since the trace of T equals 0, one has $T_{11} = 0$ and $T_{12} = \pm 1$: only

the positive choice, which gives rise to $T_{\mathcal{X}}$ itself, is relevant. Continuity in $\boldsymbol{\mathcal{U}}_3$ and smoothness in $\boldsymbol{\mathcal{U}}_3 \setminus \{T_{\mathcal{X}}\}$ of the functions $\boldsymbol{\omega}_\pm$ are now easy.

For item (c), suppose that $\det(T - \boldsymbol{\omega}_+(T)I) = 0$ (the case $\det(T - \boldsymbol{\omega}_-(T)I) = 0$ is similar). By construction, $\boldsymbol{\omega}_+(T)$ must be a common eigenvalue of T and $\hat{T}$. Expand the characteristic polynomial of T along the first row to obtain

$$\det(\lambda I - T) = (\lambda - T_{11}) \det(\lambda I - \hat{T}) + (-T_{12}^2\lambda + T_{12}^2 T_{33}).$$

A common eigenvalue annihilates the two determinants, thus $T_{12}^2(T_{33} - \boldsymbol{\omega}_+(T)) = 0$. Since $T_{12} \approx 1$ and $\boldsymbol{\omega}_+(T)$ equals an eigenvalue of Λ we have $\boldsymbol{\omega}_+(T) = T_{33} = 0$. Now $\det \hat{T} = 0$, which implies $T_{23} = 0$. Notice that $T_{12} = 0$, $T_{23} \neq 0$ implies that some sub-eigenvalue equals ± 1: this possibility was excluded above.

For item (d), if T_{33} is equidistant from the sub-eigenvalues, $T \in \mathcal{T}_\Lambda$ must be of the form

$$\begin{pmatrix} -2d & b & 0 \\ b & d & c \\ 0 & c & d \end{pmatrix}.$$

Now $\det(T - \lambda) = \lambda^3 - \lambda = \lambda^3 + (-3d^2 - b^2 - c^2)\lambda - 2dc^2 + db^2 + 2d^3$, so $d(b^2 - 2c^2 + 2d^2) = 0$ and $b^2 + c^2 + 3d^2 = 1$. If $d = 0$, T belongs to $\mathbf{c}_v$: notice that $T_{\mathcal{X}}$ corresponds to $\theta = 0$ in the parametrization of $\mathbf{c}_v$. If $d \neq 0$, T lies in one of two curves in $\mathcal{T}_\Lambda$ which may be assumed to be disjoint from $\boldsymbol{\mathcal{U}}_3$.

Continuity of the functions $\boldsymbol{\omega}_\pm$ allows for a choice of $\boldsymbol{\mathcal{U}}_3$ satisfying the final condition. ■

Let $\mathcal{U}_3 = \phi^{-1}(\boldsymbol{\mathcal{U}}_3)$. From item (c) above, $(x, y) \in \mathcal{U}_3$ implies $x > 0$. Let $r_h \subset \mathbb{R}^2$ be the horizontal axis and $r_v \subset \mathbb{R}^2$ be the vertical line $x = 2$. Let $\mathcal{U}_{3,+} = \mathcal{U}_3 \cap \{(x, y) \mid x \leq 2\}$, $\mathcal{U}_{3,-} = \mathcal{U}_3 \cap \{(x, y) \mid x \geq 2\}$, $\boldsymbol{\mathcal{U}}_{3,\pm} = \phi(\mathcal{U}_{3,\pm})$.

It is convenient to work with a domain for (x, y) coordinates which is more explicit than the set $\mathcal{U}_3$.

Definition 3.2 *Let $\mathcal{R}_a = [2 - a, 2 + a] \times [-a/10, a/10] \subset \mathcal{U}_3$ be a rectangle centered in $p_{\mathcal{X}} = (2, 0)$. The rectangle $\mathcal{R}_a$ is split by r_v in two closed rectangles $\mathcal{R}_{a,+} = \mathcal{R}_a \cap \mathcal{U}_{3,+}$ and $\mathcal{R}_{a,-} = \mathcal{R}_a \cap \mathcal{U}_{3,-}$.*

From [8], it follows easily that $\mathcal{U}_3$ can be taken to contain $\mathcal{R}_{1/10}$: using this result, as we shall see, all the subsequent constructions are compatible with $a = 1/10$. To make this paper self-contained the working hypothesis is only $a \leq 1/10$, $\mathcal{R}_a \subset \mathcal{U}_3$.

We rephrase Proposition 3.1 in (x, y)-variables. Write $\omega_\pm(x, y) = \boldsymbol{\omega}_\pm(\phi(x, y))$ so that the functions $\omega_\pm$ are continuous in $\mathcal{R}_a$ with a non-smooth point $p_{\mathcal{X}}$. Set $\boldsymbol{\mathcal{R}}_a = \phi(\mathcal{R}_a)$.

Proposition 3.3 *The diffeomorphism $\phi : \mathcal{R}_a \to \boldsymbol{\mathcal{R}}_a \subset \boldsymbol{\mathcal{U}}_3$ yields bijections from $r_h \cap \mathcal{R}_a$ to $\mathbf{c}_h \cap \boldsymbol{\mathcal{R}}_a$ and from $r_v \cap \mathcal{R}_a$ to $\mathbf{c}_v \cap \boldsymbol{\mathcal{R}}_a$. The functions $\omega_\pm(x, y)$ are even with respect to y, i.e, $\omega_\pm(x, y) = \omega_\pm(x, -y)$. For points $(x, y) \in \mathcal{R}_{a,+}$ (resp. $\mathcal{R}_{a,-}$), $(\phi(x, y))_{33}$ is to the right (resp. left) of $(\omega_+(x, y) + \omega_-(x, y))/2$.*

Signs in the notation $\mathcal{R}_{a,\pm}$ indicate which among $\omega_\pm$ is closer to $(\phi(x, y))_{33}$: unfortunately, they are the reverse of what their position might suggest.

Proof: We already saw in Proposition 2.1 that $\operatorname{sign}((\phi(x, y))_{23}) = \operatorname{sign}(y)$. Clearly, $\phi(2, (\tan\theta)/\sqrt{2})$ equals the matrix used to parametrize $\mathbf{c}_v$ in the statement of Proposition 3.1. Evenness of $\omega_\pm(x, y)$ in y is immediate from the explicit form of ϕ in Proposition 2.1. Finally, to decide which sub-eigenvalue of $T = \phi(x, y)$ is closer to T_{33}, compare $\operatorname{tr} \hat{T} = T_{22} + T_{33} = \omega_+(x, y) + \omega_-(x, y)$ with $2T_{33}$:

$$(2T_{33} - \operatorname{tr} T) r_1^2 r_2^2 = (x - 2)(x + 2)(x^2 y^2 + 8x^2 y^4 - 4 + 4y^2).$$

For x slightly smaller (resp. larger) than 2, the expression is positive (resp. negative) and thus T_{33} is to the right (resp. left) of $(\omega_+(x,y)+\omega_-(x,y))/2$. Since the only points where $T_{33} = (\omega_+(x,y)+\omega_-(x,y))/2$ are those in $r_v \cap \mathcal{R}_a$ and $\mathcal{R}_a$ is connected the result follows. ■

We need more precise estimates for the sub-eigenvalues $\omega_\pm$ near $p_\mathcal{X} = (2,0)$.

Definition 3.4 *The wedge* $\mathcal{X}_0$ *is* $\{(x,y) \in \mathcal{R}_a \mid |y| \geq |x-2|/10\}$*; set* $\mathcal{X}_{0,\pm} = \mathcal{X}_0 \cap \mathcal{R}_{a,\pm}$.

Figure 1 contains some of the geometric objects defined in this section; the triangles $\mathcal{D}_{0,\pm}$ will be defined in the next section.

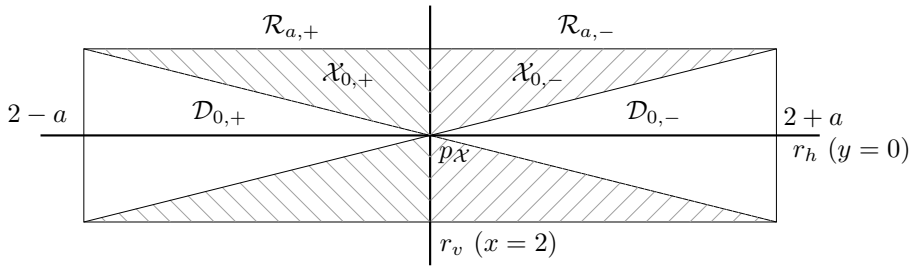

Figure 1: The rectangle $\mathcal{R}_a$; in $\mathcal{R}_{a,+}$, $\omega = \omega_+$ and in $\mathcal{R}_{a,-}$, $\omega = \omega_-$.

Lemma 3.5 *Near the point* $p_\mathcal{X} = (2,0)$ *the functions* $\omega_\pm$ *have a cone-like behavior:*

$$\omega_\pm = \frac{(x-2) \pm \sqrt{(x-2)^2 + 32y^2}}{4} + O(d^2)$$

where $d^2 = (x-2)^2 + y^2$*. In* $\mathcal{R}_{a,\pm}$*,* $y^2 \leq |\omega_\pm(x,y)| \leq 2|y|$*; in* $\mathcal{X}_{0,\pm}$*,* $|y|/5 \leq |\omega_\pm(x,y)| \leq 2|y|$.

Proof: As in Proposition 2.1, set $r_1^2 = 4 + x^2 + 4x^2y^2$. The sub-eigenvalues solve $r_1^2\omega^2 + (4-x^2)\omega - 4x^2y^2 = 0$,

$$\omega_\pm = \frac{-4 + x^2 \pm \sqrt{\Delta}}{2r_1^2}, \tag{3}$$

where $\Delta = \Delta(x,y) = ((x+2)^2 + 8x^2y^2)((x-2)^2 + 8x^2y^2) \geq 0$, $\Delta = O(d^2)$. In particular, $\Delta = 0$ in $\mathcal{R}_a$ only for $p_\mathcal{X} = (2,0)$. The expression for $\omega_\pm$ in the statement of the lemma follows directly from

$$\Delta - 16((x-2)^2 + 32y^2) = O(d^3).$$

For $y > 0$, the inequality $\omega_+ \leq 2|y|$ is equivalent to $\Delta \leq \left(4r_1^2 y + 4 - x^2\right)^2$ which is in turn equivalent to

$$8r_1^2 y(4 - x^2 + 8y + 8x^2y^3) \geq 0,$$

which is clearly true for $0 < x \leq 2$. The case $y < 0$ follows from the fact that ω_+ is even in y. A similar factorization holds for ω_-. Also, we have $\omega_+(x,y)\omega_-(x,y) = -4x^2y^2/r_1^2$ and $|\omega_-(x,y)|, |\omega_+(x,y)| \leq 1$. Since $a \leq 1/10$ we have that $4x^2/r_1^2 \leq 1$ for $(x,y) \in \mathcal{R}_a$. The estimate for $\omega_\pm$ in $\mathcal{X}_{0,\pm}$ is now somewhat cumbersome but straightforward. ■

We also need estimates for partial derivatives of $\omega_\pm$ in terms of x and y.

Lemma 3.6 *For all $(x,y) \in \mathcal{R}_a - \{p_\mathcal{X}\}$ the partial derivatives $(\omega_\pm)_x$ and $(\omega_\pm)_y$ satisfy $0 \le (\omega_\pm)_x < 1$ and $|(\omega_\pm)_y| < 7/3$. The equality $(\omega_+)_x = 0$ (resp. $(\omega_-)_x = 0$) holds exactly for $y = 0$, $x < 2$ (resp. $x > 2$); for $y \neq 0$, $\pm y(\omega_\pm)_y > 0$.*

Proof: The partial derivatives of $\omega_\pm$ are

$$(\omega_\pm)_x = \frac{8x}{r_1^4\sqrt{\Delta}}\left(\left((1+2y^2)\sqrt{\Delta}\right) \pm \left(-4+x^2+8y^2+6x^2y^2+16x^2y^4\right)\right),$$
$$(\omega_\pm)_y = \frac{4x^2y}{r_1^4\sqrt{\Delta}}\left(\left((4-x^2)\sqrt{\Delta}\right) \pm \left(16+24x^2+x^4+32x^2y^2+8x^4y^2\right)\right),$$

which are well defined provided $\Delta \neq 0$, i.e., outside of $p_\mathcal{X} = (2,0)$. Also,

$$\left((1+2y^2)\sqrt{\Delta}\right)^2 - \left(-4+x^2+8y^2+6x^2y^2+16x^2y^4\right)^2 = 8y^2r_1^4 \ge 0$$

whence $(1+2y^2)\sqrt{\Delta} \ge \left|-4+x^2+8y^2+6x^2y^2+16x^2y^4\right|$ and $(\omega_\pm)_x \ge 0$; equality implies $y = 0$. Also, $(\omega_\pm)_x \le 16x(1+2y^2)/r_1^4 < 1$.

In the rectangle $\mathcal{R}_a$,

$$115 < 16+24x^2+x^4+32x^2y^2+8x^4y^2 < 142, \quad \left|(4-x^2)\sqrt{\Delta}\right| < 1/4$$

and the signs of $(\omega_\pm)_y$ are settled. Since

$$\frac{y^2}{\Delta} \le \frac{1}{(x+2)^2+8x^2y^2}\,\frac{y^2}{8x^2y^2} < \frac{1}{360}$$

we also have

$$|(\omega_\pm)_y| \le \frac{|y|}{\sqrt{\Delta}}\,\frac{4x^2\cdot\left(16+24x^2+x^4+32x^2y^2+8x^4y^2+\left|(4-x^2)\sqrt{\Delta}\right|\right)}{r_1^4} < \frac{7}{3}.$$

■

4 Wilkinson's step

Take $\boldsymbol{\omega}(T)$ to be the sub-eigenvalue of T closer to T_{33}; in case of a draw, we arbitrarily choose $\boldsymbol{\omega}(T) = \boldsymbol{\omega}_+(T)$. *Wilkinson's step* is the map $\mathbf{W}(T) = \mathbf{F}(\boldsymbol{\omega}(T), T)$. From now on we shall work in (x,y) coordinates, i.e., with $W(x,y) = \phi^{-1}(\mathbf{W}(\phi(x,y))) = F(\omega(x,y),x,y)$ where $\omega(x,y) = \boldsymbol{\omega}(\phi(x,y))$. Write $W_\pm(x,y) = F(\omega_\pm(x,y),x,y)$. From Propositions 2.2 and 3.1, the maps $W_\pm : \mathcal{R}_a \to \mathcal{U}_1$ are continuous and $W : \mathcal{R}_a \to \mathcal{U}_1$ is well defined with step discontinuities along $r_v \cap \mathcal{R}_a$. From Proposition 2.2 and the fact that $\omega_-(x,y) \le 0 \le \omega_+(x,y)$,

$$W_\pm(x,y) = (X_\pm(x,y), Y_\pm(x,y)) = \left(\frac{1+\omega_\pm(x,y)}{1-\omega_\pm(x,y)}\,x, \frac{\pm\omega_\pm(x,y)}{1+\omega_\pm(x,y)}\,y\right) \qquad (4)$$

and therefore $W_\pm$ are smooth functions in $\mathcal{R}_a \smallsetminus \{p_\mathcal{X}\}$. Evenness of $\omega_\pm$ with respect to y (Proposition 3.3) yields $X_\pm(x,y) = X_\pm(x,-y)$, $Y_\pm(x,y) = -Y_\pm(x,-y)$.

The rectangles $\mathcal{R}_a = [2-a, 2+a] \times [-a/10, a/10]$ for $a \le 1/10$ are invariant under W; furthermore, the maps $W_\pm$ are injective on $\mathcal{R}_{a,\pm}$. Figure 2 provides strong evidence to these facts, proved in Propositions 4.1 and 4.3.

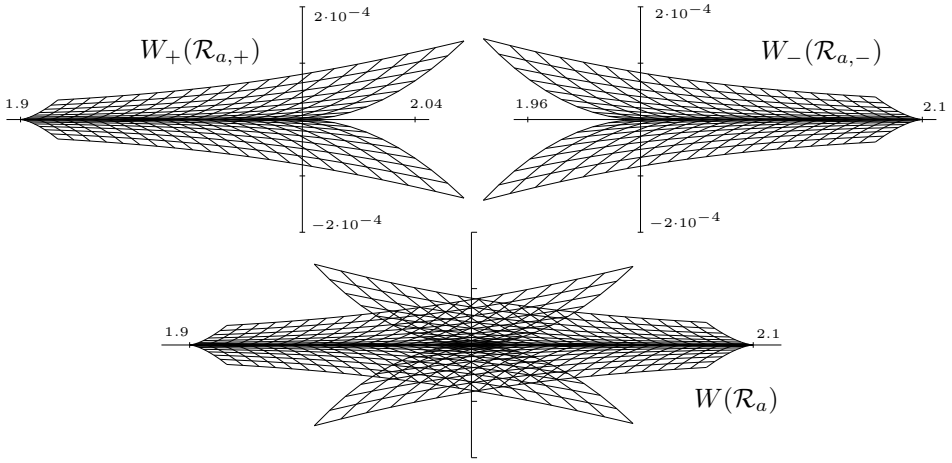

Figure 2: $W_+(\mathcal{R}_{a,+})$, $W_-(\mathcal{R}_{a,-})$ and $W(\mathcal{R}_a)$, $a = 1/10$; vertical scales are stretched.

Proposition 4.1 *Take $a \leq 1/10$ satisfying also $\mathcal{R}_a \subseteq \mathcal{U}_3$. Then $W(\mathcal{R}_a) \subseteq \mathcal{R}_a$. Moreover, $|Y(x,y)| \leq |y|/49$ for all $(x,y) \in \mathcal{R}_a$, $X_+(x,y) \geq x$ for all $(x,y) \in \mathcal{R}_{a,+}$ and $X_-(x,y) \leq x$ for all $(x,y) \in \mathcal{R}_{a,-}$. Also, $W_\pm(\{2 \mp a\} \times [-a/10, a/10]) \subset \mathcal{R}_{a,\pm}$, $W_\pm(r_v) \subset \mathcal{R}_{a,\mp}$.*

Proof: From Lemma 3.5, we have $|\omega_\pm(x,y)| \leq 2a/10 \leq 1/50$ and therefore $|Y(x,y)| \leq |y|/49$. We now prove that $W_+(\mathcal{R}_{a,+}) \subset \mathcal{R}_a$. Clearly, $(x,y) \in \mathcal{R}_{a,+}$ implies $|Y(x,y)| \leq a/10$ so we must prove that $2 - a \leq X_+(x,y) \leq 2 + a$. From equation (4), $X_+(x,y) \geq x$ with equality exactly when $y = 0$. Since $(\omega_+)_x(x,y) \geq 0$ we have $X_x(x,y) \geq 0$ and it therefore suffices to prove the two claims in the statement, i.e., $X_+(2-a,y) < 2$ and $X_+(2,y) < 2 + a$. For $a \leq 1/10$, the inequalities follow from $\omega_+(x,y) \leq 2a/10 < a/(4+a) < a/(4-a)$. Similar checks apply to W_-. ■

Each rectangle $\mathcal{R}_{a,\pm}$ is not invariant. Denote the interior of a set $X \subset \mathbb{R}^2$ by $\mathrm{int}(X)$. Given $b \geq 0$, $b < a$, let $\mathcal{D}_{b,+}$ (resp. $\mathcal{D}_{b,-}$) be the triangle defined by $2 - a \leq x \leq 2 - b - 10|y|$ (resp. $2 + b + 10|y| \leq x \leq 2 + a$). Notice that $\mathrm{int}(\mathcal{X}_0) = \mathcal{R}_a \setminus (\mathcal{D}_{0,+} \cup \mathcal{D}_{0,-})$. Let $\mathcal{Y}_0 \subset \mathcal{X}_0$ be the thinner open wedge defined by $|y| > 10|x - 2|$.

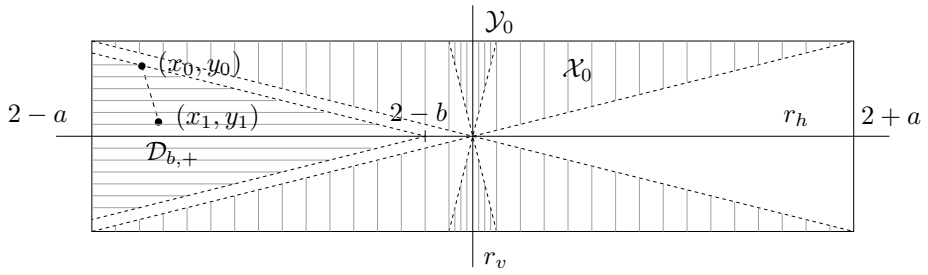

Figure 3: $(x_0, y_0) \in \mathcal{D}_{b,+}$ implies $(x_1, y_1) = W_+(x_0, y_0) \in \mathcal{D}_{b,+}$.

Proposition 4.2 *The triangles $\mathcal{D}_{b,\pm}$ are invariant, i.e., $W_+(\mathcal{D}_{b,+}) \subseteq \mathcal{D}_{b,+}$ and $W_-(\mathcal{D}_{b,-}) \subseteq \mathcal{D}_{b,-}$. If $p \neq (2 \pm b, 0), (2 \pm a, 0)$ is a boundary point of $\mathcal{D}_{b,\pm}$ then $W_\pm(p) \in \mathrm{int}(\mathcal{D}_{b,\pm})$.*

Finally, $W_+(\mathcal{Y}_0) \subset \mathrm{int}(\mathcal{D}_{0,-})$ *and* $W_-(\mathcal{Y}_0) \subset \mathrm{int}(\mathcal{D}_{0,+})$.

This will imply (Proposition 5.1) that points in $\mathcal{D}_{0,\pm}$ and $\mathcal{Y}_0$ have cubic convergence to points in r_h different from $p_{\mathcal{X}}$.

Proof: We prove that $W_+(\mathcal{D}_{b,+}) \subseteq \mathcal{D}_{b,+}$ by computing the slope α of the segment joining (x_0, y_0) and

$$(x_1, y_1) = W_+(x_0, y_0) = \left(\frac{1 + \omega_+(x_0, y_0)}{1 - \omega_+(x_0, y_0)}\, x_0, \frac{\omega_+(x_0, y_0)}{1 + \omega_+(x_0, y_0)}\, y_0 \right),$$

given by

$$\alpha = \frac{-y_0}{\omega_+} \frac{1 - \omega_+}{2x_0(1 + \omega_+)}$$

where ω_+ stands for $\omega_+(x_0, y_0)$ (see Figure 3). By Lemma 3.5, $|\omega_\pm| \leq 2|y|$: simple algebra then shows that, for $a \leq 1/10$, we have $|\alpha| > 1/10$ and the segment is steeper than the non-vertical sides of $\mathcal{D}_{b,+}$. Since y_0 and y_1 have the same sign, invariance of $\mathcal{D}_{b,+}$ follows. The argument for $\mathcal{D}_{b,-}$ is similar.

The verification that $W_\pm(\mathcal{Y}_0) \subset \mathrm{int}(\mathcal{D}_{0,\mp})$ uses estimates of the form $|y|/2 < |\omega_\pm(p)| < 1/20$ in the closure of $\mathcal{Y}_0$; details are left to the reader. ■

Proposition 4.3 *Each map* $W_\pm : \mathrm{int}(\mathcal{R}_{a,\pm}) \setminus r_h \to \mathcal{R}_a$ *is an orientation preserving diffeomorphism to its image.*

Proof: We consider the Jacobian matrix

$$DW_\pm(x, y) = \begin{pmatrix} \dfrac{2(\omega_\pm)_x}{(1 - (\omega_\pm))^2} x + \dfrac{1 + (\omega_\pm)}{1 - (\omega_\pm)} & \dfrac{2(\omega_\pm)_y}{(1 - (\omega_\pm))^2} x \\ \\ \pm \dfrac{(\omega_\pm)_x}{(1 + (\omega_\pm))^2} y & \pm \dfrac{(\omega_\pm)_y}{(1 + (\omega_\pm))^2} y \pm \dfrac{(\omega_\pm)}{1 + (\omega_\pm)} \end{pmatrix}. \quad (5)$$

From Lemma 3.6, $(X_\pm)_x > 0$ and $(Y_\pm)_y \geq 0$ with equality precisely for $y = 0$. Similarly, for $y \neq 0$, $\mathrm{sign}((X_\pm)_y) = \mathrm{sign}((Y_\pm)_x) = \pm\mathrm{sign}(y)$.

To prove that $W_\pm$ are local diffeomorphisms in $\mathrm{int}(\mathcal{R}_{a,\pm}) \setminus r_h$, write

$$\det DW_\pm(x, y) = \pm \frac{1}{1 - \omega_\pm^2} \left(\frac{2(\omega_\pm)_x \omega_\pm x}{1 - \omega_\pm} + (\omega_\pm)_y y + \omega_\pm(1 + \omega_\pm) \right).$$

From Lemma 3.6, all terms in the sum in parenthesis have the same sign and thus $\det DW_\pm(x, y) > 0$ if $y \neq 0$.

We now prove the injectivity of W_+ on $\mathrm{int}(\mathcal{R}_{a,+}) \setminus r_h$. From symmetry, it suffices to prove the injectivity of W_+ on $\mathcal{R}_{a,++} = \{(x, y) \in \mathrm{int}(\mathcal{R}_{a,+}) \mid y \geq 0\}$. In other words, given $(x_1, y_1) \in \mathcal{R}_a$, $y_1 > 0$, we must prove that there exists at most one point $(x, y) \in \mathcal{R}_{a,++}$ with $W_+(x, y) = (x_1, y_1)$. Let $\gamma : [0, 1] \to \mathbb{R}^2$ be the piecewise affine counterclockwise parametrization of the boundary of $\mathcal{R}_{a,++}$ with $\gamma(0) = \gamma(1) = (2 - a, 0)$, $\gamma(1/4) = (2, 0)$, $\gamma(1/2) = (2, a/10)$ and $\gamma(3/4) = (2 - a, a/10)$. Since $\det DW_+(x, y) > 0$ for $y > 0$ and $y_1 > 0$, (x_1, y_1) is a regular value of W_+ and, assuming $(x_1, y_1) \notin (W_+ \circ \gamma)([0, 1])$, the number of solutions of $W_+(x, y) = (x_1, y_1)$ is given by the winding number c_1 of $W_+ \circ \gamma$ around (x_1, y_1). Recall that a simple way to compute $2c_1$ is to count with signs the intersections of $W_+ \circ \gamma$ with the vertical line through (x_1, y_1). From the signs of the entries of DW_+, the x coordinate of $(W_+ \circ \gamma)(t)$ is strictly increasing from $t = 0$ to $t = 1/2$ and strictly decreasing from $t = 1/2$ to $t = 1$. Thus, there are at most two intersection points and $|c_1| \leq 1$ implying injectivity. If $(x_1, y_1) \in (W_+ \circ \gamma)([0, 1])$, the argument above applies to nearby points and the result follows by continuation outside r_h. The proof of the analogous statement for W_- is similar. ■

Actually, the maps $W_\pm : \mathcal{R}_{a,\pm} \to \mathcal{R}_a$ are homeomorphisms to their respective images; we omit details.

We conclude the section with some estimates on $DW_\pm$ which will be used in the last section. A vector $v = (v_1, v_2)$ is *near-horizontal* if $v_1 > 0$ and $|v_2|/v_1 < 1/25$.

Lemma 4.4 *Take $p \in \mathcal{R}_a$, $p \neq p_\mathcal{X}$. Let $v = (v_1, v_2)$ be a near-horizontal vector. Then $\tilde{v} = DW_\pm(p)v = (\tilde{v}_1, \tilde{v}_2)$ is also near-horizontal and $0 < v_1/2 < \tilde{v}_1 < 10v_1$.*

Proof: From the expressions of the entries of

$$DW_\pm = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$$

in equation (5) above and the estimates in Lemma 3.5 and Lemma 3.6,

$$0.96 < m_{11} < 5.5, \quad |m_{12}| < 10.5, \quad |m_{21}| < 0.0105, \quad |m_{22}| < 0.045.$$

The claims now follow from easy computations. ■

5 Convergence rates of Wilkinson's shift strategy

We now consider the asymptotic behavior of Wilkinson's shift strategy. Given a **W**-orbit (T_k), $T_{k+1} = \mathbf{W}(T_k)$, we study the decay of the entry $(T_k)_{23}$. In (x, y) coordinates, the W-orbits (p_k) are defined by $p_{k+1} = W(p_k)$, $p_0 \in \mathcal{R}_a$; the relevant issue is the convergence rate of the y coordinate since, from Proposition 2.1, $(\phi(x, y))_{32}/y = 2r_1/r_2^2 > 0$ is bounded and bounded away from 0 in $\mathcal{R}_a$.

A W-orbit (p_k) has *strictly quadratic (resp. cubic) convergence* if there exist constants $c, C > 0$ such that, for all $k \in \mathbb{N}$,

$$c|y_k|^r \leq |y_{k+1}| \leq C|y_k|^r$$

for $r = 2$ (resp. $r = 3$) (here $p_k = (x_k, y_k)$).

Proposition 5.1 *Let $p \in \mathcal{R}_a \setminus r_h$, $p_k = W^k(p)$. If $p_k \in \mathcal{D}_{0,\pm}$ for some k then the W-orbit (p_k) has strictly cubic convergence. Otherwise $p_k \in \mathcal{X}_0$ for all k and convergence is strictly quadratic.*

Proof: The case $y_k = 0$ is trivial. Assume without loss that $p_k \in \mathcal{D}_{0,+}$. From Proposition 4.2, there exists $b > 0$ such that $p_{k+1} \in \mathcal{D}_{b,+}$ and therefore $p_j \in \mathcal{D}_{b,+}$ for all $j > k$. From Proposition 3.1, ω_+ is an even, smooth function in the y-variable in $\mathcal{D}_{b,+}$. From compactness and Lemma 3.5, given $b > 0$ there exists $C > 0$ such that for all $(x, y) \in \mathcal{D}_{b,+}$ we have $y^2 \leq \omega_+(x, y) \leq Cy^2$. From equation (4),

$$|y|^3/2 \leq |Y_+(x, y)| = \frac{\omega_+}{1 + \omega_+}|y| \leq 2C|y|^3$$

for $(x, y) \in \mathcal{D}_{b,+}$, yielding strictly cubic convergence.

We now consider orbits in the wedge $\mathcal{X}_0 = \mathcal{X}_{0,+} \cup \mathcal{X}_{0,-}$. From Lemma 3.5, $|y|/5 \leq |\omega_\pm(x, y)| \leq 2|y|$ for $(x, y) \in \mathcal{X}_{0,\pm}$: strictly quadratic convergence now follows from

$$|y|^2/10 \leq |Y_+(x, y)| = \frac{\omega(x, y)}{2 + \omega(x, y)}|y| \leq 4|y|^2,$$

completing the proof. ■

Notice that the constant C in the proof depends on b and therefore the rate of cubic convergence is not uniform in $\mathcal{D}_{0,+}$, consistent with strictly quadratic convergence for orbits in $\mathcal{X}_0$.

Corollary 5.2 *Given $p \in \mathcal{R}_a \setminus r_h$, consider the W-orbit $p_k = W^k(p)$. The following conditions are equivalent:*

(a) $\lim_{k\to\infty} p_k = p_{\mathcal{X}}$*;*

(b) $p_k \in \mathcal{X}_0$ *for all* k*;*

(c) the W-orbit (p_k) has strictly quadratic convergence.

Proof: The estimate $|Y(x,y)| \le |y|/50$ guarantees convergence to some point p_∞ in r_h. Orbits contained in $\mathcal{X}_0$ must then converge to $p_{\mathcal{X}}$. Conversely, if $p_\infty \neq p_{\mathcal{X}}$ then $p_\infty \in \text{int}(\mathcal{D}_{b,\pm})$ for some $b > 0$ and $p_k \in \mathcal{D}_{b,\pm}$ for sufficiently large k. ■

Let $\mathcal{X} \subset \mathcal{R}_a$ be the set of points p for which $\lim p_k = p_{\mathcal{X}}$. From Proposition 4.2, $\mathcal{X} \subset \mathcal{X}_0 \setminus \mathcal{Y}_0$ and therefore $r_v \cap \mathcal{X} = r_h \cap \mathcal{X} = \{p_{\mathcal{X}}\}$. We still need to prove that $\mathcal{X} \neq \{p_{\mathcal{X}}\}$: this and more will be done in this section.

Figure 4 shows $\mathcal{X}$ extended to a rectangle much larger than $\mathcal{R}_{1/10}$: numerical evidence indicates that even in such larger regions the qualitative descriptions remain valid.

A compact set $K \subset \mathbb{R}$ is a *Cantor set* if K has empty interior and no isolated points. As we shall prove in Theorem 6.1, horizontal sections of $\mathcal{X}$ are Cantor sets. The set $\mathcal{X}$ is the union of *uncountably many* arcs, disjoint except at $p_{\mathcal{X}}$. Each arc intersects a horizontal line in a single point.

The Hausdorff dimension of the middle-third Cantor is $\log 2/\log 3 \approx 0.63$ ([5] and [6] contain a thorough discussion of Hausdorff dimension). More generally, self-similar Cantor sets have positive Hausdorff dimension. The horizontal sections of $\mathcal{X}$ are much thinner: they have Hausdorff dimension 0. That is why the fine structure is invisible in this figure, unlike most figures of Cantor sets in the quoted books.

Numerical evidence also indicates that the northwest-southeast leg of set $\mathcal{X}$ is the union of a family of analytic curves, tangent (not crossing) at $p_{\mathcal{X}}$: we shall not pursue this matter further.

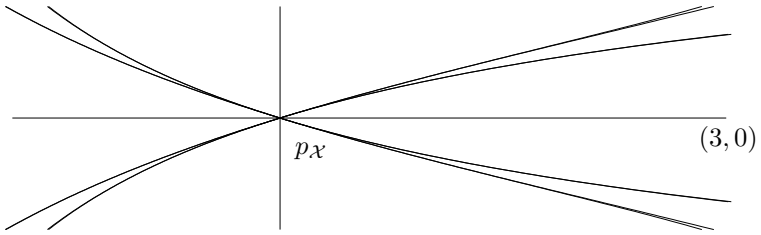

Figure 4: The set $\mathcal{X}$ near $p_{\mathcal{X}}$; in scale.

We provide another description of $\mathcal{X}$ as the intersection of a nested sequence of compact sets $\mathcal{X}_n$. From Corollary 5.2, $p = p_0 \in \mathcal{X}$ if and only if $p_k = W^k(p) \in \mathcal{X}_0$ for all k. Define $\mathcal{X}_n \subset \mathcal{R}_a$ to be the set of points $p \in \mathcal{R}_a$ such that $p_k \in \mathcal{X}_0$ for $0 \le k \le n$. Thus $\mathcal{X}_0 = \mathcal{X}_0$, $\mathcal{X}_{n+1} = W^{-1}(\mathcal{X}_n) \subset \mathcal{X}_n$ and $\mathcal{X} = \bigcap_n \mathcal{X}_n$. Figure 5 indicates the first few sets. As the diagram suggests, the interior of $\mathcal{X}_n$ has 2^{n+1} connected components which we now describe. In a sense, the Cantor sets on horizontal lines are limits of this successive doubling of components.

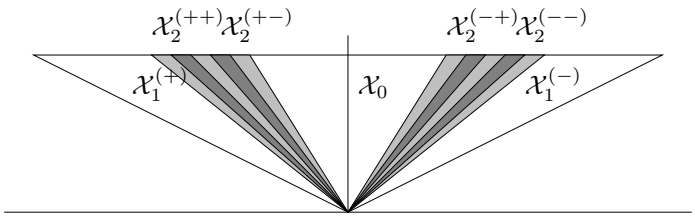

Figure 5: The upper halves of the sets $\mathcal{X}_0$, $\mathcal{X}_1^{\pm}$ and $\mathcal{X}_2^{\pm\pm}$; schematic.

Decompose the sets $\mathcal{X}_n$ by tracking on which side of r_v the points p_k lie. A *sign sequence* of length n is a function $\tau : \{0, \dots, n-1\} \to \{+, -\}$ or, equivalently, a string of n signs $(\tau(0), \tau(1), \dots, \tau(n-1))$. Define

$$\mathcal{X}_n^\tau = \{p \in \mathcal{R}_a \mid p_k \in \mathcal{X}_{0,\tau(k)}, 0 \le k < n; p_n \in \mathcal{X}_0\}.$$

Figure 6 shows a simple example: the upper half of $\mathcal{X}_1^{(+)}$ is a curvilinear triangle with base contained in the top side of $\mathcal{R}_{a,+}$ and vertex in $p_\mathcal{X}$; the upper half of $\mathcal{X}_1^{(-)}$ is a similar triangle in $\mathcal{R}_{a,-}$. Also, $\mathcal{X}_1^{(+)} = W_+^{-1}(\mathcal{X}_0) \cap \mathcal{R}_{a,+} \subset \mathcal{X}_0$.

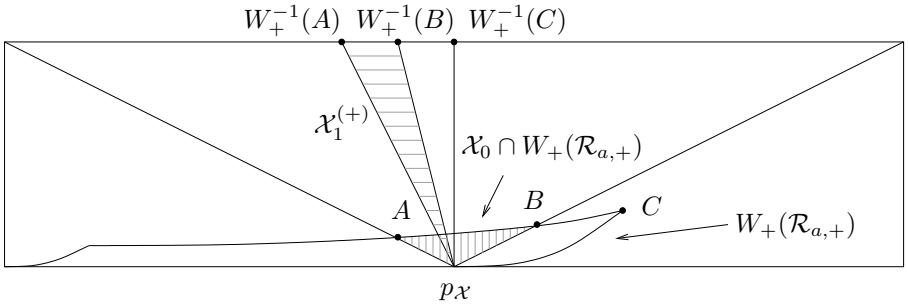

Figure 6: The upper halves of $\mathcal{X}_0 \cap W_+(\mathcal{R}_{a,+})$ and $\mathcal{X}_1^{(+)} = W_+^{-1}(\mathcal{X}_0)$; schematic.

We also consider *infinite sign sequences*, i.e., functions $\sigma : \mathbb{N} \to \{+, -\}$, and the corresponding sets

$$\mathcal{X}^\sigma = \bigcap_{n \in \mathbb{N}} \mathcal{X}_n^{\sigma|_{\{0,1,\dots,n-1\}}}.$$

Let Σ be the set of infinite sign sequences: Σ admits the natural metric $d(\sigma_1, \sigma_2) = 2^{-n}$ where n is the smallest number for which $\sigma_1(n) \neq \sigma_2(n)$. The metric space Σ is homeomorphic to the middle-third Cantor set contained in $[0, 1]$. The sets $\mathcal{X}^\sigma$ are compact, the intersection of two distinct such sets is $\{p_\mathcal{X}\}$ and the union of all $\mathcal{X}^\sigma$ is $\mathcal{X}$. We conclude this section by proving that the intersection of each $\mathcal{X}^\sigma$ with a horizontal line is not empty.

Let $\ell_\pm$ be the sides $y = \pm a^2$, $-a \le x \le a$, of $\mathcal{R}_a$.

Proposition 5.3 *Let $\gamma_h : [0, 1] \to \mathcal{R}_a \setminus \{p_\mathcal{X}\}$ be a parametrized curve with $\gamma_h(0) \in \mathcal{D}_{0,+}$, $\gamma_h(1) \in \mathcal{D}_{0,-}$. Let σ be an infinite sign sequence: there exists $t \in [0, 1]$ such that $\gamma_h(t) \in \mathcal{X}^\sigma$.*

Notice that we do not claim that such t is unique: this requires stronger hypothesis and will be discussed in the next section.

Proof: We first prove by induction on n that the connected component containing $p_\mathcal{X}$ of each set $\mathcal{X}_n^\tau$ has elements on the sides $\ell_\pm$ (here τ is a sign sequence of length

n). The case $n = 0$ is trivial and the case $n = 1$ has already been discussed (Figure 6). Assume the connected component of $\mathcal{X}_n^\tau$ containing $p_\mathcal{X}$ to have a point in ℓ_+. Consider a path $\gamma_v : [0,1] \to \mathcal{X}_n^\tau$ with $\gamma_v(0) = p_\mathcal{X}$, $\gamma_v(1) \in \ell_+$. The image of γ_v must intersect the curve $W_+(\ell_+)$: let t_0 be the smallest t in this intersection. The image of the restriction of γ_v to $[0, t_0]$ is contained in $W_+(\mathcal{R}_{a,+})$. Since W_+ is a homeomorphism on its image (Proposition 4.3), there exists $\gamma_{v+} : [0, t_0] \to \mathcal{R}_{a,+}$ with $W_+(\gamma_{v+}(t)) = \gamma_v(t)$ for all t. Also, the image of γ_{v+} is contained in $\mathcal{X}_{n+1}^{(+,\tau)}$. A similar construction works for ℓ_- and for $\mathcal{X}_{n+1}^{(-,\tau)}$, completing the proof of the claim.

Consider an infinite sign sequence σ and its restrictions $\tau_n = \sigma|_{\{0,1,\ldots,n-1\}}$. For each n let

$$K_n = \{t \in [0,1] \mid \gamma_h(t) \in \mathcal{X}_n^{\tau_n}\}.$$

The sets K_n are nested, compact and nonempty and therefore their intersection is also nonempty. Any t in the intersection satisfies $\gamma_h(t) \in \mathcal{X}^\sigma$. ■

As we shall see in the last section, the sets $\mathcal{X}_n^\tau$ and $\mathcal{X}^\sigma$ are connected but the proof of this fact requires more careful estimates.

6 Geometry of $\mathcal{X}$

The main result of this section, Theorem 6.1, is that $\mathcal{X}$ is very thin, almost as thin as a finite union of curves. More precisely, $\mathcal{X}$ has Hausdorff dimension 1.

Denote the length of an interval I by $\mu(I)$. Given an infinite sign sequence σ, define $\sigma^\sharp$ by $\sigma^\sharp(k) = \sigma(k+1)$ (this is the standard shift operator in symbolic dynamics). Similarly, for a sign sequence τ of length $n+1$ let $\tau^\sharp$ be the sign sequence of length n defined by $\tau^\sharp(k) = \tau(k+1)$.

Theorem 6.1 *For any infinite sign sequence σ, the set $\mathcal{X}^\sigma$ is a curve $\mathcal{X}^\sigma = \{(g^\sigma(y), y), y \in [-a/10, a/10]\}$ where $g^\sigma : [-a/10, a/10] \to [2-a, 2+a]$ is a Lipschitz function. Wilkinson's step takes curves to curves: $W(\mathcal{X}^\sigma) \subset \mathcal{X}^{\sigma^\sharp}$. The set $\mathcal{X}$ has Hausdorff dimension 1 and the intersection of $\mathcal{X}$ with any horizontal line is a Cantor set of Hausdorff dimension 0.*

We need a few preliminary definitions. A *near-horizontal curve* is a C^1 function $\gamma : I \to \mathcal{R}_a \setminus \{p_\mathcal{X}\}$ such that, for all t in the interval I, the tangent vector $\gamma'(t)$ is near-horizontal (as defined at the end of Section 4). A near-horizontal curve $\tilde{\gamma}$ is *standard* if the first coordinate of $\tilde{\gamma}(t)$ is $x = t$. For any near-horizontal curve γ there exists a unique strictly increasing C^1 function α, the *standard reparametrization* of γ, for which $\tilde{\gamma} = \gamma \circ \alpha$ is standard. The *height* of a near-horizontal curve γ is $y_* = \tilde{\gamma}(2)$ so that γ crosses r_v at $(2, y_*)$.

Near-horizontality is preserved by Wilkinson's step: from Lemma 4.4, if $\gamma : I \to \mathcal{R}_{a,+} \setminus \{p_\mathcal{X}\}$ (resp. $\mathcal{R}_{a,-} \setminus \{p_\mathcal{X}\}$) is a near-horizontal curve then so is $W_+ \circ \gamma$ (resp. $W_- \circ \gamma$). This process squeezes near-horizontal curves towards the line r_h. The constant $1/25$ in the definition is somewhat arbitrary but it can not be replaced by very small numbers since W does not take horizontal lines to horizontal lines.

Recall that $\mathcal{Y}_0 \subset \mathcal{X}_0$ is an open wedge with the property that $p \in \mathcal{Y}_0$ implies $W_\pm(p) \in \mathcal{D}_{0,\mp}$. For a sign sequence τ of length n, define

$$\mathcal{Y}_n^\tau = \{p \in \mathcal{R}_a \mid p_k \in \mathcal{X}_{0,\tau(k)}, 0 \le k < n; p_n \in \mathcal{Y}_0\}.$$

In particular, the orbit (p_k) with $p_0 \in \mathcal{Y}_n^\tau$ escapes $\mathcal{X}_0$ starting from $k = n+1$. Also, $\mathcal{Y}_n^\tau$ is an open subset of $\mathcal{X}_n^\tau$ disjoint from $\mathcal{X}_{n+1}$.

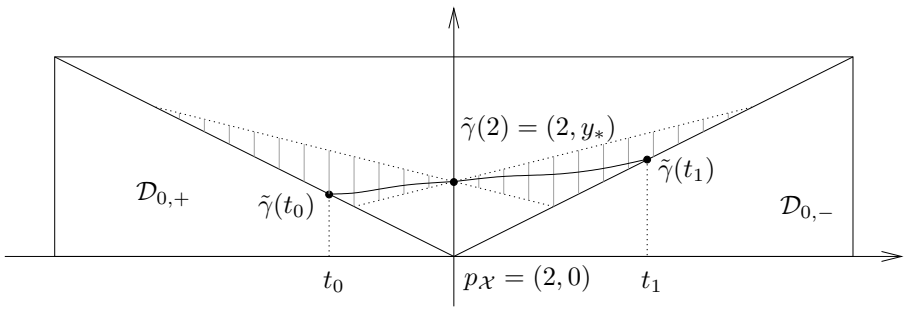

Figure 7: A near-horizontal curve (not in scale).

Lemma 6.2 *Let $\tilde\gamma : [t_0, t_1] \to \mathcal{R}_a \setminus \{p_{\mathcal{X}}\}$ be a standard near-horizontal curve. Assume that $\tilde\gamma(t_0) \in \mathcal{D}_{0,+}$, $\tilde\gamma(t_1) \in \mathcal{D}_{0,-}$. Let $y_* = \tilde\gamma(2)$ be the height of $\tilde\gamma$. Then, for $\tilde\gamma(t) = (t, y) \in \mathcal{X}_0$,*

$$|y_*|/3 < |y| < 3|y_*|.$$

Let τ be a sign sequence of length n. The sets

$$I_n^\tau = \{t \in [t_0, t_1] \mid \tilde\gamma(t) \in \mathcal{X}_n^\tau\}, \quad J_n^\tau = \{t \in [t_0, t_1] \mid \tilde\gamma(t) \in \mathcal{Y}_n^\tau\}$$

are intervals and their lengths satisfy

$$\frac{1}{10^{n+1}} \left|\frac{y_*}{90}\right|^{2^n} < \mu(J_n^\tau) < \mu(I_n^\tau) < 2^n \left|40 y_*\right|^{2^n} < 2^{(-2^{(n-1)})}.$$

Proof: As a basis for an inductive proof, we first consider the case $n = 0$. Set

$$I_0 = \{t \in [t_0, t_1] \mid \tilde\gamma(t) \in \mathcal{X}_0\}, \quad J_0 = \{t \in [t_0, t_1] \mid \tilde\gamma(t) \in \mathcal{Y}_0\}.$$

Draw lines through the point $(2, y_*)$ with linear coefficients $\pm 1/25$, as in Figure 7, and compute their intersections with the diagonals of $\mathcal{R}_a$. Since these diagonals are steeper than $\tilde\gamma$, the intersection of the image of $\tilde\gamma$ with $\mathcal{X}_0$ is contained in the shaded triangles. An elementary geometric argument proves the first claim, verifies that I_0 and J_0 are intervals and obtains the estimates for their lengths.

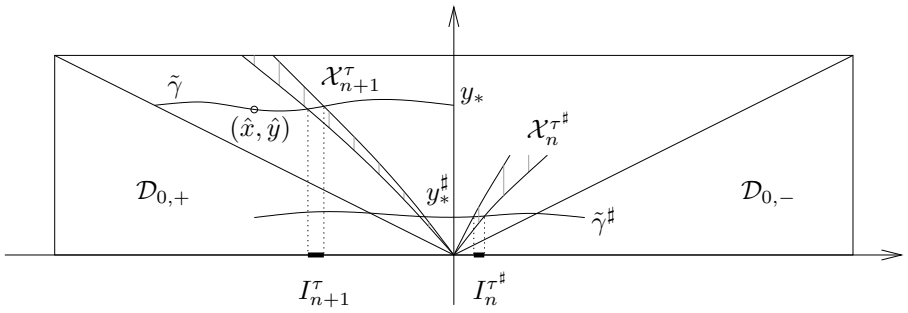

Figure 8: A near-horizontal curve (not in scale).

We now do the induction step. For a sign sequence τ of length $n + 1$ with $\tau(0) = +$ (the other case is similar), consider

$$I_{n+1}^\tau = \{t \in [t_0, t_1] \mid \tilde\gamma(t) \in \mathcal{X}_{n+1}^\tau\}$$

and γ_+, the restriction of $\tilde{\gamma}$ to $[t_0, 2]$. Set $\gamma^\sharp = W_+ \circ \gamma_+$ (see Figure 8): as remarked above, $\gamma^\sharp$ is a near-horizontal curve. Let $\alpha^\sharp : [t_0^\sharp, t_1^\sharp] \to [t_0, 2]$ be the standard reparametrization of $\gamma^\sharp$ so that $\tilde{\gamma}^\sharp = \gamma^\sharp \circ \alpha^\sharp$ is a standard near-horizontal curve with height $y_*^\sharp = \tilde{\gamma}^\sharp(2)$. Notice that $\tilde{\gamma}^\sharp(t_0^\sharp) \in \mathcal{D}_{0,+}$ and $\tilde{\gamma}^\sharp(t_1^\sharp) \in \mathcal{D}_{0,-}$ so, by the induction hypothesis, the sets

$$I_n^{\tau^\sharp} = \{t \in [t_0^\sharp, t_1^\sharp] \mid \tilde{\gamma}^\sharp(t) \in \mathcal{X}_n^{\tau^\sharp}\}, \quad J_n^{\tau^\sharp} = \{t \in [t_0^\sharp, t_1^\sharp] \mid \tilde{\gamma}^\sharp(t) \in \mathcal{Y}_n^{\tau^\sharp}\}$$

are intervals whose lengths $\ell_I^\sharp, \ell_J^\sharp$ satisfy

$$\frac{1}{10^{n+1}} \left|\frac{y_*^\sharp}{90}\right|^{2^n} < \ell_J^\sharp < \ell_I^\sharp < 2^n \left|40 y_*^\sharp\right|^{2^n}. \tag{6}$$

Clearly, $t \in I_n^{\tau^\sharp}$ (resp. $J_n^{\tau^\sharp}$) if and only if $\alpha^\sharp(t) \in I_{n+1}^\tau$ (resp. J_{n+1}^τ), proving that I_{n+1}^τ and J_{n+1}^τ are intervals; let ℓ_I, ℓ_J be their lengths. By Lemma 4.4, $1/10 < (\alpha^\sharp)'(t) < 2$ for all $t \in I_n$ and therefore

$$(1/10)\ell_J^\sharp < \ell_J < \ell_I < 2\ell_I^\sharp. \tag{7}$$

Let $\tilde{\gamma}(\alpha^\sharp(2)) = (\hat{x}, \hat{y})$ so that $W_+(\hat{x}, \hat{y}) = (2, y_*^\sharp)$. From the first claim, $|y_*|/3 < |\hat{y}| < 3|y_*|$. By Lemma 3.5, $(1/5)|\hat{y}| < \omega_+(\hat{x}, \hat{y}) < 2|\hat{y}|$. Thus, by equation 4, $(1/10)\hat{y}^2 < |y_*^\sharp| < 3\hat{y}^2$. Combining these estimates,

$$\left|\frac{y_*}{90}\right| > \left|\frac{y_*^\sharp}{90}\right|^2, \quad \left|40 y_*^\sharp\right| < |40 y_*|^2. \tag{8}$$

The required estimate now follows from estimates 6, 7, 8 and the fact that $|y_*| \leq a/10 \leq 1/100$. ■

Proof of Theorem 6.1: We organize the proof by stating a few claims, most of which in terms of a standard near-horizontal curve $\tilde{\gamma}$; notation will be borrowed from Lemma 6.2. For a sign sequence τ of length n, consider the sign sequences $(\tau, +)$ and $(\tau, -)$ of length $n+1$.

(a) *The intervals $I_{n+1}^{(\tau,+)}$ and $I_{n+1}^{(\tau,-)}$ are contained in different connected components of $I_n^\tau \smallsetminus J_n^\tau$.*

This claim is proved by induction. The case $n = 0$ follows from the inclusions $\mathcal{X}_1^\pm \subset \mathcal{X}_{0,\pm} \smallsetminus \mathcal{Y}_0$. The induction step employs the same construction used in Lemma 6.2: for $\tilde{\gamma}^\sharp = W^+ \circ \tilde{\gamma} \circ \alpha^\sharp$, the standard reparametrization $\alpha^\sharp$ maps the intervals constructed from $\tau^\sharp$ to the corresponding intervals for τ.

(b) *For an infinite sign sequence σ there is a unique $t^\sigma \in [t_0, t_1]$ with $\tilde{\gamma}(t^\sigma) \in \mathcal{X}^\sigma$.*

This follows directly from Lemma 6.2 since

$$\lim_{n\to\infty} \mu(I_n^{\sigma|_{\{0,1,\ldots,n-1\}}}) = 0.$$

In particular, a horizontal line at height y (which is near-horizontal) meets $\mathcal{X}^\sigma$ at a single point $(g^\sigma(y), y)$. The Lipschitz estimate $|g^\sigma(y_1) - g^\sigma(y_2)| \leq 25|y_1 - y_2|$ likewise follows by taking lines of slope less than $1/25$.

(c) *Let σ_1 and σ_2 be distinct sign sequences and n be the smallest integer for which $\sigma_1(n) \neq \sigma_2(n)$. Then*

$$\frac{1}{10^{n+1}} \left|\frac{y_*}{90}\right|^{2^n} < |t^{\sigma_1} - t^{\sigma_2}| < 2^n \left|40 y_*\right|^{2^n} < 2^{(-2^{(n-1)})}.$$

Let τ be the restriction of σ_1 or σ_2 to $\{0, 1, \ldots, n-1\}$. Without loss, $t^{\sigma_1} \in I_{n+1}^{(\tau,+)}$ and $t^{\sigma_2} \in I_{n+1}^{(\tau,-)}$: Lemma 6.2 and claim (a) obtain the inequalities.

In particular, the map from the Cantor set Σ to $\mathcal{K} = \{t \in [t_0, t_1] \mid \tilde{\gamma}(t) \in \mathcal{X}\}$ taking σ to t_σ is a homeomorphism.

We now construct a very thin open cover of $\mathcal{K}$. For a positive integer n let $S_n^+ \subset \Sigma$ be the set of the 2^n sign sequences σ for which $k \geq n$ implies $\sigma(k) = +$. Let $B_r(p)$ be the ball of radius r and center p.

(d) *Take $r > 0$ and let n be a positive integer such that $r > 2^{(-2^{(n-1)})}$. Then the balls $B_r(t^{\sigma^+}) \subset \mathbb{R}$, $\sigma^+ \in S_n^+$, form an open cover of $\mathcal{K}$.*

From step (b), given $t \in \mathcal{K}$ there is some $\sigma \in \Sigma$ for which $t = t^\sigma$. Let $\sigma_n \in S_n^+$ be defined by

$$\sigma_n(k) = \begin{cases} \sigma(k), & 0 \leq k < n, \\ +, & k \geq n \end{cases}$$

so that $\sigma_n \in S_n^+$. From claim (c), $|t^{\sigma_n} - t^\sigma| < 2^{(-2^{(n-1)})}$ and therefore $t \in B_r(t^{\sigma_n})$.

The open cover of $\mathcal{K}$ above is now adapted to yield a thin open cover of $\mathcal{X}$. For a positive integer m, let $Y_m \subset [-a/10, a/10]$ be a set of $2m$ equally spaced points so that given $y \in [-a/10, a/10]$ there is $y_0 \in Y_m$ with $|y - y_m| < a/(10m)$. Also, for positive integers n and m, let $X_{n,m} \subset \mathcal{X}$ be the set of $m2^{(n+1)}$ elements of the form $(g^\sigma(y), y)$, $\sigma \in S_n^+$, $y \in Y_m$.

(e) *Consider $r > 0$ and positive integers n and m such that $r > 2^{(1-2^{(n-1)})}$ and $r > 1/m$. Then the balls $B_r(p) \subset \mathbb{R}^2$, $p \in X_{n,m}$, form an open cover of $\mathcal{X}$.*

Given $p = (g^\sigma(y), y)$ let y_0 be the point of Y_m closest to y and let $\sigma_n \in S_n^+$ be defined as in claim (d). Let $p_1 = (g^{\sigma_n}(y), y)$ and $p_0 = (g^{\sigma_n}(y_0), y_0)$. Notice that $p_0 \in X_{n,m}$. In order to prove that p belongs to $B_r(p_0)$ we estimate $|p - p_1|$ and $|p_1 - p_0|$. From claim (d), $|p - p_1| < r/2$. Since $a \leq 1/10$, $|y - y_0| < a/(10m) \leq r/100$; from the Lipschitz estimate after claim (b), $|g^{\sigma_0}(y) - g^{\sigma_0}(y_0)| \leq r/4$ and therefore $|p_1 - p_0| < r/2$.

(f) *The set $\mathcal{K}$ has Hausdorff dimension* 0.

Given $r > 0$, let $n = \lceil 1 + \log_2(-\log_2 r) \rceil$. Claim (d) obtains on open cover of $\mathcal{K}$ by $2^n < -4 \log_2 r$ balls of radius r. For all $d > 0$ we have

$$\lim_{r \to 0} (-4 \log_2 r) r^d = 0.$$

Thus, $\mu_d(\mathcal{K})$, the Hausdorff measure of dimension d of $\mathcal{K}$, equals 0.

(g) *The set $\mathcal{X}$ has Hausdorff dimension* 1.

Since $\mathcal{X}$ contains curves this dimension is at least one. It suffices therefore to prove that $\mu_d(\mathcal{X}) = 0$ for all $d > 1$.

Given $r > 0$, let $n = \lceil 2 + \log_2(-\log_2 r) \rceil$ and $m = \lceil 1/r \rceil$. Claim (e) obtains an open cover of $\mathcal{X}$ by $m2^{n+1} < -32(\log_2 r) r^{-1}$ balls of radius r. For all $d > 1$ we have

$$\lim_{r \to 0} -32(\log_2 r) r^{-1} r^d = 0.$$

This implies $\mu_d(\mathcal{X}) = 0$, proving the claim and completing the proof of the theorem. ■

7 Conclusion

Recall that $\boldsymbol{\omega}_+(T)$ and $\boldsymbol{\omega}-(T)$ are the roots of a quadratic polynomial with smooth coefficients (in T). The algorithm to obtain $\boldsymbol{\omega}$ given T is thus multivalued: in variables x and y, the domains of choice of each sign are rectangles sharing the line

r_v. In a sense, dynamical systems techniques arose naturally in this paper because orbits may switch sides arbitrarily often and according to arbitrary patterns (sign sequences). An outstanding example of the use of sign sequences to label orbits are kneading sequences ([11]). Many numerical algorithms do not exhibit this kind of complexity.

Unsurprisingly, the set $\mathcal{X}$ of atypical points is more complicated than an algebraic surface. This is why Hausdorff dimension becomes the relevant tool to quantify the size of $\mathcal{X}$. Theorem 6.1 shows that $\mathcal{X}$ is very thin.

A matrix is an *AP-matrix* if its spectrum contains an arithmetic progression with three terms and is *AP-free* otherwise. This paper concentrated on the simplest examples of AP-matrices, with eigenvalues -1, 0 and 1. We conjecture that the presence of three term arithmetic progressions in the spectrum (and whether the three eigenvalues are consecutive) dictates the possible rates of convergence of Wilkinson's shift. In a forthcoming paper we show that for AP-free tridiagonal matrices, cubic convergence of Wilkinson's shift indeed holds ([9], [10]).

Acknowledgements: The editorial board of FoCM did an extraodinary job in articulating a diverse collection of referees which led to a far more readable text. We are very grateful to these editors and referees. The authors acknowledge support from CNPq, CAPES, IM-AGIMB and Faperj.

References

[1] Batterson, S. and Smillie, J., Rayleigh quotient iteration for nonsymmetric matrices, Math. of Comp., 55, 169-178, 1990.

[2] de Boor, C. and Golub, G. H., The numerically stable reconstruction of a Jacobi matrix from spectral data, Lin. Alg. Appl. 21, 245-260, 1978.

[3] Deift, P., Nanda, T., Tomei, C., Differential equations for the symmetric eigenvalue problem, SIAM J. Num. Anal. 20, 1-22, 1983.

[4] Demmel, J. W., Applied Numerical Linear Algebra, SIAM, Philadelphia, 1997.

[5] Falconer, K. J., Fractal geometry: mathematical foundations and applications, John Willey, Chichester, 2003.

[6] Federer, H., Geometric measure theory, Classics in Mathematics, Springer-Verlag, New York, 1996.

[7] Gragg, W. B. and Harrod, W. J., The numerically stable reconstruction of Jacobi matrices from spectral data, Numer. Math. 44, 317-335, 1984.

[8] Leite, R. S., Saldanha, N.C. and Tomei, C., An atlas for tridiagonal isospectral manifolds, Linear Algebra Appl. 429, 387-402, 2008.

[9] Leite, R. S., Saldanha, N.C. and Tomei, C., The asymptotic of Wilkinson's shift: cubic convergence for generic spectra, in preparation.

[10] Leite, R. S., Saldanha, N.C. and Tomei, C., The asymptotics of Wilkinson's shift iteration, preprint, `http://www.arxiv.org/abs/math.NA/0412493`.

[11] Milnor, J. and Thurston, W., On iterated maps of the interval, in Dynamical Systems (ed. Alexander, J. C.), Lecture Notes in Math. 1342, 465-563, New York: Springer-Verlag, 1988.

[12] Moser, J., Finitely many mass points on the line under the influence of an exponential potential, in Dynamic systems theory and applications (ed. J. Moser), Lecture Notes in Phys. 38, 467-497, New York, 1975.

[13] Parlett, B. N., The Symmetric Eigenvalue Problem, Prentice-Hall, Englewood Cliffs, NJ 1980.

[14] Tomei, C., The Topology of Manifolds of Isospectral Tridiagonal Matrices, Duke Math. J., 51, 981-996, 1984.

[15] Wilkinson, J. H., The algebraic eigenvalue problem, Oxford University Press, 1965.

Ricardo S. Leite, Departamento de Matemática, UFES
Av. Fernando Ferrari, 514, Vitória, ES 29075-910, Brazil

Nicolau C. Saldanha and Carlos Tomei, Departamento de Matemática, PUC-Rio
R. Marquês de S. Vicente 225, Rio de Janeiro, RJ 22453-900, Brazil

rsleite@pq.cnpq.br
nicolau@mat.puc-rio.br; http://www.mat.puc-rio.br/∼nicolau/
tomei@mat.puc-rio.br