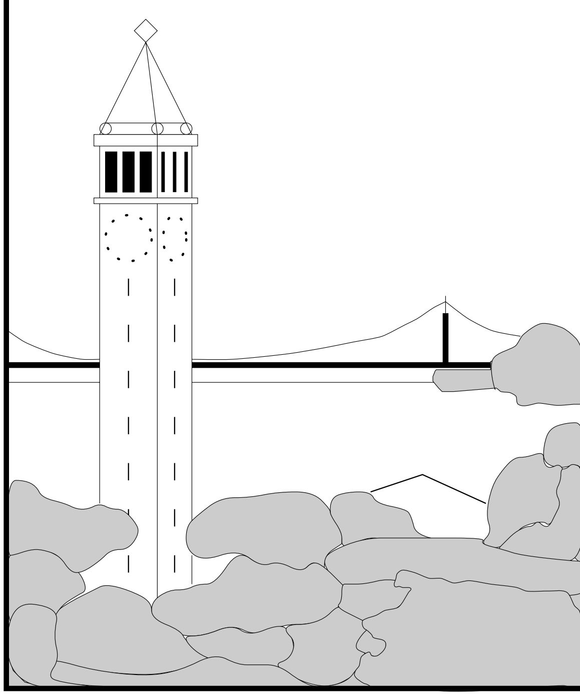
Relations Between the Field of Values of a Matrix and Those of its Schur Complements

Ren-Cang Li
Department of Mathematics
University of California at Berkeley
Berkeley, California 94720
li@math.berkeley.edu



Report No. UCB//CSD-94-849

December 1994

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Relations Between the Field of Values of a Matrix and Those of its Schur Complements *

Ren-Cang Li
Department of Mathematics
University of California at Berkeley
Berkeley, California 94720

June 14, 1992

Computer Science Division Technical Report UCB//CSD-94-849, University of California, Berkeley, CA 94720, December, 1994.

Abstract

Relations between the field of values of a matrix A and those of its Schur complements are established. This work began with an attempt to get rid of pivoting from Gauss elimination under certain circumstances when the field of values $\mathcal{F}(A)$ does not contain the origin. The upper bound proved in this paper must be improved before it is of more practical use. However, the proof of the upper bound does provide an intuition on how a tight upper bound looks like.

*This material is based in part upon work supported by Argonne National Laboratory under grant No. 20552402 and the University of Tennessee through the Advanced Research Projects Agency under contract No. DAAL03-91-C-0047, by the National Science Foundation under grant No. ASC-9005933, and by the National Science Infrastructure grants No. CDA-8722788 and CDA-9401156.

1 Kahan's Theorems

Let A be an $n \times n$ complex matrix. The *field of values* of A is defined to be the set

$$\mathcal{F}(A) \stackrel{\text{def}}{=} \{x^* A x / x^* x : 0 \neq x \text{ is a } n\text{-dimensional vector}\}.$$

(Here the superscript * means taking conjugate transpose.) Toeplitz–Hausdorff Theorem says that $\mathcal{F}(A)$ *is convex*. Now partition A as

$$A = \begin{pmatrix} H & R \\ L & V \end{pmatrix}, \tag{1}$$

where H is of $m \times m$ ($1 \leq m \leq n-1$). If H is invertible,

$$X \stackrel{\text{def}}{=} V - LH^{-1}R \tag{2}$$

is called the *Schur complement* of H in A. The following theorem establishes a relation between $\mathcal{F}(A)$ and $\mathcal{F}(X)$.

Theorem 1 *Let $\alpha \in \mathcal{F}(X)$. Then there exist a $\beta \in \mathcal{F}(A)$ and a positive number γ with $1 \leq \gamma \leq 1 + \|LH^{-1}\|_2^2$, where $\|\cdot\|_2$ is the spectral norm of a matrix, such that $\alpha = \gamma\beta$.*

Proof: Note

$$\begin{aligned} &\begin{pmatrix} I & 0 \\ -LH^{-1} & I \end{pmatrix} \begin{pmatrix} H & R \\ L & V \end{pmatrix} \begin{pmatrix} I & -H^{-*}L^* \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} H & R \\ 0 & V - LH^{-1}R \end{pmatrix} \begin{pmatrix} I & -H^{-*}L^* \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} H & R - HH^{-*}L^* \\ 0 & X \end{pmatrix} \stackrel{\text{def}}{=} C. \end{aligned} \tag{3}$$

Denote

$$M = \begin{pmatrix} I & 0 \\ -LH^{-1} & I \end{pmatrix}.$$

Then $C = MAM^*$ by (3). For any n–dimensional vector y, letting $x = M^*y$, we get then

$$\frac{y^*Cy}{y^*y} = \frac{x^*Ax}{x^*x} \cdot \frac{x^*x}{y^*y}. \tag{4}$$

Consider now those vector y having form

$$y = \begin{pmatrix} 0 \\ z \end{pmatrix}, \tag{5}$$

where z is of $n - m$–dimension. Then

$$y^*Cy = z^*Xz, \quad x = \begin{pmatrix} -H^{-*}L^*z \\ z \end{pmatrix}. \tag{6}$$

Hence $\|x\|_2^2 = \|z\|_2^2 + \|LH^{-1}z\|_2^2 \leq (1 + \|LH^{-1}\|_2^2)\|z\|_2^2$. Since $\|y\|_2 = \|z\|_2$,

$$1 \leq \frac{x^*x}{y^*y} \leq 1 + \|LH^{-1}\|_2^2.$$

By the definition of the field of values of a matrix, the conclusion of the theorem follows from (4). ∎

It has been proved that the field of values of any 2×2 matrix is an *ellipse*. More precisely, let $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and let U be the 2×2 unitary matrix such that

$$U^*TU = \begin{pmatrix} \lambda_1 & \delta \\ 0 & \lambda_2 \end{pmatrix},$$

where λ_1 and λ_2 are the eigenvalues of T, δ is nonnegative and equals to $(|a|^2 + |b|^2 + |c|^2 + |d|^2 - |\lambda_1|^2 - |\lambda_2|^2)^{1/2}$. The field of values $\mathcal{F}(T)$ is the ellipse with two foci λ_1 and λ_2 and semi-minor $|\delta|/2$. (It is a straight line segment joining λ_1 and λ_2 if $\delta = 0$, i.e., T is a normal matrix.) On the other hand, an ellipse on the complex plane corresponding to infinitely many 2×2 matrix unless the ellipse degenerates to a point.

There is only one Schur complement in a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ which is $d - \frac{bc}{a}$ provided $a \neq 0$.

For any vector $x \neq 0$, partition it conformly to (1) as $x = (x_1^T, x_2^T)^T$. (Here the superscript ^{T} means taking transpose.) Note

$$x^*Ax = x_1^*Hx_1 + x_1^*Rx_2 + x_2^*Lx_1 + x_2^*Vx_2.$$

We claim that there are complex numbers a, b, c, d and ξ_j with $\|x_j\|_2 = |\xi_j|$ such that

$$\begin{aligned} x_1^*Hx_1 &= a|\xi_1|^2, \\ x_1^*Rx_2 &= b\bar{\xi}_1\xi_2, \\ x_2^*Lx_1 &= c\xi_1\bar{\xi}_2, \\ x_2^*Vx_2 &= d|\xi_2|^2. \end{aligned}$$

As a matter of fact, we could simply take $\xi_j = \|x_j\|_2$, then solve for a, b, c, d in the following way:

- *The case* $x_1 \neq 0$ and $x_2 \neq 0$: Then $\xi_1 \neq 0$ and $\xi_2 \neq 0$. Hence
$$a = \frac{x_1^*Hx_1}{|\xi_1|^2},\ b = \frac{x_1^*Rx_2}{\bar{\xi}_1\xi_2},\ c = \frac{x_2^*Lx_1}{\xi_1\bar{\xi}_2},\ d = \frac{x_2^*Vx_2}{|\xi_2|^2}.$$

- *The case* $x_1 = 0$ and $x_2 \neq 0$: Then $\xi_1 = 0$ and $\xi_2 \neq 0$. Set $a = b = c = 0$ and $d = x_2^*Vx_2/|\xi_2|^2$.

- *The case* $x_1 \neq 0$ and $x_2 = 0$: Then $\xi_1 \neq 0$ and $\xi_2 = 0$. Set $a = x_1^*Hx_1/|\xi_1|^2$ while set $b = c = d = 0$.

Claim: *The field of values of the* 2×2 *matrix* $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, *which is an ellipse, is contained in* $\mathcal{F}(A)$.

Proof: The case when $x_1 = 0$ or $x_2 = 0$ is trivial. In the following we consider the case when none of the two is zero. For any $g = (\zeta_1, \zeta_2)^T$, let $\rho_j = \zeta_j/\xi_j$. It is easy to verify

$$\frac{g^*Tg}{g^*g} = \frac{\begin{pmatrix} \rho_1 x_1 \\ \rho_2 x_2 \end{pmatrix}^* A \begin{pmatrix} \rho_1 x_1 \\ \rho_2 x_2 \end{pmatrix}}{|\rho_1|^2\|x_1\|_2^2 + |\rho_2|^2\|x_2\|_2^2} \in \mathcal{F}(A).$$

$\blacksquare$

Recall the equations (4), (5) and (6). Consider now $x_2 = z$ (with $z^*z = 1$) and $x_1 = -H^{-*}L^*z$. For the present case $x_1^*Hx_1 + x_2^*Lx_1 = 0$. Thus $a|\xi_1|^2 + c\xi_1\bar{\xi}_2 = 0$. Assume, for the moment, $\xi_1 \neq 0$. Then $\bar{\xi}_1 = -\frac{c}{a}\bar{\xi}_2$, provided $a \neq 0$. By (4), we see ($z^*z = y^*y = 1 \Rightarrow |\xi_2| = 1$)

$$z^*Xz = x^*Ax = d - \frac{bc}{a},$$

if $a \neq 0$, which is guaranteed if we assume $0 \notin \mathcal{F}(H)$. If, however, $\xi = 0$, then by the construction of the 2×2 matrix T, $x_1 = 0$ which implies $L^*z = 0$. Thus $z^*Xz = x^*Ax = x_2^*Vx_2 = d$, which can be regarded as the Schur complement in a matrix like $\begin{pmatrix} d & 0 \\ 0 & d \end{pmatrix}$. Thus we have proved

Theorem 2 *If $0 \notin \mathcal{F}(H)$, then any point in $\mathcal{F}(X)$ is a Schur complement in a 2×2 matrix whose field of values is contained completely in $\mathcal{F}(A)$.*

2 The Region of Schur Complements of All 2×2 Matrices with a Fixed Field of Values

Lemma 1 *Let T be a 2×2 matrix, and D a 2×2 diagonal unitary matrix. Then T and DTD^* have the same Schur complement.*

Given an ellipse $\mathcal{E}(\lambda_1, \lambda_2, m)$ with the two foci λ_1 and λ_2 and semi-minor m, all possible 2×2 matrices whose fields of values are $\mathcal{E}(\lambda_1, \lambda_2, m)$ are

$$U \begin{pmatrix} \lambda_1 & \delta \\ 0 & \lambda_2 \end{pmatrix} U^*,$$

where U runs over all 2×2 unitary matrices, and $\delta = 2m$. Since any 2×2 unitary matrix U can be decomposed as

$$U = \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} c & -s \\ \bar{s} & \bar{c} \end{pmatrix},$$

where $|d_1| = |d_2| = 1$ and $|c|^2 + |s|^2 = 1$. By Lemma 1, to study Schur complements of all possible 2×2 matrices with the field of values $\mathcal{E}(\lambda_1, \lambda_2, m)$, it suffices for us to consider these matrices

$$\begin{aligned} & \begin{pmatrix} c & -s \\ \bar{s} & \bar{c} \end{pmatrix} \begin{pmatrix} \lambda_1 & \delta \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \bar{c} & s \\ -\bar{s} & c \end{pmatrix} \\ = & \begin{pmatrix} |c|^2\lambda_1 + |s|^2\lambda_2 - c\bar{s}\delta & cs(\lambda_1 - \lambda_2) + c^2\delta \\ \bar{c}\bar{s}(\lambda_1 - \lambda_2) - \bar{s}^2\delta & |s|^2\lambda_1 + |c|^2\lambda_2 + c\bar{s}\delta \end{pmatrix}, \end{aligned} \tag{7}$$

whose only Schur complement is

$$\frac{\lambda_1\lambda_2}{|c|^2\lambda_1 + |s|^2\lambda_2 - c\bar{s}\delta}. \tag{8}$$

Lemma 2

$$\begin{aligned} \mathcal{E}(\lambda_1, \lambda_2, m) &= \{|c|^2\lambda_1 + |s|^2\lambda_2 - c\bar{s}\delta \,:\, |c|^2 + |s|^2 = 1\} \\ &= \{|s|^2\lambda_1 + |c|^2\lambda_2 - c\bar{s}\delta \,:\, |c|^2 + |s|^2 = 1\}. \end{aligned}$$

From now on, for convenience, we will not distinguish a complex number and the unique point it represents on the complex plane. For a set $\mathcal{D}$ consisting of complex numbers, the notation $|\mathcal{D}|$ is defined as

$$|\mathcal{D}| = \max\{|z| \,:\, z \in \mathcal{D}\}.$$

We assume $\mathcal{E}(\lambda_1, \lambda_2, m)$ does not contain the origin, i.e.,

$$\mathcal{E}(\lambda_1, \lambda_2, m) \not\ni 0. \tag{9}$$

Our interest is the region $\mathcal{R}(\lambda_1, \lambda_2, m)$ of all possible values of (8) for all possible c and s subject to $|c|^2 + |s|^2 = 1$. By Lemma 2, we see

$$\mathcal{R}(\lambda_1, \lambda_2, m) = \{\lambda_1\lambda_2/z \,:\, z \in \mathcal{E}(\lambda_1, \lambda_2, m)\}. \tag{10}$$

Now, we try to find an upper bound for $|\mathcal{R}(\lambda_1, \lambda_2, m)|$ Let t_1 and t_2 be two tangent lines of $\mathcal{E}(\lambda_1, \lambda_2, m)$ coming from the origin. The two lines divide the complex plane into two sectors. $\mathcal{E}(\lambda_1, \lambda_2, m)$ lies in the smaller

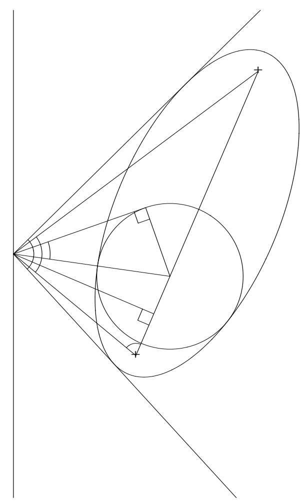


Figure 1: An Ellipse

one. Let θ be the angle of the smaller sector. (Note $0 \leq \theta < \pi$! $\theta = 0$ if $\frac{\lambda_2}{\lambda_1}$ is real.) Draw two lines ℓ_1 and ℓ_2 connecting the origin and λ_1, the origin and λ_2, respectively. Let ψ be the smaller angle between ℓ_1 and ℓ_2 ($0 < \psi \leq \theta$ and $\psi = \theta$ if $m = 0$). Let ℓ be the line joining λ_1 and λ_2. Consider now a point in $\mathcal{R}(\lambda_1, \lambda_2, m)$, which is of form (8). Its absolute value reaches its maximum when the absolute value of its denominator reaches its minimum, which means, by Lemma 2, the maximum occurs at the closest point of $\mathcal{E}(\lambda_1, \lambda_2, m)$ to the origin. Let c_0 and s_0 with $|c_0|^2 + |s_0|^2 = 1$ be the numbers for which

$$\left| |c_0|^2\lambda_1 + |s_0|^2\lambda_2 - c_0\bar{s}_0\delta \right| = \min_{|c|^2+|s|^2=1} \left\{ \left| |c|^2\lambda_1 + |s|^2\lambda_2 - c\bar{s}\delta \right| \right\}.$$

Then it is easy to see $c_0\bar{s}_0\delta = 2|c_0 s_0|m$. In another word, $c_0\bar{s}_0\delta$ has to be real and positive. Draw a circle with center $|c_0|^2\lambda_1 + |s_0|^2\lambda_2$ and radius $2|c_0 s_0|m$. The following facts are easy to establish:

- The circle lies inside the ellipse $\mathcal{E}(\lambda_1, \lambda_2, m)$ completely;

- The center of the circle lies on the line ℓ joining the two foci λ_1 and λ_2;
- The circle is tangent to the boundary of the ellipse at two points one of which is closer to the origin. Let P denote the closer point:
    1. The point P lies on the line segment joining the origin and $|c_0|^2\lambda_1 + |s_0|^2\lambda_2$, the center of the circle;
    2. $P = |c_0|^2\lambda_1 + |s_0|^2\lambda_2 - c_0\bar{s}_0\delta$, in another word, P is the closest point to the origin among all other points of $\mathcal{E}(\lambda_1, \lambda_2, m)$.

Without loss of generality, we may assume $|\lambda_1| \geq |\lambda_2|$. Rewrite (8) into

$$\frac{\lambda_1\lambda_2}{|c_0|^2\lambda_1 + |s_0|^2\lambda_2 - c_0\bar{s}_0\delta} = \lambda_1 \cdot \frac{\lambda_2}{|c_0|^2\lambda_1 + |s_0|^2\lambda_2} \cdot \frac{1}{1 - \frac{c_0\bar{s}_0\delta}{|c_0|^2\lambda_1 + |s_0|^2\lambda_2}}. \quad (11)$$

By drawing a perpendicular line from the origin to the line ℓ, one can easily see that[1]

$$\left|\frac{\lambda_2}{|c_0|^2\lambda_1 + |s_0|^2\lambda_2}\right| \leq \frac{1}{\cos(\psi/2)}.$$

One the other hand,

$$\frac{c_0\bar{s}_0\delta}{|c_0|^2\lambda_1 + |s_0|^2\lambda_2} \leq \sin\frac{\phi}{2}.$$

[1] By the elementary knowledge of triangular algebra, we know (refer to Figure 1)

$$\cos\alpha = \frac{|\lambda_2|^2 + |\lambda_1 - \lambda_2|^2 - |\lambda_1|^2}{2|\lambda_2(\lambda_1 - \lambda_2)|}. \quad (12)$$

If $\cos\alpha \leq 0$, i.e., $\pi/2 \leq \alpha < \pi$, then

$$\left|\frac{\lambda_2}{|c_0|^2\lambda_1 + |s_0|^2\lambda_2}\right| \leq 1;$$

otherwise, $\cos\alpha > 0$, i.e., $0 < \alpha < \pi/2$, then

$$\left|\frac{\lambda_2}{|c_0|^2\lambda_1 + |s_0|^2\lambda_2}\right| \leq \frac{1}{\sin\alpha} \leq \frac{1}{\cos(\psi/2)},$$

since $\alpha \geq \pi/2 - \psi/2$.

where $\phi/2 \equiv \max\{\arcsin \frac{r}{|z|}$: r and z is the radius and the center of a circle inside $\mathcal{E}(\lambda_1, \lambda_2, m)\}$. Clearly

$$\psi \leq \theta, \ \phi \leq \theta.$$

Hence it follows from (11) that

$$\left|\frac{\lambda_1\lambda_2}{|c|^2\lambda_1 + |s|^2\lambda_2 - c\bar{s}\delta}\right| \leq \frac{1}{\cos\frac{\psi}{2}(1-\sin\frac{\phi}{2})} \cdot \max\{|\lambda_1|, |\lambda_2|\}. \qquad (13)$$

Theorem 3 *An upper bound for* $|\mathcal{R}(\lambda_1, \lambda_2, m)|$ *is*

$$\frac{1}{\cos\frac{\psi}{2}(1-\sin\frac{\phi}{2})} \cdot \max\{|\lambda_1|, |\lambda_2|\} \leq \frac{1}{\cos\frac{\theta}{2}(1-\sin\frac{\theta}{2})} \cdot \max\{|\lambda_1|, |\lambda_2|\}.$$

In the case when $\mathcal{E}(\lambda_1, \lambda_2, m)$ is a circle, i.e., $\lambda_1 = \lambda_2 = \lambda$, we have an exactly answer:

$$|\mathcal{R}(\lambda, \lambda, m)| = \frac{|\lambda|^2}{|\lambda| - m} = \frac{1}{1 - \sin\frac{\phi}{2}} \cdot |\lambda|,$$

which means the upper bound by Theorem 3 overestimates it by a factor

$$\left(\cos\frac{\psi}{2}\right)^{-1} \geq 1.$$

3 An Application

In this section, we present an upper bound for $|\mathcal{F}(X)|$ (refer to (2)) by simply applying Theorems 2 and 3. To this end, we assume

$$0 \notin \mathcal{F}(A). \qquad (14)$$

It follows from (14) that $0 \notin \mathcal{F}(H) \subset \mathcal{F}(A)$. Hence Theorem 3 applies to any ellipses inside $\mathcal{F}(A)$ and Theorem 2 applies to $\mathcal{F}(X)$. Let t_1 and t_2 be the two tangent lines to the boundary of $\mathcal{F}(A)$ from the origin. Then $\mathcal{F}(A)$ lies in the smaller sector of the complex plane divided by the two lines t_1 and t_2. Let θ be the angle of the smaller sector, and $\phi/2 \equiv \max\{\arcsin \frac{r}{|z|}$: r and z is the radius and the center of a circle inside $\mathcal{F}(A)\}$.

Theorem 4 *Under the assumption of (14), we have*

$$|\mathcal{F}(X)| \le \frac{1}{\cos\frac{\theta}{2}(1-\sin\frac{\phi}{2})} \cdot |\mathcal{F}(A)| \le \frac{1}{\cos\frac{\theta}{2}(1-\sin\frac{\theta}{2})} \cdot |\mathcal{F}(A)|. \qquad (15)$$

The inequality (15) is very pessimistic when θ comes very close to π. As a matter of fact, if $\theta = \pi - \epsilon$ with ϵ very small, one can verify that the factor before $|\mathcal{F}(A)|$ satisfies

$$h(\theta) \stackrel{\text{def}}{=} \frac{1}{\cos\frac{\theta}{2}(1-\sin\frac{\theta}{2})} = \frac{1}{2\sin\frac{\epsilon}{2}\sin^2\frac{\epsilon}{4}} \sim \frac{16}{\epsilon^3}.$$

However, if θ is relative away from π, the inequality (15) will give a reasonable estimate of the magnitude for $|\mathcal{F}(X)|$. The following table illustrates roughly how fast $h(\theta)$ grows as θ approaches π.

θ	0	$\pi/4$	$\pi/2$	$3\pi/4$	$9\pi/10$	$99\pi/100$
$h(\theta)$	1	1.7534	4.8284	$3.4329 \cdot 10^{+01}$	$5.1922 \cdot 10^{+02}$	$5.1606 \cdot 10^{+05}$

Appendix 1: Compute the Closest Point of $\mathcal{E}(\lambda_1, \lambda_2, m)$ to the Origin.

We assume that (9) holds throughout.

We will not deal with the *trivial* case $\lambda_1 = \lambda_2$, i.e., $\mathcal{E}(\lambda_1, \lambda_2, m)$ is a circle.

In Section 2, we have learned several properties associated with the closest point of $\mathcal{E}(\lambda_1, \lambda_2, m)$ to the origin. Mathematically, the closest point is unique and can be found by solving certain equations. As a matter of fact, the equation to be solved eventually end up with an algebraic equation of order 4 (Two real roots of which correspond to the closest point to the origin and the furthest point from the origin, respectively; the other two roots are complex conjugates.). In this appendix, we present a numerical method based on *Newton iteration* to compute the closest point. (The furthest point can be computed in a similar way.) Recall that

$$\mathcal{E}(\lambda_1, \lambda_2, m) = \{|c|^2\lambda_1 + |s|^2\lambda_2 - c\bar{s}\delta \,:\, |c|^2 + |s|^2 = 1\}$$

by Lemma 2. A short argument will lead to

Proposition 1 *The shortest distance between the points of $\mathcal{E}(\lambda_1, \lambda_2, m)$ and the origin is the minimal values of the function*

$$f(t) = |t\lambda_1 + (1-t)\lambda_2| - 2m\sqrt{t(1-t)}, \quad 0 \le t \le 1;$$

and if $f(t_{\rm min}) = \min\limits_{0\le t\le 1} f(t)$, then the closest point $P_{\rm clt}$ to the origin is

$$P_{\rm clt} = f(t_{\rm min})\frac{t_{\rm min}\lambda_1 + (1-t_{\rm min})\lambda_2}{|t_{\rm min}\lambda_1 + (1-t_{\rm min})\lambda_2|}.$$

The longest distance between the points of $\mathcal{E}(\lambda_1, \lambda_2, m)$ and the origin is the maximal values of the function

$$F(t) = |t\lambda_1 + (1-t)\lambda_2| + 2m\sqrt{t(1-t)}, \quad 0 \le t \le 1;$$

and if $F(t_{\rm max}) = \max\limits_{0\le t\le 1} F(t)$, then the furthest point $P_{\rm ftt}$ form the origin is

$$P_{\rm ftt} = F(t_{\rm max})\frac{t_{\rm max}\lambda_1 + (1-t_{\rm max})\lambda_2}{|t_{\rm max}\lambda_1 + (1-t_{\rm max})\lambda_2|}.$$

Let $t_{\rm min}$, $P_{\rm clt}$ and $t_{\rm max}$, $P_{\rm ftt}$ be as defined in Proposition 1. Set[2]

$$b = m,\ c = \frac{|\lambda_2 - \lambda_1|}{2},\ a = \sqrt{b^2+c^2}. \tag{16}$$

The following proposition restricts the possible values of $t_{\rm min}$ and $t_{\rm max}$.

Proposition 2 $\frac{1}{2}\left(1-\frac{c}{a}\right) \le t_{\rm min}, t_{\rm max} \le \frac{1}{2}\left(1+\frac{c}{a}\right)$.

Proof: We give a proof for $t_{\rm min}$, only. As an exercise, the reader is asked to do the other. First of all, we claim *the line segment joining the origin and*

[2] With these parameters, the equation that describes the boundary of the ellipse $\mathcal{E}(\lambda_1, \lambda_2, m)$ can now be written as

$$|z-\lambda_1| + |z-\lambda_2| = 2a.$$

$t_{\min}\lambda_1 + (1 - t_{\min})\lambda_2$ *is perpendicular to the boundary of the ellipse at* P_{clt}. This can be easily seen. To see what are the possible values that $t_{\min}$ can take, we shall determine how big the distance between each of the two foci and $t_{\min}\lambda_1 + (1 - t_{\min})\lambda_2$ could be. To this end, let's perform a transformation (a shift and a rotation, generally) on the complex plane such that the foci of the ellipse are transformed to $(-c, 0)$ and $(c, 0)$, respectively. Suppose P_{clt} is transformed to $(a\cos\beta, b\sin\beta)$ with $0 \le \beta < 2\pi$. Assume, for the moment, $\beta \ne 0, \pi$. The point $t_{\min}\lambda_1 + (1 - t_{\min})\lambda_2$ after the transformation can be located by finding the intersection of the new x-axis and the line passing through P_{clt} and perpendicular to the boundary of the ellipse. It is easy to see that the equation of the line is

$$\frac{y - b\sin\beta}{x - a\cos\beta} \cdot \frac{b\cos\beta}{-a\sin\beta} = -1.$$

Letting $y = 0$ gives $x = \frac{c^2}{a}\cos\beta$. In the other word, The point $t_{\min}\lambda_1 + (1 - t_{\min})\lambda_2$ is transformed to $(\frac{c^2}{a}\cos\beta, 0)$. It is easily verified that this remains true even for $\beta = 0, \pi$. Therefore the distance between each focus and $t_{\min}\lambda_1 + (1 - t_{\min})\lambda_2$ is between $c - \frac{c^2}{a}$ and $c + \frac{c^2}{a}$. Now, note

$$\begin{aligned} |t_{\min}\lambda_1 + (1 - t_{\min})\lambda_2 - \lambda_1| &= 2(1 - t_{\min})c, \\ |t_{\min}\lambda_1 + (1 - t_{\min})\lambda_2 - \lambda_2| &= 2t_{\min}c, \end{aligned}$$

from which the desired result follows. ∎

The following proposition is easy to establish by using geometrical arguments.

Proposition 3

- The most general case $b > 0$ and $c > 0$:

    1. *If* $|\lambda_1| = |\lambda_2|$, *then* $t_{\min} = \frac{1}{2}$, *and there are two* $t_{\max}$ *one of which lies in the open interval* $(\frac{1}{2}(1 - \frac{c}{a}), \frac{1}{2})$ *while the other in the open interval* $(\frac{1}{2}, \frac{1}{2}(1 + \frac{c}{a}))$. *And moreover the sum of the two* $t_{\max}$ *is equal to 1;*

2. *If* $|\lambda_1| > |\lambda_2|$, *then* $\frac{1}{2}\left(1-\frac{c}{a}\right) \le t_{\min} < \frac{1}{2}$, *and* $\frac{1}{2} < t_{\max} \le \frac{1}{2}\left(1+\frac{c}{a}\right)$;
3. *If* $|\lambda_1| < |\lambda_2|$, *then* $\frac{1}{2} < t_{\min} \le \frac{1}{2}\left(1+\frac{c}{a}\right)$, *and* $\frac{1}{2}\left(1-\frac{c}{a}\right) \le t_{\max} < \frac{1}{2}$;

- The case $c = 0$: $t_{\min} = t_{\max} = \frac{1}{2}$;
- The case $b = 0$ and $c > 0$:

1. *If* $|\lambda_1| = |\lambda_2|$, *then* $t_{\min} = \frac{1}{2}$, *and* $t_{\max} = 0$ *and* 1;
2. *If* $|\lambda_1| > |\lambda_2|$, *then* $t_{\min} = 0$ *and* $t_{\max} = 1$;
3. *If* $|\lambda_1| < |\lambda_2|$, *then* $t_{\min} = 1$ *and* $t_{\max} = 0$.

Set $\lambda_j = x_j + iy_j$ where x_j, y_j for $j = 1, 2$ are real and $i = \sqrt{-1}$. Let (recall $m = b$ in (16))

$$\begin{aligned} f_1(t) &= |t\lambda_1 + (1-t)\lambda_2| \qquad (17) \\ &= \sqrt{(tx_1 + (1-t)x_2)^2 + (ty_1 + (1-t)y_2)^2}, \\ f_2(t) &= 2m\sqrt{t(1-t)}. \qquad (18) \end{aligned}$$

Then $f(t) = f_1(t) - f_2(t)$. Taking derivatives gives

$$\begin{aligned} f_1'(t) &= \frac{(x_1-x_2)(tx_1+(1-t)x_2)+(y_1-y_2)(ty_1+(1-t)y_2)}{\sqrt{(tx_1+(1-t)x_2)^2+(ty_1+(1-t)y_2)^2}}, \\ f_1''(t) &= -\frac{\left[(x_1-x_2)(tx_1+(1-t)x_2)+(y_1-y_2)(ty_1+(1-t)y_2)\right]^2}{\left[(tx_1+(1-t)x_2)^2+(ty_1+(1-t)y_2)^2\right]^{3/2}} \\ &\quad + \frac{(x_1-x_2)^2+(y_1-y_2)^2}{\sqrt{(tx_1+(1-t)x_2)^2+(ty_1+(1-t)y_2)^2}} \\ &= \frac{(x_2y_1-x_1y_2)^2}{\left[(tx_1+(1-t)x_2)^2+(ty_1+(1-t)y_2)^2\right]^{3/2}} > 0, \\ f_2'(t) &= \frac{m(1-2t)}{\sqrt{t(1-t)}}, \\ f_2''(t) &= -\frac{m}{2[t(1-t)]^{3/2}} < 0. \end{aligned}$$

From these formula it follows $f''(t) = f_1''(t) - f_2''(t) > 0$. Since $f'(t) = f_1'(t) - f_2'(t) \to -\infty$ as $t \to 0^+$, and $f'(t) = f_1'(t) - f_2'(t) \to +\infty$ as $t \to 1^-$, we see

$$f'\left(\frac{1}{2}\left(1 - \frac{c}{a}\right)\right) < 0,\ f'\left(\frac{1}{2}\left(1 + \frac{c}{a}\right)\right) > 0.$$

$f'(t)$ has exactly one zero between $\frac{1}{2}\left(1 - \frac{c}{a}\right)$ and $\frac{1}{2}\left(1 + \frac{c}{a}\right)$ which is $t_{\min}$.

Newton iteration can now be applied to find the point of interest. One remaining question is how to get a good initial guess. The following way is used in my MATLAB code. It is assumed $|\lambda_1| \geq |\lambda_2|$ (otherwise, simply swap λ_1 and λ_2). Then $\frac{1}{2}\left(1 - \frac{c}{a}\right) \leq t_{\min} < \frac{1}{2}$. Our motivation for choosing an initial guess is based on Propositions 2 and 3 and the observation that the t making $|t\lambda_1 + (1-t)\lambda_2|$ reach its minimum should be close to $t_{\min}$. Here are the formulas for τ, an initial guess of $t_{\min}$ (refer to Figure 1 and the equation (12)).

- If $\cos\alpha \leq 0$, set $\tau = \frac{1}{2}\left(1 - \frac{c}{a}\right)$;
- If $\cos\alpha > 0$, set $\tau = \frac{|\lambda_2|\cos\alpha}{2c}$.

Appendix 2: The Tangent Lines of the Ellipse $\mathcal{E}(\lambda_1, \lambda_2, m)$ from the Origin.

Let the assignments to x_j, y_j, a, b and c in Appendix 1 hold throughout this appendix. Assume also (9) holds throughout. The slopes k of the two tangent lines from the origin to the boundary of the ellipse are the roots of the following algebraic equation of order 2:

$$\begin{aligned} [(4a^2 - (x_1+x_2)^2 - (y_1-y_2)^2]k^2 \quad &+ \quad 4(x_1y_2 + x_2y_1)k \\ &+ \quad [(4a^2 - (x_1-x_2)^2 - (y_1+y_2)^2] = 0. \end{aligned} \tag{19}$$

A by-product of this is that *the assumption (9) holds if and only if (19) has real solutions.*

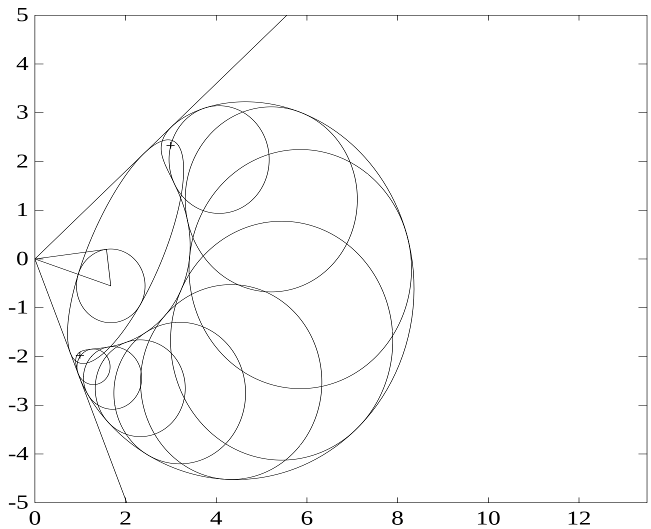


Figure 2: $\mathcal{E}(1-2i, 3+2.3i, 1)$

Appendix 3: Some Typical Graphes for $\mathcal{R}(\lambda_1, \lambda_2, m)$.

This appendix displays four graphes Figures 2–5 for $\mathcal{R}(\lambda_1, \lambda_2, m)$ in different situations. They are telling us what a typical $\mathcal{R}(\lambda_1, \lambda_2, m)$ looks like.

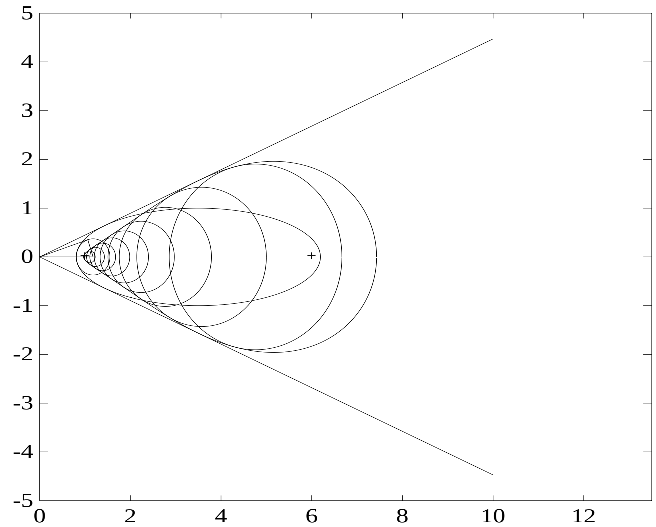


Figure 3: $\mathcal{E}(1, 6, 1)$

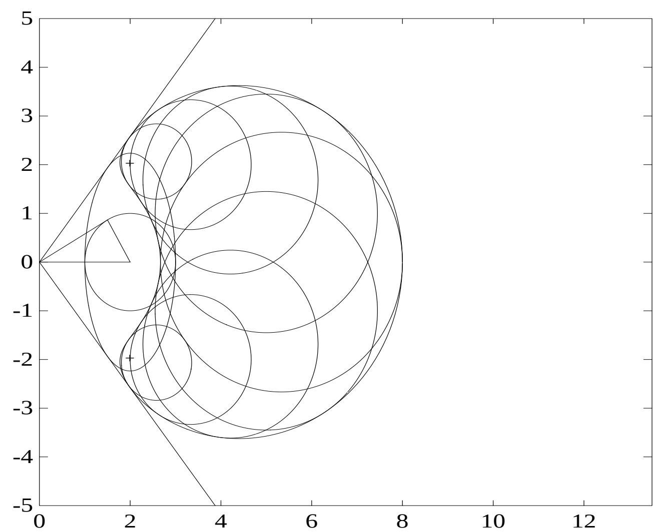


Figure 4: $\mathcal{E}(2 - 2i, 2 + 2i, 1)$

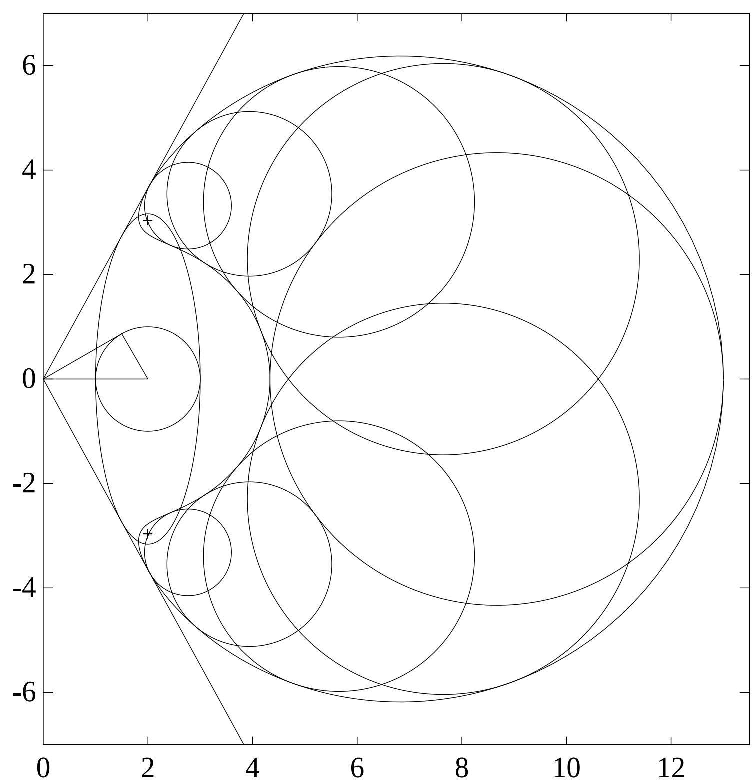


Figure 5: $\mathcal{E}(2+3i, 2-3i, 1)$