Linear Algebra and its Applications 435 (2011) 494–513

Contents lists available at ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

Characterization and construction of the nearest defective matrix via coalescence of pseudospectral components

Rafikul Alam [a], Shreemayee Bora [a], Ralph Byers [b,1], Michael L. Overton [c,*]

[a] Indian Institute of Technology Guwahati, Guwahati 7810390, Assam, India
[b] Department of Mathematics, University of Kansas, Lawrence, KS 66045, USA
[c] Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA

ARTICLE INFO

Submitted by V. Mehrmann

Dedicated to Pete Stewart on the occasion of his 70th birthday.

Keywords:
Multiple eigen values
Saddle point
Pseudospectrum

ABSTRACT

Let A be a matrix with distinct eigenvalues and let $w(A)$ be the distance from A to the set of defective matrices (using either the 2-norm or the Frobenius norm). Define Λ_ϵ, the ϵ-pseudospectrum of A, to be the set of points in the complex plane which are eigenvalues of matrices $A+E$ with $\|E\| < \epsilon$, and let $c(A)$ be the supremum of all ϵ with the property that Λ_ϵ has n distinct components. Demmel and Wilkinson independently observed in the 1980s that $w(A) \geqslant c(A)$, and equality was established for the 2-norm by Alam and Bora in 2005. We give new results on the geometry of the pseudospectrum near points where first coalescence of the components occurs, characterizing such points as the lowest generalized saddle point of the smallest singular value of $A - zI$ over $z \in \mathbb{C}$. One consequence is that $w(A) = c(A)$ for the Frobenius norm too, and another is the perhaps surprising result that the minimal distance is attained by a defective matrix in all cases. Our results suggest a new computational approach to approximating the nearest defective matrix by a variant of Newton's method that is applicable to both generic and nongeneric cases. Construction of the nearest defective matrix involves some subtle numerical issues which we explain, and we present a simple backward error analysis showing that a certain singular vector residual measures how close the computed

[*] Corresponding author. Partially supported by the National Science Foundation under award DMS-0714321.
E-mail addresses: rafik@iitg.ac.in (R. Alam), shbora@iitg.ac.in (S. Bora), overton@cs.nyu.edu (M.L. Overton).
[1] Deceased.

0024-3795/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2010.09.022

matrix is to a truly defective matrix. Finally, we present a result giving lower bounds on the angles of wedges contained in the pseudospectrum and emanating from generic coalescence points. Several conjectures and questions remain open.

1. Introduction and history

A matrix is defective if it is not diagonalizable. Given a complex n-by-n matrix A, we consider the quantity

$$w(A) = \inf \{\|A - B\| \, |B \text{ is defective} \},$$

where we restrict the norm to be the 2-norm or the Frobenius norm. In other words, $w(A)$ is the distance to the set of matrices whose Jordan canonical form has a block of size 2 or more, or equivalently, which have a nonlinear elementary divisor. An eigenvalue associated with such a Jordan block is called defective as its geometric multiplicity (the number of linearly independent eigenvectors associated with it) is less than its algebraic multiplicity. By a multiple eigenvalue we mean one whose algebraic multiplicity is greater than one. The distance to the set of matrices with a defective eigenvalue is the same as the distance to the set of matrices with a multiple eigenvalue, since an arbitrarily small perturbation to a matrix with a nondefective multiple eigenvalue makes the eigenvalue defective. In this paper we show for the first time (in Theorem 6 below) that as long as A has distinct eigenvalues, a defective matrix B attaining the infimum in the definition of $w(A)$ always exists. We refer to such a B as the nearest defective matrix, although it is not necessarily unique.

The search for insight into the distance $w(A)$ goes back to the 1960s. In his classic work [27], Wilkinson defined the condition number of a simple eigenvalue λ as $1/|y^*x|$, where y and x are, respectively, normalized left and right eigenvectors associated with λ; this is infinite for a double defective eigenvalue as $y^*x = 0$. He observed that even if the eigenvalues are well separated from each other, they can still be very ill-conditioned, and gave an example of a matrix A illustrating the point. He then wrote “It might be expected that there is a matrix close to A which has some nonlinear elementary divisors and we can readily see that this is true ...".

In his Ph.D. thesis [9], Demmel introduced $w(A)$ under the name *diss*(A, *path*) as well as a second quantity that we will denote by $c(A)$ under the name *diss*(A, *region*). The former is defined to be the distance from a fixed matrix A to the nearest matrix with multiple eigenvalues, *path* referring to the path traveled by the eigenvalues in the complex plane under a smoothly varying perturbation to A, and *diss* being an abbreviation for *dissociation*. The second quantity is defined as the largest ϵ such that “the area swept out by the eigenvalues under perturbation" – that is the set of z in the complex plane that are eigenvalues of matrices differing from A by norm at most ϵ, the set now commonly known as the ϵ-pseudospectrum of A – consists of n disjoint regions, or connected components. Demmel observed that for all norms, $w(A) \geqslant c(A)$, because under the first definition, two eigenvalues must travel to the same point z under the *same* perturbation, while under the second definition, two eigenvalues must travel to the same point z under the *same size* perturbation. He indicated that $w(A) > c(A)$ for a specially chosen norm, and mentioned that it is an open question as to whether equality holds in the case of the 2-norm and the Frobenius norm. Demmel discussed these issues further in [10], where the first definition *diss*(A, *path*) remains unchanged, but the second definition *diss*(A, *region*) is replaced by a more informal discussion of pseudospectral components.[2]

About the same time Wilkinson made a detailed study of the distance to the nearest defective matrix in [29]. He wrote: “A problem of primary interest to us is the distance, measured in the 2-norm,

[2] Demmel’s quantities *diss*(A, *path*) and *diss*(A, *region*) were actually defined more generally in terms of coalescence of eigenvalues from two specified partitions of the spectrum of A, and their associated pseudospectral components, but by taking the minimum of these quantities over all partitions, we obtain the ones that are relevant in our context.

of our matrix A from matrices having a multiple eigenvalue" and "We expect that [when the eigenvalue condition number is large] A will be, at least in some sense, close to a matrix having a double eigenvalue. A major objective of this paper is to quantify this statement." He went on to give many examples and bounds, but there is, surprisingly, still no mention of pseudospectra. The same is true of another well-known paper of Wilkinson on the same subject [28]. In their book [25], Trefethen and Embree discuss the early history of pseudospectra and speculate as to why, with his life-long interest in eigenvalues and rounding errors, Wilkinson did not arrive at the idea of pseudospectra much earlier than he did.

In fact, it was only in his last paper [30] that Wilkinson discussed the notion of pseudospectra, under the name "fundamental domain." He defined $D(\epsilon)$, for any operator norm, as the set of points in the complex plane satisfying $\|(A-zI)^{-1}\|^{-1} \leqslant \epsilon$, establishing that this precisely identifies all z which can be induced as eigenvalues by perturbations E with $\|E\| \leqslant \epsilon$. He emphasized "the sheer economy of this theorem." This observation may be called the fundamental theorem of pseudospectra: that, for operator norms, the definition of pseudospectra in terms of norm-bounded perturbations is equivalent to the definition using the norm of the resolvent, which reduces, in the case of the 2-norm, to $\sigma_n(A-zI) \leqslant \epsilon$, where σ_n denotes smallest singular value. The importance of this equivalence has been emphasized by Trefethen for many years; Wilkinson's emphatic comments in this regard seem to be the first in the literature, although Varah [26] comes close: his observation is explicit only in one direction, however, and is only for the 2-norm.

Wilkinson's paper [30] continues: "The behaviour of the domain $D(\epsilon)$ as ϵ increases from zero is of fundamental interest to us ... When ϵ is sufficiently small [and A has distinct eigenvalues], $D(\epsilon)$ consists of n isolated domains ... A problem of basic interest to us is the smallest value of ϵ for which one of these domains coalesces with one of the others...." Like Demmel, he was aware that coalescence of two components of $D(\epsilon)$ at z for a particular ϵ shows that each of two eigenvalues can be moved to the same z by a perturbation of norm ϵ but that this does not imply that a single perturbation exists that can induce a double eigenvalue. Towards the end of the paper (pp. 272–273) he gave an example suggesting that, for the ∞-norm, $w(A)$ might be larger than $c(A)$, and also hinted that equality might hold, possibly in general, for the 2-norm.

At a conference at the National Physical Laboratory in memory of Wilkinson, Demmel [11] returned to the subject of the nearest defective matrix. Specifically, he observed that the fact that a matrix with an ill-conditioned eigenvalue must be near a defective matrix is a special case of the more general property of many computational problems that the distance from a particular problem instance to a nearest ill-posed problem is inversely related to the condition number of the problem instance. He also commented that "a simple guaranteed way to compute the distance to the nearest defective matrix remains elusive."

In summary, Demmel and Wilkinson independently observed in the 1980s that, in general, $w(A) \geqslant c(A)$, and Demmel raised the question of whether $w(A) = c(A)$ for the 2-norm and the Frobenius norm. Subsequent work on the nearest defective matrix, notably by Lippert and Edelman [16] and Malyshev [19], did not address this question, which was finally answered affirmatively for the 2-norm by Alam and Bora [1]. That the same equality holds for the Frobenius norm is proved for the first time in the present paper (Theorem 6). Furthermore, neither Demmel nor Wilkinson made a clear statement as to whether equality might hold for ℓ_p operator norms with $p \neq 2$, and indeed this remains an open question. It is conjectured by Alam and Bora [1, p. 294] that $w(A) = c(A)$ for all operator norms.

For more information on the history of the distance to the nearest defective matrix, including a comprehensive catalogue of lower and upper bounds, see [2].

2. Coalescence of pseudospectra and generalized saddle points

We assume throughout that $A \in \mathbb{C}^{n\times n}$ is fixed and has n distinct eigenvalues, with $n > 1$. Following [25], the open ϵ-pseudospectrum of a matrix $A \in \mathbb{C}^{n\times n}$ is defined, for $\epsilon > 0$, by

$$\Lambda_\epsilon = \{z \in \mathbb{C} \,|\det(A+E-zI) = 0 \text{ for some } E \text{ with } \|E\| < \epsilon\}.$$

It is easily proved using the singular value decomposition (SVD) that, for both the 2-norm and the Frobenius norm,

$$\Lambda_\epsilon = \{z \in \mathbb{C} \,|\sigma_n(A-zI) < \epsilon\},$$

where σ_n denotes smallest singular value. For each $\epsilon > 0$, the pseudospectrum Λ_ϵ has at most n (connected) components, each of which is an open set and contains at least one eigenvalue of A.

It is a consequence of continuity of eigenvalues that for ϵ small enough, Λ_ϵ has n distinct components. Define

$$c(A) = \sup\{\epsilon \mid \Lambda_\epsilon \text{ has } n \text{ components}\}.$$

Note that, by definition, $c(A)$ is the same for both the 2-norm and the Frobenius norm. Clearly, if Λ_ϵ has n components, then for all perturbation matrices $E \in \mathbb{C}^{n\times n}$ with $\|E\| < \epsilon$, the matrix $A+E$ has n distinct eigenvalues and, in particular, $A+E$ is not defective. Hence, as discussed in the previous section, it is intuitively clear that, for all norms,

$$w(A) \geqslant c(A).$$

Theorem 1 below states that $w(A) = c(A)$ for the 2-norm and gives a characterization of a nearest matrix with multiple eigenvalues, but first we state a key lemma used in the proof. Whenever we refer to singular vectors, we mean with unit length in the 2-norm, as is standard.

Lemma 1. *Suppose $A - zI$ has smallest singular value $\epsilon > 0$, with corresponding left and right singular vectors u and v satisfying $(A - zI)v = \epsilon u$. Then z is an eigenvalue of $B = A - \epsilon uv^*$ with geometric multiplicity one and corresponding left and right eigenvectors u and v, respectively. Furthermore, if $u^*v = 0$, then z has algebraic multiplicity greater than one, so it is a defective eigenvalue.*

Proof. Let $A - zI$ have singular value decomposition $U\Sigma V^*$, where u and v are, respectively, the last columns of U and V. Clearly $A - zI - \epsilon uv^*$ has nullity one, so z is an eigenvalue of B with geometric multiplicity one, with

$$(B - zI)v = (A - zI)v - \epsilon u = 0, \quad (B - zI)^*u = (A - zI)^*u - \epsilon v = 0.$$

The last part follows from the well-known property that the left and right eigenvectors corresponding to a simple eigenvalue cannot be mutually orthogonal [13, Lemma 6.3.10]. □

Theorem 1 (R. Alam and S. Bora). *For the 2-norm, $w(A) = c(A)$. Furthermore, there exists at least one point $\tilde{z}$, which we call a first-coalescence point, which lies in the closure of two distinct components of $\Lambda_{\tilde{\epsilon}}$, where $\tilde{\epsilon} = \sigma_n(A - \tilde{z}I) = c(A)$. Let $A - \tilde{z}I$ have singular value decomposition $A = U\Sigma V^*$ and define $B = A - \tilde{\epsilon}\, UDV^*$, where D is the rank-one matrix* $\operatorname{diag}(0, \ldots, 0, 1)$ *if the multiplicity of the smallest singular value of $A - \tilde{z}I$ is one and the rank-two matrix* $\operatorname{diag}(0, \ldots, 0, 1, 1)$ *otherwise. Then $\tilde{z}$ is a multiple eigenvalue of B and no other matrix with multiple eigenvalues is closer to A in the 2-norm. If the multiplicity of the smallest singular value of $A - \tilde{z}I$ is one, the corresponding left and right singular vectors u and v satisfy $u^*v = 0$, so B is defective, and is a nearest defective matrix in the Frobenius norm as well as the 2-norm, so $w(A) = c(A)$ for both norms in this case. In general, for the Frobenius norm, $w(A) \geqslant c(A)$.*

This result is given in [1, Theorem 5.1]. See also [6], where a different proof of the inequality $w(A) \geqslant c(A)$ is provided.

An important part of Theorem 1 is the orthogonality of the left and right singular vectors in the case that the smallest singular value of $A - \tilde{z}I$ is simple. Let us define $f : \mathbb{R}^2 \to \mathbb{R}$ by $f(x, y) = \sigma_n(A - (x + iy)I)$. If $\sigma_{n-1}(A - \tilde{z}I) > \sigma_n(A - \tilde{z}I) = f(\tilde{x}, \tilde{y}) > 0$, with $\tilde{z} = \tilde{x} + i\tilde{y}$, then f is differentiable, in fact real analytic, at $(\tilde{x}, \tilde{y})$, with

$$\frac{\partial f}{\partial x}(\tilde{x}) = -\Re(u^*v), \quad \frac{\partial f}{\partial y}(\tilde{y}) = \Im(u^*v),$$

where u and v are, respectively, the left and right singular vectors corresponding to $\sigma_n(A - \tilde{z}I)$ and $\Re$ and $\Im$, respectively, denote real and imaginary parts.[3] Identifying $\mathbb{C}$ with $\mathbb{R}^2$, we may rewrite $f(\Re(z), \Im(z))$ as $f(z) = \sigma_n(A - zI)$ and rewrite the partial derivative formulas concisely as

[3] This is a consequence of the equivalence of the SVD to a $2n$ by $2n$ Hermitian eigenvalue problem and results on derivatives of eigenvalues that go back to Rayleigh.

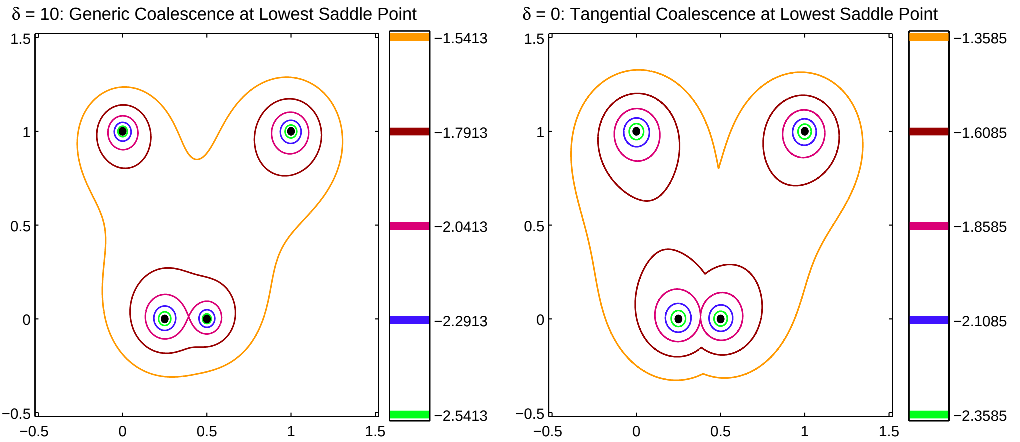

Fig. 1. Coalescence of pseudospectral components for the matrix $A(\delta)$. Solid curves denote contours of $f(x, y) = \sigma_n(A - (x + iy)I)$. For small contour levels ϵ, the ϵ-pseudospectrum consists of 4 components, one surrounding each eigenvalue (shown as black dots). When the contour level is increased to $\tilde{\epsilon} = c(A)$ (the middle value plotted), coalescence of components occurs. On the left, the case $\delta = 10$, for which the first-coalescence point is a smooth saddle point of f. On the right, the case $\delta = 0$, for which $A(\delta)$ is block diagonal and coalescence takes place tangentially, so that the first-coalescence point is a nonsmooth saddle point of f. The legend showing the contour levels uses a logarithmic scale (base 10). The critical value $\tilde{\epsilon}$ was computed in both cases using Method k described in Section 4.

$$\nabla f(\tilde{z}) = -v^*u.$$

Thus, the assumption that the smallest singular value of $A - \tilde{z}I$ is simple implies that $\tilde{z}$ is a stationary point of the function $f(z) = \sigma_n(A - zI)$, or, more precisely, a saddle point, since it is obviously neither a local minimizer nor a local maximizer. Thus, if it were the case that the smallest singular value of $A - \tilde{z}I$ is always simple, the first-coalescence point $\tilde{z}$ could be characterized simply as a lowest saddle point of the smooth function $f(z)$. Indeed, Lippert and Edelman [16] claimed that, in the case of the Frobenius norm, the nearest defective matrix is $A - \sigma uv^*$, where σ, u and v are, respectively, the smallest singular value and corresponding left and right singular vectors of $A - \tilde{z}I$, with the orthogonality property $u^*v = 0$, and where $\tilde{z}$ is the lowest critical point of $f(z)$, provided that the smallest singular value of $A - \tilde{z}I$ is simple.

It sometimes happens that the smallest singular value of $A - \tilde{z}I$ is double, that is $\sigma_{n-1}(A - \tilde{z}I) = \sigma_n(A - \tilde{z}I)$. In this case, the function $f(z)$ is usually[4] not differentiable at $\tilde{z}$. This case always occurs when A is normal, in which case the boundaries of the pseudospectral components are circles and coalescence of components can only occur when two component boundaries are tangent to each other. It can also occur when A is not normal.

The two cases are well illustrated by the example

$$A(\delta) = \begin{bmatrix} 0.25 & 10 & 0 & \delta \\ 0 & i & 0 & 0 \\ 0 & 0 & 0.5 & 10 \\ 0 & 0 & 0 & 1+i \end{bmatrix}$$

with $\delta = 10$ (a typical example whose first-coalescence point $\tilde{z}$ is a smooth saddle point with $\sigma_n(A - \tilde{z}I)$ simple) and $\delta = 0$ (a nongeneric example for which, because of the block diagonal structure, coalescence takes place tangentially and for which $\sigma_n(A - \tilde{z}I)$ is double). Both cases are, respectively, illustrated, courtesy of EigTool [31], in the left and right sides of Fig. 1.

In the double singular value case, construction of a nearest matrix with multiple eigenvalues is easy for the 2-norm: simply subtract a *rank-two* term from A instead of a rank-one term, as stated

[4] Conjecture 1 below speculates that f is *never* differentiable at $\tilde{z}$ in this case.

in Theorem 1; however, construction of a nearest defective matrix, or a nearest matrix with multiple eigenvalues for the Frobenius norm, is not so simple. We address this issue in the next section.

We will first establish in this section that although f may not be differentiable at a first-coalescence point $\tilde{z}$, nonetheless such a point is *always* a lowest saddle point of $f(z) = \sigma_n(A - zI)$, provided we generalize the familiar notion of smooth saddle point to a possibly nonsmooth saddle point, as follows. We continue to identify $\mathbb{C}$ with $\mathbb{R}^2$ in the following definition.

Definition 1. The *Clarke generalized gradient* [8,4] of a locally Lipschitz function $\phi : \mathbb{C} \to \mathbb{R}$ at a point $\hat{z}$ is the set

$$\partial\phi(\hat{z}) = \text{convex hull } \left\{ \lim_{z_k \to \hat{z}} \nabla\phi(z_k) \,|\phi \text{ is differentiable at } z_k \right\}.$$

A *stationary point* of ϕ is a point $\hat{z}$ at which $0 \in \partial\phi(\hat{z})$. A *saddle point* of ϕ is a stationary point of ϕ that is not a local extremum. If ϕ is differentiable at a saddle point $\hat{z}$ and $\partial\phi(\hat{z}) = \{\nabla\phi(\hat{z})\}$, then $\hat{z}$ is said to be a *smooth saddle point*; otherwise, it is said to be a *nonsmooth* saddle point. A *lowest* saddle point is a saddle point $\hat{z}$ for which $\phi(\hat{z})$ is minimal.

We will need the following lemma of Burke, Lewis and Overton (see [5, p. 88] and its corrigendum).

Lemma 2. *If $z \in \text{cl}\,\Lambda_\epsilon$ and if $v^*(A - zI)v \neq 0$ for some right singular vector v of $A - zI$ corresponding to $\sigma_n(A - zI)$, then the open disk with center v^*Av and radius $|z - v^*Av|$ is contained in Λ_ϵ.*

Denoting the open disk in the complex plane with center $c \in \mathbb{C}$ and containing $b \in \mathbb{C}$ on its boundary by $D(c, b)$, Lemma 2 states that $z \in \Lambda_\epsilon$ implies that the disk $D(v^*Av, z) \subset \Lambda_\epsilon$, where v is a right singular vector corresponding to $\sigma_n(A - zI)$. (If $c = b$, then $D(c, b) = \emptyset$.) Generically, points $z \in \mathbb{C}$ are *nondegenerate* in the sense that there is a right singular vector v corresponding to the smallest singular value of $A - zI$ for which $v^*(A - zI)v \neq 0$, i.e. for which $D(v^*Av, z) \neq \emptyset$. However, matrices typically also have degenerate points including all smooth stationary points of $f(z) = \sigma_n(A - zI)$.

The main result of the next theorem, the existence of tangent disks inside the coalescing pseudospectral components, is illustrated by the example shown on the right side of Fig. 1.

Theorem 2 (Tangent disks). *Let $\tilde{z}$ be a first-coalescence point satisfying $\tilde{\epsilon} = f(\tilde{z})$, $\tilde{z} \in \text{cl}\,\Omega$, $\tilde{z} \in \text{cl}\,\widehat{\Omega}$, $\tilde{z} \notin \Omega$, and $\tilde{z} \notin \widehat{\Omega}$, where Ω and $\widehat{\Omega}$ are distinct components of $\Lambda_{\tilde{\epsilon}}$. Suppose that there does not exist a sequence of points $z_k \in \Lambda_{\tilde{\epsilon}}$ converging to $\tilde{z}$ such that* $\lim_{k\to\infty} \nabla f(z_k) = 0$. *Then we have the following:*

1. *Let $z_k \in \Omega$ and $\hat{z}_k \in \widehat{\Omega}$ be sequences both converging to $\tilde{z}$. Let v_k and $\hat{v}_k$ be right singular vectors corresponding to $\sigma_n(A - z_kI)$ and $\sigma_n(A - \hat{z}_kI)$, respectively, and assume, by taking subsequences if necessary, that v_k and $\hat{v}_k$ converge to limits v and $\hat{v}$, respectively. Then v and $\hat{v}$ are both right singular vectors for $\sigma_n(A - \tilde{z}I)$, and Ω and $\widehat{\Omega}$, respectively, contain nonempty disks $D(v^*Av, \tilde{z})$ and $D(\hat{v}^*A\hat{v}, \tilde{z})$, whose closures are mutually tangent at $\tilde{z}$.*
2. *The first-coalescence point $\tilde{z}$ is not in the closure of any other component of $\Lambda_{\tilde{\epsilon}}$.*
3. *The right singular vectors v and $\hat{v}$ are linearly independent, and the smallest singular value of $A - \tilde{z}I$ has multiplicity two.*
4. *Let $g = -v^*u$ and $\hat{g} = -\hat{v}^*\hat{u}$, where $u = (A - \tilde{z}I)v/\tilde{\epsilon}$ and $\hat{u} = (A - \tilde{z}I)\hat{v}/\tilde{\epsilon}$ are, respectively, left singular vectors for $\sigma_n(A - \tilde{z}I)$. Then g and $\hat{g}$ are, respectively, limits of $\nabla f(z_k)$ and $\nabla f(\hat{z}_k)$, and they satisfy $\mu g + (1 - \mu)\hat{g} = 0$ for some $\mu \in (0, 1)$.*

Proof. Since singular values are continuous, it follows that

$$\sigma_n(A - \tilde{z}I) = \lim_{k\to\infty} \sigma_n(A - z_kI) = \lim_{k\to\infty} \|(A - z_kI)v_k\|_2 = \|(A - \tilde{z}I)v\|_2 \,.$$

Hence, v is a right singular vector of $A - \tilde{z}I$ corresponding to its smallest singular value. By Lemma 2, each of the disks $C_k = D(v_k^*Av_k, z_k)$ lies in some component of $\Lambda_{\tilde{\epsilon}}$. Each of the points $z_k \in \Omega$ lies on the boundary of its disk C_k, so each C_k lies in Ω. The radius of C_k is

$$|z_k - v_k^*Av_k| = |v_k^*(A - z_kI)v_k| = f(z_k)\,|v_k^*u_k| = f(z_k)\,|\nabla f(z_k)|$$

where u_k is the left singular vector corresponding to v_k and the right-hand side is well defined because $z_k \in \Omega$, so $\sigma_n(A - z_kI)$ is simple. Both $f(z_k) = \sigma_n(A - z_kI)$ and $\nabla f(z_k)$ are bounded away from zero, the former as it converges to $\tilde{\epsilon}$ and the latter by assumption. Hence, the limiting disk

$$C = \lim_{k\to\infty} C_k = D(v^*Av, \tilde{z})$$

exists, has positive radius and is contained in Ω.

Similarly, by choosing a sequence of points $\hat{z}_k \in \widehat{\Omega}$ one can infer that there is a nonempty disk $\widehat{C} = D(\hat{v}^*A\hat{v}, \tilde{z}) \subset \widehat{\Omega}$ where $\hat{v}$ is a right singular vector of $A - \tilde{z}I$ corresponding to its smallest singular value. The point $\tilde{z}$ is a common boundary point of both the disks $C \subset \Omega$ and $\widehat{C} \subset \widehat{\Omega}$. This proves the first claim.

It follows that $\tilde{z}$ cannot lie in the closure of a third component of $\Lambda_{\tilde{\epsilon}}$, because that would also contain an open disk whose boundary includes $\tilde{z}$. However, at least two of any three open disks that share a common boundary point have nonempty intersection. This contradicts the fact that each disk is contained in a separate component. This proves the second claim.

The disks C and $\widehat{C}$ are disjoint and, in particular, their centers, v^*Av and $\hat{v}^*A\hat{v}$, are different. So, the unit-length vectors v and $\hat{v}$ cannot be scalar multiples of each other. Hence, v and $\hat{v}$ are linearly independent right singular vectors corresponding to $\sigma_n(A - \tilde{z}I)$ and therefore $\sigma_n(A - \tilde{z}I)$ has multiplicity at least two. The multiplicity cannot be more than two, because in that case, there would be a rank-three perturbation matrix E with $\|E\|_2 = \tilde{\epsilon}$ for which $A + E$ would have eigenvalue $\tilde{z}$ of multiplicity at least three. However, it follows from the minimality property of $w(A)$ that for all $\delta \in (0, 1), A + \delta E$ has simple eigenvalues each of which lies in a different component of $\Lambda_{\delta\tilde{\epsilon}}$. This in turn implies that every neighborhood of $\tilde{z}$ has nonempty intersection with at least three components of $\Lambda_{\delta\tilde{\epsilon}}$ in contradiction to the first claim. This proves the third claim.

The function $f(z) = \sigma_n(A - zI)$ is smooth at $z \in \Omega$ and $z \in \widehat{\Omega}$, because the smallest singular value is simple in these open sets. We have

$$f(z_k)\nabla f(z_k) = -f(z_k)v_k^*u_k = -v_k^*(A - z_kI)v_k = z_k - v_k^*Av_k,$$

the “spoke" of C_k, say s_k, that radiates from its center to the boundary point. Similarly,

$$f(\hat{z}_k)\nabla f(\hat{z}_k) = \hat{z}_k - \hat{v}_k^*A\hat{v}_k,$$

the spoke of $\widehat{C}_k$, say $\hat{s}_k$, that radiates from its center to the boundary point. In the limit s_k and $\hat{s}_k$ converge to the spokes of the disks C and $\widehat{C}$ that run from their centers to the common boundary point $\tilde{z}$. Thus the gradients, respectively, have limits $g = -v^*u$ and $\hat{g} = -\hat{v}^*\hat{u}$. As the closures of C and $\widehat{C}$ are tangent at $\tilde{z}$, these limits satisfy $\hat{g} = -\kappa g$ for some real $\kappa > 0$. Setting $\mu = \kappa/(\kappa + 1)$ completes the proof of the theorem. □

The saddle point result now follows easily.

Theorem 3 (Lowest saddle points)**.** *The first-coalescence points, defined in Theorem 1, are lowest saddle points of $f(z) = \sigma_n(A - zI)$ in the sense of Definition 1.*

Proof. Let $\tilde{z}$ be a first-coalescence point with $\tilde{\epsilon} = f(\tilde{z})$. Suppose there exists a sequence $z_k \in \Lambda_{\tilde{\epsilon}}$ converging to $\tilde{z}$ such that $\lim_{k\to\infty} \nabla f(z_k) = 0$. It follows that $\tilde{z}$ is a stationary point of f according to Definition 1. On the other hand, if no such sequence exists, the final claim of Theorem 2 shows that $\tilde{z}$ is a stationary point of f, again according to Definition 1. In either case, $\tilde{z}$ is in the closure of two pseudospectral components, so it can be neither a local maximum nor a local minimum of f. Hence, $\tilde{z}$ is a saddle point of f. Suppose it is not a lowest saddle point. Then there exists another saddle point y with $f(y) < \tilde{\epsilon}$, and therefore with $\sigma_n(A - yI)$ simple, and so with $\nabla f(y) = -v^*u = 0$, where u and v are corresponding left and right singular vectors of $A - yI$. It follows from Lemma 1 that $A - f(y)uv^*$ has a defective eigenvalue, contradicting the fact that $w(A) = \tilde{\epsilon}$ for the 2-norm. Thus, $\tilde{z}$ is a lowest saddle point of f. □

We conjecture the following.

Conjecture 1. *Let $\tilde{z}$ be a first-coalescence point with $\tilde{\epsilon} = f(\tilde{z})$. If there exists a sequence $z_k \in \Lambda_{\tilde{\epsilon}}$ converging to $\tilde{z}$ with $\lim_{k\to\infty} \nabla f(z_k) = 0$, then $\sigma_n(A - \tilde{z}I)$ has multiplicity one, so $\partial f(\tilde{z}) = \{\nabla f(\tilde{z})\} = \{0\}$. On the other hand, if no such sequence exists, then in addition to the conclusions of Theorem 2, we have that g and $\hat{g}$ are unique (independent of the sequences z_k and $\hat{z}_k$), and that*

$$0 \in \partial f(\tilde{z}) = \big\{\mu g + (1-\mu)\hat{g} \,|\, \mu \in [0,1]\big\}.$$

If this conjecture holds, the multiplicity of the smallest singular value at a first-coalescence point completely determines the geometry of coalescence, with a simple singular value occurring if and only if the saddle point is smooth, and a double singular value occurring if and only if the pseudospectrum contains tangent disks at the coalescence points, with the Clarke generalized gradient consisting of a line segment in $\mathbb{C}$ containing the origin in the latter case.

It is possible that z is a saddle point of f, with a sequence $z_k \in \Lambda_{f(z)}$ with $z_k \to z$ and $\nabla f(z_k) \to 0$, and with $\sigma_n(A - zI)$ having multiplicity two. For example, consider the "reverse diagonal" matrix[5] with entries 1,1,3,2 and with $z = 0$. However, 0 is not a *lowest* saddle point of f, so it is not a first-coalescence point, and hence this example is not a counterexample to Conjecture 1.

Before continuing with our development, we mention two recent papers by Boulton et al. [7] and by Lewis and Pang [17] that address related issues of coalescence of pseudospectral components. There is little overlap between these papers or between either paper and the present paper, except that characterization of coalescence via generalized saddle points is also established in [17, Section 8], using the terminology "resolvent-critical", via variational analysis of semialgebraic functions. Another tool exploited by Lewis and Pang [17] is the convexity of the field of values (numerical range) of a matrix, which we also use in the proof of Theorem 5 below.

3. Analytic paths and orthogonality of singular vectors

Let us now strengthen the assumption of Theorem 2. For sequences $z_k \in \Omega$ and $\hat{z}_k \in \widehat{\Omega}$ that lie on an *analytic path*, we can draw a stronger conclusion about the limiting singular vectors.

Theorem 4 (Analytic path)**.** *Suppose that the assumption of Theorem 2 holds, and suppose that $z_k = z(t_k), \hat{z}_k = z(\hat{t}_k)$, where $z(t) : (0,1) \to \mathbb{C}$ is an analytic path with $z(\tilde{t}) = \tilde{z}$ for exactly one $\tilde{t} \in (0,1)$, with $z(t) \in \Omega$ for $t \in (0,\tilde{t})$, $z(t) \in \widehat{\Omega}$ for $t \in (\tilde{t},1)$, and $z'(\tilde{t}) \neq 0$, and where t_k is a real sequence converging to $\tilde{t}$ from below and $\hat{t}_k$ is a real sequence converging to $\tilde{t}$ from above. As previously, let v and $\hat{v}$, respectively, be limits of v_k, the right singular vectors of $\sigma_n(A - z_kI)$, and $\hat{v}_k$, the right singular vectors of $\sigma_n(A - \hat{z}_kI)$. Then v and $\hat{v}$ are orthogonal, that is $v^*\hat{v} = 0$. Likewise the corresponding limiting left singular vectors $u = (A - \tilde{z}I)v/\tilde{\epsilon}$ and $\hat{u} = (A - \tilde{z}I)\hat{v}/\tilde{\epsilon}$ satisfy $u^*\hat{u} = 0$. Furthermore, there exist $\omega \in \mathbb{C}$ and $\hat{\omega} \in \mathbb{C}$ with $|\omega| = |\hat{\omega}| = 1$ such that*

$$(\omega u)^*(\hat{\omega}\hat{v}) + (\hat{\omega}\hat{u})^*(\omega v) = 0.$$

Proof. We first note that such analytic paths $z(t)$ exist: for example, we could define the path to be the line segment joining the centers of the two tangent disks whose existence was established in Theorem 2. Consider the family of matrices $A - z(t)I$. By [3,15][6] there exists an "analytic SVD", that is, matrices $X(t)$, $\Delta(t)$ and $Y(t)$ which are analytic with respect to t, satisfying

$$A - z(t)I = X(t)\Delta(t)Y(t)^*, \quad X(t)^*X(t) = I, \quad Y(t)^*Y(t) = I,$$

and with $\Delta(t)$ real and diagonal. Thus the absolute values of the diagonal entries of $\Delta(t)$ are the singular values of $A - z(t)I$. We know from Theorem 2 that the multiplicity of the smallest singular

[5] This example was provided by J.-M. Gracia to show the necessity of the corrigendum in [5].

[6] The assumption in [3] that the matrix family is real is not necessary; see [15, Section II.6.1]. What is essential is that the parameter t is real.

value is two at $t = \tilde{t}$, so by permuting the signed singular values and multiplying by a constant diagonal unitary matrix if necessary, we may assume that the following two conditions hold:

1. The last two diagonal entries of $\Delta(\tilde{t})$ are each equal to $\tilde{\epsilon} = \sigma_n(A - \tilde{z}I)$.
2. Let $x, \hat{x}$ and $y, \hat{y}$ be the last two columns of $X(\tilde{t})$ and $Y(\tilde{t})$, respectively. Then $\zeta x^* \hat{y}$ is real, where $\zeta = z'(\tilde{t})$. (This is accomplished, if necessary, by replacing x by $\text{sign}(\zeta x^* \hat{y})x$ and y by $\text{sign}(\zeta x^* \hat{y})y$, where "sign" denotes complex sign.)

Differentiating the decomposition $A - z(t)I = X(t)\Delta(t)Y(t)^*$ with respect to t at $\tilde{t}$, we have

$$-\zeta I = X'(\tilde{t})\Delta(\tilde{t})Y(\tilde{t})^* + X(\tilde{t})\Delta'(\tilde{t})Y(\tilde{t})^* + X(\tilde{t})\Delta(\tilde{t})(Y'(\tilde{t}))^*.$$

Multiplying this equation by $X(\tilde{t})^*$ from the left and by $Y(\tilde{t})$ from the right, we have

$$-\zeta X(\tilde{t})^* Y(\tilde{t}) = X(\tilde{t})^* X'(\tilde{t})\Delta(\tilde{t}) + \Delta'(\tilde{t}) + \Delta(\tilde{t})Y'(\tilde{t})^* Y(\tilde{t}).$$

Setting $R = X(\tilde{t})^* X'(\tilde{t})$ and $S = Y'(\tilde{t})^* Y(\tilde{t})$, and equating the $(n-1, n)$ and $(n, n-1)$ entries of this matrix equation, we have

$$\begin{aligned} -\zeta x^* \hat{y} &= \tilde{\epsilon}(R_{n-1,n} + S_{n-1,n}) \\ -\zeta \hat{x}^* y &= \tilde{\epsilon}(R_{n,n-1} + S_{n,n-1}). \end{aligned}$$

Differentiating $X(t)X(t)^* = I$ and $Y(t)Y(t)^* = I$ at $t = \tilde{t}$ we see that R and S are skew-Hermitian matrices. Therefore the right-hand sides of the two equations above are the negative-complex-conjugates of each other, and hence so are the left hand sides. But the left-hand side of the first equation is real, so both sides of both equations must be real. Adding the equations, we have, since ζ is nonzero, $x^* \hat{y} + \hat{x}^* y = 0$.

Since $\sigma_n(A - z_k I)$ is simple and $z_k = z(t_k)$, the right singular vectors of $A - z_k I$ satisfy $\omega_k v_k = y_m(t_k)$ with $|\omega_k| = 1$, where $y_m(t_k)$ is the mth column of $Y(t_k)$ and, by continuity of $\Delta(t)$ at $\tilde{t}$, m is either $n-1$ or n. Since v_k converges to v, a right singular vector corresponding to $\sigma_n(A - \tilde{z}I)$, we have either $\omega v = y$ or $\omega v = \hat{y}$, where $|\omega| = 1$. Similarly, since the right singular vectors $\hat{v}_k$ of $A - \hat{z}_k I$ converge to $\hat{v}$, we have either $\hat{\omega}\hat{v} = y$ or $\hat{\omega}\hat{v} = \hat{y}$, where $|\hat{\omega}| = 1$. By Theorem 2, v and $\hat{v}$ are linearly independent, so one must be a unit modulus multiple of y and the other of $\hat{y}$. Hence, they are mutually orthogonal, and without loss of generality we can take $\omega v = y$, $\hat{\omega}\hat{v} = \hat{y}$. Thus the left singular vectors satisfy $\omega u = x$ and $\hat{\omega}\hat{u} = \hat{x}$ and are also mutually orthogonal. The final claim follows from the property $x^* \hat{y} + \hat{x}^* y = 0$. □

It follows easily from this result that there *always* exists an orthogonal pair of left and right singular vectors at a first-coalescence point. Recall that, as throughout the paper, when we refer to singular vectors, we mean with unit length in the 2-norm.

Theorem 5 (Orthogonal left and right singular vectors)**.** *Let $\tilde{z}$ be a first-coalescence point with $\tilde{\epsilon} = f(\tilde{z})$. Then there exist left and right singular vectors $\tilde{u}$ and $\tilde{v}$ satisfying $(A - \tilde{z}I)\tilde{v} = \tilde{\epsilon}\tilde{u}$ with $\tilde{u}^*\tilde{v} = 0$.*

Proof. First suppose that the assumption of Theorem 2 does not hold, so that there is a sequence $z_k \in \Lambda_{\tilde{\epsilon}}$ converging to $\tilde{z}$ such that $\lim_{k\to\infty} \nabla f(z_k) = 0$. Then, by selecting a subsequence, we may assume that the sequence of left and right singular vectors u_k and v_k corresponding to $\sigma_n(A - z_k I)$ converge to limits u and v, respectively, and these must be, respectively, left and right singular vectors corresponding to $\sigma_n(A - \tilde{z}I)$. Since $\nabla f(z_k) = -v_k^* u_k \to 0$, we have $\tilde{u}^*\tilde{v} = 0$ by setting $\tilde{u} = u, \tilde{v} = v$.

Suppose now that no such sequence z_k exists, so that Theorems 2 and 4 apply. Then, because of the orthogonality properties $u^*\hat{u} = v^*\hat{v} = 0$ established in Theorem 4, all left and right singular vectors $\tilde{u}$ and $\tilde{v}$ for $\sigma_n(A - \tilde{z}I)$ have the form $\tilde{u} = cu + s\hat{u}$, $\tilde{v} = cv + s\hat{v}$ for some $|c|^2 + |s|^2 = 1$, and hence

$$\tilde{u}^*\tilde{v} = \begin{bmatrix} c \\ s \end{bmatrix}^* [u \;\; \hat{u}]^* [v \;\; \hat{v}] \begin{bmatrix} c \\ s \end{bmatrix}.$$

Thus the set of possible values for the inner products $\tilde{u}^*\tilde{v}$ is exactly the field of values of a fixed 2-by-2 matrix. The field of values is convex [14] and $c = 1, s = 0$ gives u^*v while $c = 0, s = 1$ gives $\hat{u}^*\hat{v}$. By Theorem 2, zero is a convex combination of these two complex numbers. Hence, zero is in the field of values, proving that c, s exist defining $\tilde{u}, \tilde{v}$ with $\tilde{u}^*\tilde{v} = 0$. In fact, using the final result of Theorem 4, we have the explicit formulas $c = \pm\,\omega\sqrt{\mu}$ and $s = \pm\,\hat{\omega}\sqrt{1-\mu}$, where μ is defined by the last part of Theorem 2. □

This result immediately leads to answers to several questions left open in [1]. First, it shows that, for the Frobenius norm as well as for the 2-norm, $w(A) = c(A)$. Second, it shows that as long as A has distinct eigenvalues (as assumed throughout the paper) the minimal distance to the set of defective matrices is *attained*, which was previously known only for the case of a simple smallest singular value.

Theorem 6 (Nearest defective matrix in Frobenius norm). *For the Frobenius norm, $w(A) = c(A)$. Furthermore, for both the 2-norm and the Frobenius norm, the minimal distance to the set of defective matrices is attained.*

Proof. We know from Theorem 1 that $w(A) \geqslant c(A)$. Let $B = A - \tilde{\epsilon}\tilde{u}\tilde{v}^*$ with $\tilde{u}^*\tilde{v} = 0$ as defined in Theorem 5. By Lemma 1, B is defective, and for both norms, $\|A - B\| = \tilde{\epsilon} = c(A) = w(A)$. □

An obvious class of matrices for which $\sigma_n(A - \tilde{z}I)$ has multiplicity two consists of those matrices A satisfying

$$A = Q\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}Q^*,$$

with $Q^*Q = I$ and with the property that the two components Ω and $\widehat{\Omega}$ in whose closure $\tilde{z}$ lies contain one eigenvalue of A_1 and one eigenvalue of A_2, respectively. The normal matrices form a subclass of this class, since normal matrices are unitarily diagonalizable. A nonnormal example was already given in Section 2 and illustrated on the right side of Fig. 1. Intuitively, tangential coalescence at a first-coalescence point $\tilde{z}$, with a double minimum singular value, must apply for this class of matrices since the ϵ-pseudospectra of A are simply the union of the ϵ-pseudospectra of A_1 and of A_2, and these cannot "interact" for $\epsilon < \tilde{\epsilon} = f(\tilde{z})$. We have verified this experimentally for many examples. In the case that A is normal, its ϵ-pseudospectra are simply the union of disks of radius ϵ around its eigenvalues, so tangential coalescence is clear. Assuming then that the assumption of Theorem 2 applies to this class, and taking $Q = I$ for simplicity, each of the right singular vectors v_k is a block vector with its nonzeros corresponding to the block of A whose eigenvalue is in Ω, and therefore its limit v has the same property, as does the limit u of the left singular vectors. Likewise $\hat{v}$ and $\hat{u}$ are block vectors with nonzeros corresponding to the other block. Therefore, $u^*\hat{v} = \hat{u}^*v = 0$. Clearly, this extends to any unitary Q. Notice that the condition $u^*\hat{v} = \hat{u}^*v = 0$ is *stronger* than the property stated at the end of Theorem 4.

Note further that the same condition holds for a broader class of matrices, as illustrated by the following example. Consider a matrix A from the class just described, with $n = 4$ and a unique first-coalescence point $\tilde{z}$ with $\tilde{\epsilon} = f(\tilde{z})$, and set

$$C = A + \delta(u_1 + u_2)(v_1 + v_2)^*$$

where v_1, v_2 are right singular vectors of $A - \tilde{z}I$ corresponding to the *largest* two singular values and u_1, u_2 are the corresponding left singular vectors. Thus, v_1, v_2, v and $\hat{v}$ are all mutually orthonormal, as are u_1, u_2, u and $\hat{u}$. By construction, $(C - \tilde{z}I)v = \tilde{\epsilon}u$ and $(C - \tilde{z}I)\hat{v} = \tilde{\epsilon}\hat{u}$, so $\tilde{z}$ is a first-coalescence point for C as long as $|\delta|$ is sufficiently small, with $u, \hat{u}, v, \hat{v}$ unchanged. It is easily verified experimentally, by choosing A to have randomly generated blocks A_1 and A_2, modified if necessary so that Ω and $\widehat{\Omega}$ contain eigenvalues from different blocks, that constructing C in this way typically results in a matrix that is *not* unitarily block diagonalizable. (For $n = 4$ this is easily verified by computing the Schur factorizations corresponding to all possible orderings of the eigenvalues; in MATLAB, this can be done using the `ordschur` function.)

These observations raise the question:

Question 1. Suppose that the assumption of Theorem 2 holds, or, more restrictively, that the assumption of Theorem 4 holds. Does it follow that the property $u^*\hat{v} = \hat{u}^*v = 0$ must hold?

Notice that if the answer to Question 1 is positive, the formula for $\tilde{u}$ and $\tilde{v}$ in Theorem 5 holds for *any* unit scalars ω and $\hat{\omega}$.

We conclude this section by noting that, if in addition to making the assumption of Theorem 4, we also assume that the right singular vectors for the *second smallest* singular values of $A - z_kI$ and $A - \hat{z}_kI$ converge, then the roles of the limiting singular vectors are reversed in the following sense.

Theorem 7 (Second smallest singular values)**.** *Suppose that the assumption of Theorem 4 holds, and suppose further that the sequences of left and right singular vectors for $\sigma_{n-1}(A - z_kI)$ converge. Then these limiting singular vectors, respectively, have the form $\psi\hat{u}$ and $\psi\hat{v}$, with $|\psi| = 1$. Likewise if the sequences of left and right singular vectors for $\sigma_{n-1}(A - \hat{z}_kI)$ converge, then they have the form $\hat{\psi}u$ and $\hat{\psi}v$, with $|\hat{\psi}| = 1$. Furthermore, the gradients $\nabla\sigma_{n-1}(A - z_kI)$ and $\nabla\sigma_{n-1}(A - \hat{z}_kI)$ converge to $-\hat{v}^*\hat{u}$ and $-v^*u$, respectively. Finally, the convex combinations $\mu\nabla\sigma_n(A - z_kI) + (1 - \mu)\nabla\sigma_{n-1}(A - z_kI)$ and $(1 - \mu)\nabla\sigma_n(A - \hat{z}_kI) + \mu\nabla\sigma_{n-1}(A - \hat{z}_kI)$ both converge to zero, where μ is as in Theorem 2.*

Proof. Let $v, v_k, \hat{v}$ and $\hat{v}_k$ be as in Theorem 4. Note that $v^*\hat{v} = 0$. For k large enough, $\sigma_{n-1}(A - z_kI)$ is simple and hence the corresponding right singular vectors are unique up to unit modulus scalars. Further, for each k these vectors are orthogonal to v_k and hence the limits of these vectors are orthogonal to v. Since the multiplicity of $\sigma_n(A - \tilde{z}I)$ is two and $v^*\hat{v} = 0$, these limits must be unit modulus multiples of $\hat{v}$. Similarly, the limits of right singular vectors corresponding to $\sigma_{n-1}(A - \hat{z}_kI)$ are unit modulus multiples of v. The corresponding results for the left singular vectors follow. The formulas for the gradients are then immediate, and the last statement follows from Theorem 2. □

This theorem will be useful in the formulation of a numerical method in the next section.

4. Numerical approximation of first-coalescence points

We have seen in Section 2 that, in all cases, nearest defective matrices are determined by first-coalescence points, and these are lowest generalized saddle points of $f(z) = \sigma_n(A - zI)$. There are two key issues in the numerical computation of such points : global and local. We will not address the first, which is how to identify a *lowest* saddle point; however, recent work has been done in this direction by Lewis and Pang [18] (based on ideas in [21]) and Mengi [20] (based on ideas in [19]). We will address the second issue: supposing that we have an approximation to a desired saddle point $\tilde{z}$, how do we compute it accurately?

In the generic case that $\sigma_n(A - \tilde{z}I)$ is simple, the obvious answer is to apply Newton's method to the equation $\nabla f(z) = 0$. To avoid confusion we revert to using $\mathbb{R}^2$ instead of $\mathbb{C}$: thus the equation to be solved is $g(x, y) = \nabla f(x, y) = 0$, where $f(x, y) = \sigma_n(A - (x + iy)I)$, and the desired solution is $(\tilde{x}, \tilde{y})$, where $\tilde{z} = \tilde{x} + i\tilde{y}$. We know that $\nabla f(x, y) = -[\Re(v^*u); \Im(v^*u)]$, where u and v are, respectively, left and right singular vectors corresponding to $\sigma_n(A - (x + iy)I)$. To implement Newton's method we also need the Hessian matrix of second derivatives of $f(x, y)$, say $H(x, y)$, whose entries are given in the following lemma. The proof applies more general results in [24] to the parametrization $A - (x + iy)I$.

Lemma 3 (J.-G. Sun)**.** *Suppose that σ is a positive, simple singular value of $A - zI$ with corresponding left and right singular vectors u and v. Define $D = (\sigma^2I - (A - zI)^*(A - zI))^\dagger$ and $E = (\sigma^2I - (A - zI)(A - zI)^*)^\dagger$ where $\dagger$ indicates the Moore–Penrose pseudo-inverse. The second partial derivatives of $\sigma = \sigma(z) = \sigma(x + iy)$ are*

$$
\begin{aligned}
\frac{\partial^2\sigma}{\partial x^2} &= \sigma u^* D u + \sigma v^* E v + 2\Re\left(v^*(A-zI)Du\right) + \sigma^{-1}\left(\Im(u^*v)\right)^2 \\
\frac{\partial^2\sigma}{\partial x \partial y} &= 2\Im(v^*(A-zI)Du) + \sigma^{-1}\Re(u^*v)\Im(u^*v) \\
\frac{\partial^2\sigma}{\partial y^2} &= \sigma u^* D u + \sigma v^* E v - 2\Re\left(v^*(A-zI)Du\right) + \sigma^{-1}\left(\Re(u^*v)\right)^2
\end{aligned}
$$

where $\Re(\cdot)$ *and* $\Im(\cdot)$ *indicate real and imaginary parts, respectively*.

Newton's method applied to $g(x, y) = 0$ is quadratically convergent to $(\tilde{x}, \tilde{y})$ as long as $\sigma_n(A - \tilde{z}I)$ is simple and $H(\tilde{x}, \tilde{y})$ is nonsingular. A practical implementation needs to enforce reduction of some merit function, such as $\|g(x, y)\|_2$. This is easily done by a backtracking line search: if the Newton update $[x; y] - \alpha H(x, y)^{-1} g(x, y)$ with $\alpha = 1$ does not result in reduction in $\|g\|$, replace α by $\alpha/2$, repeating until a reduction is obtained. Let us refer to this algorithm as Method g. Its weakness is that convergence to the desired saddle point $\tilde{z} = \tilde{x} + i\tilde{y}$ may not occur if the initial approximation is not good enough, and in particular, if the singular value separation $\sigma_{n-1}(A - \tilde{z}I) - \sigma_n(A - \tilde{z}I)$ is small, the norm of the Hessian $H(\tilde{x}, \tilde{y})$ is large and the radius of convergence of Newton's method is small.

Now let us consider the case where $\sigma_n(A - \tilde{z}I)$ is double and assume that Theorem 2 applies. We refer to this case henceforth as the nongeneric case, equivalently the case where the coalescence of pseudospectra is tangential. Method g is almost certain to fail in this situation, even when initialized very close to $\tilde{z}$, since not only is ∇f undefined at $\tilde{z}$, but also (by the assumption of the theorem) there is no sequence z_k converging to $\tilde{z}$ with $\nabla f(z_k) \to 0$. However, Theorem 7 suggests applying Newton's method to the following function mapping $\mathbb{R}^3$ to $\mathbb{R}^3$,

$$
h(x, y, \mu) = \begin{bmatrix} \mu \nabla \sigma_{n-1}(A - (x+iy)I) + (1-\mu)\nabla\sigma_n(A - (x+iy)I) \\ \sigma_{n-1}(A - (x+iy)I) - \sigma_n(A - (x+iy)I) \end{bmatrix} = 0.
$$

Note that the definition of h is consistent with the corresponding equation at the end of Theorem 7 that holds on Ω, not the one that holds on $\widehat{\Omega}$. This is an arbitrary choice that must be made one way or the other as in practice one does not know in which component an iterate lies. Inside Ω and $\widehat{\Omega}$, we know that the smallest singular value is simple, and sufficiently close to $\tilde{z}$, the second smallest singular value must also be simple, so the function h is well defined. We may therefore compute its Jacobian (derivative), again using the formulas for second derivatives of simple singular values in Lemma 3. Again we may use Newton's method with a backtracking line search. This time the Newton equation has three variables (the corrections to x, y and μ), but it is natural to carry out the line search in the (x, y) space, defining μ (given x and y) by the least squares approximation to the first two equations in $h(x, y, \mu) = 0$, projecting μ if necessary to lie in [0,1]. We call this Method h.

For sufficiently good starting points, we generally observe quadratic convergence of Method h, but it is not obvious how to state and prove a quadratic convergence theorem, as the singular values are not differentiable at $\tilde{z}$ and hence h is not well defined in the limit. One possibility is to follow the approach in [22], exploiting the existence of an analytic SVD along lines passing through $\tilde{z}$, as was done in the proof of Theorem 4. This allows one to define, for any real $\theta \in [0, \pi)$, a function $h_\theta(t, \mu) = h(\tilde{x} + t\cos\theta, \tilde{y} + t\sin\theta, \mu)$ which is differentiable in (t, μ). Following [22], given any iterate $[x; y; \mu]$ close to the optimal point $[\tilde{x}; \tilde{y}; \tilde{\mu}]$ with $z = x + iy$ in Ω or $\widehat{\Omega}$, one can prove a quadratic contraction in the error provided that the Jacobian of h_θ is invertible at $(0, \tilde{\mu})$ for $\theta = \arg(x - \tilde{x} + i(y - \tilde{y}))$. To some extent this explains quadratic convergence in practice, which typically takes place very rapidly, although to prove a quadratic convergence theorem one would need to know that the inverse of the Jacobian of h_θ is bounded with respect to θ: there seems no reason to suppose this is the case [12]. Another issue is that the iterates $x + iy$ may not remain in Ω and $\widehat{\Omega}$. The function h remains well defined anywhere that the singular values are distinct, but we know nothing about the properties of sets on which this is true outside Ω and $\widehat{\Omega}$. On the other hand, given the tangent disks geometry, the closer the iterates are to $\tilde{z}$, the more likely it seems that they will indeed lie in the pseudospectral components Ω and $\widehat{\Omega}$. In any case, we routinely observe quadratic convergence for Method h in the nongeneric case, as we do for the method to be described next.

Clearly, Method h will fail for generic matrices for which the smallest singular value $\sigma_n(A-\tilde{z}I)$ is simple. We therefore introduce a third method which combines the advantages of Methods g and h. Consider the following function mapping $\mathbb{R}^3$ to $\mathbb{R}^3$,

$$k(x,y,\mu)=\begin{bmatrix}\mu\nabla\sigma_{n-1}(A-(x+iy)I)+(1-\mu)\nabla\sigma_n(A-(x+iy)I)\\ \mu(\sigma_{n-1}(A-(x+iy)I)-\sigma_n(A-(x+iy)I))\end{bmatrix}=0.$$

The only difference between the functions k and h is that the third component of k has a multiplicative factor μ. This formulation exploits the well-known concept of *complementarity* familiar from constrained optimization: for $k(x,y,\mu)$ to be zero, either μ is zero, in which case the first two equations imply that $g(x,y)=0$, or μ is not zero, in which case $h(x,y,\mu)=0$. The Jacobian of k is readily computed (again assuming distinct singular values), and the resulting backtracking implementation of Newton's method is very similar to the one outlined for the function h. We call this Method k. The beauty of Method k is that it is applicable to *both* the generic (multiplicity one) and nongeneric (multiplicity two) cases.

To complete the description of these three local methods, we need to define the starting point and the termination criteria. Addressing the latter point first, we terminate the iterations when either the norm of the function g, h or k drops below 10^{-15} (a very demanding tolerance) or the line search fails to return a reduction in the norm, indicating that the limiting accuracy has been reached. Regarding the starting point, a good choice is as follows. Let p_j be the eigenvalue condition number for the eigenvalue λ_j of A (recall that p_j^{-1} is the modulus of the inner product of normalized left and right eigenvectors for λ_j). We define the starting point $z_0=x_0+iy_0$ by the following convex combination of two eigenvalues,

$$z_0=\frac{p_{\hat{j}}\lambda_{\hat{k}}+p_{\hat{k}}\lambda_{\hat{j}}}{p_{\hat{j}}+p_{\hat{k}}},$$

where the index pair $(\hat{j},\hat{k})$ minimizes $|\lambda_j-\lambda_k|/(p_j+p_k)$ over all distinct pairs of eigenvalues. The starting guess is derived from first-order perturbation bounds for simple eigenvalues. If A is perturbed to $A+E$ and $\|E\|=\eta$ is small, then for an eigenvalue λ_j of A there exists an eigenvalue $\tilde{\lambda}_j$ of $A+E$ such that

$$|\tilde{\lambda}_j-\lambda_j|\leqslant p_j\eta+O(\eta^2).$$

Thus for sufficiently small η, the component of Λ_η containing λ_j is approximately a disk of radius $p_j\eta$ centered at λ_j. Consequently, a point of coalescence of two components of Λ_η containing eigenvalues, say, λ_j and λ_k is expected to be approximated by the point of coalescence of the disks $|z-\lambda_j|<p_j\eta$ and $|z-\lambda_k|<p_k\eta$. This gives the starting guess z_0 which is the point of coalescence of the disks for the eigenvalues $\lambda_{\hat{j}}$ and $\lambda_{\hat{k}}$.

We now illustrate the behavior of the three methods on some examples, using the standard MATLAB function `svd` to compute the singular value decomposition. We start with the same family illustrated in Section 2,

$$A(\delta)=\begin{bmatrix}0.25 & 10 & 0 & \delta\\ 0 & i & 0 & 0\\ 0 & 0 & 0.5 & 10\\ 0 & 0 & 0 & 1+i\end{bmatrix}.$$

We ran Methods g, h and k on this matrix family using the following values for δ: 10^{-15}, 10^{-12}, 10^{-9}, 10^{-6}, 10^{-3} and 1. Mathematically speaking, $A(\delta)$ has generic, that is nontangential, coalescence of pseudospectra at the first-coalescence point $\tilde{z}$ for almost all $\delta>0$, with $\sigma_n(A-\tilde{z}I)$ simple, but numerically speaking, setting δ sufficiently small is effectively the same as setting it to zero, in which case $\sigma_n(A-\tilde{z}I)$ is double. Thus, it is not surprising that Method g fails to converge to the correct saddle point for $\delta=10^{-15}$. What might be surprising is that because the convergence radius is so small when

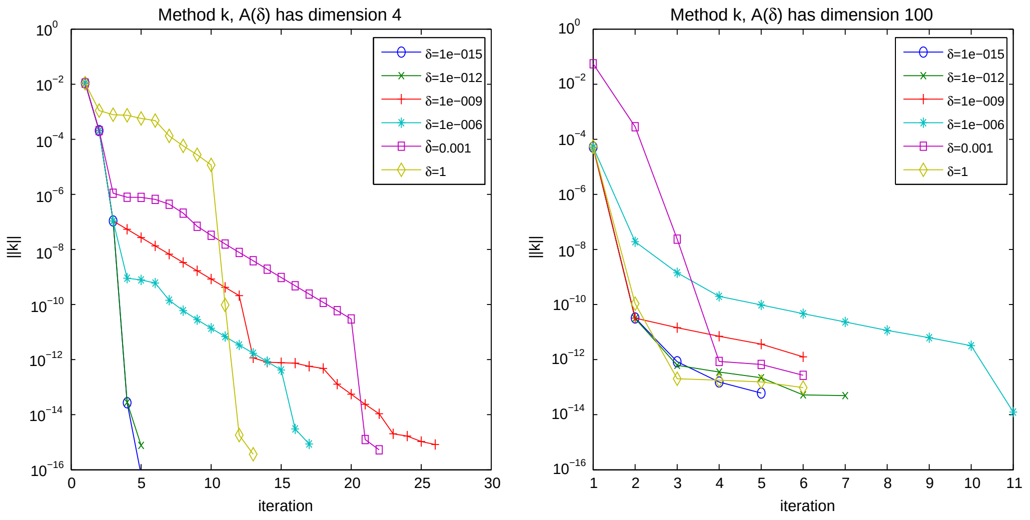

Fig. 2. The 2-norm of $k(x, y, \mu)$ as a function of the iteration count, for two matrix families $A(\delta)$ which are block diagonal when $\delta = 0$. On the left, the $n = 4$ family. On the right, the $n = 100$ family.

the singular value separation is small, Method g fails for much larger values of δ as well, including all values listed above except $\delta = 1$, *converging to a saddle point that is not the lowest saddle point*. On the other hand, Method h works successfully for $\delta = 10^{-15}$, but increasingly more poorly as δ increases. In contrast, Method k works well in *all* cases, as seen on the left side of Fig. 2, which plots $\|k\|_2$ as a function of the iteration count for each value of δ. Note the rapid, indeed quadratic, convergence for $\delta = 10^{-15}$. For the middle-sized values of δ, convergence is slower; this is because the condition number of the Jacobian of the function k is enormous. A large condition number delays convergence of Method k even for $\delta = 1$, but in that case, quadratic contractions in $\|k\|_2$ eventually occur at iterations 11 and 12. In every case convergence took place to the *lowest* saddle point.

We also ran the same experiments on a block diagonal family with $n = 100$, constructed as follows. First we randomly generated two blocks A_1 and A_2, each of order 50, and then we computed the pair $(\hat{j}, \hat{k})$ minimizing $|\lambda_j - \nu_k|$, where λ_j is an eigenvalue of A_1 and ν_k is an eigenvalue of A_2. We then set $A_0 = \mathrm{diag}(A_1, A_2 + \tau I)$, where $\tau = \lambda_{\hat{j}} - \nu_{\hat{k}} + (1 + i)/(10n)$. In this way, as we verify empirically, the smallest of the distances between all pairs of eigenvalues of A_0 is likely to be attained by one eigenvalue from each block of A_0, and, likewise, the initial point z_0 defined above is likely to be the weighted average of $\lambda_{\hat{j}}$ and $\lambda_{\hat{j}} + (1 + i)/(10n)$. We then define $A(\delta)$ to be $A_0 + \delta e_1 e_n^*$, that is the block diagonal matrix A_0 with an additional entry δ connecting the blocks in the top right corner. We then ran all three methods on this matrix for the same values of δ used previously. The right side of Fig. 2 shows the results for Method k on a typical example; for each value of δ, convergence was obtained to the same saddle point. On the other hand, Method g achieves comparable accuracy only for $\delta = 1$, and diverges to a higher saddle point for $\delta = 10^{-9}$, 10^{-12} and 10^{-15}, while Method h achieves accuracy comparable to Method k *only* for these three smallest values of δ. Because n is larger, rounding errors prevent the machine precision accuracy that we observed for $n = 4$, but quadratic contractions are still observed for the smallest and largest values of δ, while for the middle values, good accuracy is obtained but convergence is slower because the Jacobian of k is very badly conditioned.

The conclusion to be drawn from this discussion is the following. Nongeneric cases (with tangential coalescence) do not present any special difficulty as long as Method k is used, and this method works well for generic cases too. Indeed, we observe quadratic convergence in both cases. The *difficult* cases are the matrices that are *close* to nongeneric instances, and even in these cases, Method k works well, although convergence is markedly delayed by ill-conditioning of the Jacobian of the function $k(x, y, \mu)$. In contrast, in the nearly nongeneric cases (small δ), Methods g and h both fail consistently. This is because the function $g(x, y)$ is so ill-conditioned that Newton's method does not converge without an extremely good starting point, and the third component of $h(x, y, \mu)$ cannot be set exactly to zero. Thus, in summary, Method k is an effective way to stabilize Method g in the most difficult ill-conditioned cases, extending the advantage of Method h for the nongeneric case to nearly nongeneric cases.

We should acknowledge, however, that Method k is not foolproof. It is not difficult to construct small, highly ill-conditioned examples for which convergence takes place to a nonlowest saddle point or even to a local maximum of $\sigma_n(A - zI)$. In such cases, the procedure of the next section constructs a nearby defective matrix, but not the nearest one.

5. Numerical construction of the nearest defective matrix

In this section, we discuss how to construct the nearest defective matrix B to A, assuming that a lowest saddle point $\tilde{z}$ has been accurately approximated by Method k as discussed in the previous section. In order to avoid overly cluttered notation, we adopt the following conventions in this section: $\tilde{z}$ denotes the *computed* saddle point at the final iteration of Method k (after the final successful line search); $\tilde{\epsilon}$ denotes the computed smallest singular value of $A - \tilde{z}I$; u and v denote the computed left and right singular vectors corresponding to the *second smallest* computed singular value of $A - \tilde{z}I$; $\hat{u}$ and $\hat{v}$ denote the computed left and right singular vectors corresponding to the *smallest* singular value of $A - \tilde{z}I$, and μ denotes its value at the final iterate of Method k. (The reason for setting u, v, rather than $\hat{u}, \hat{v}$, to the singular vectors for the second smallest singular value is to make the usage of μ in the definition of k consistent with the usage of μ in Theorem 2.)

The first idea that comes to mind is as follows: if μ is zero (or small), set $B = A - \tilde{\epsilon}\hat{u}\hat{v}^*$; otherwise, set $B = A - \tilde{\epsilon}\tilde{u}\tilde{v}^*$ where $\tilde{u}$ and $\tilde{v}$ are defined as in Theorem 5. However, this is a problematic for several reasons. The first is that, as explained in the previous section, numerically speaking one cannot distinguish between generic and nongeneric cases, and in the cases where the smallest two singular values are not well separated it will be difficult to decide whether μ should be zero or not. The second is that the scalars ω and $\hat{\omega}$ for which the final conclusion of Theorem 4 holds are not immediately available, although any unit scalars will work if the answer to Question 1 is positive. The third reason is the most subtle. In the nongeneric case with $\tilde{z}$ exact, it does not make sense to speak of individual singular vectors corresponding to the smallest two singular values, but only of the two-dimensional subspace of singular vectors for the double singular value. Therefore, in the nongeneric case where $\tilde{z}$ is an accurate but not exact estimate of the saddle point, or the very nearly nongeneric case, the singular vectors corresponding to the smallest two singular values cannot be resolved numerically [23]. One can expect only that the subspace spanned by the computed right singular vectors corresponding to the smallest two singular values is an accurate approximation of the span of the singular vectors corresponding to the actual smallest two singular values. The same is true of the left singular vectors. Thus, the computed vectors $u, \hat{u}, v, \hat{v}$ are not likely to accurately approximate the limiting vectors defined in the proof of Theorems 2 and 4. Consequently, even if the answer to Question 1 is positive, the formula for $\tilde{u}$ and $\tilde{v}$ given in Theorem 5 may be completely wrong.

A partial remedy is to terminate Method k before $\tilde{z}$ is approximated too accurately, so that the individual singular vectors can be resolved numerically. However, a better idea is as follows. Note that the final vectors $u, \hat{u}, v, \hat{v}$ *must* satisfy the key properties $(A - \tilde{z}I)v = \tilde{\epsilon}u$, $(A - \tilde{z}I)\hat{v} = \tilde{\epsilon}\hat{u}$ and the orthogonality properties $u^*\hat{u} = v^*\hat{v} = 0$ to high precision, from standard error analysis for computing the SVD. Therefore, the desired singular vectors $\tilde{u}$ and $\tilde{v}$ for which $\tilde{u}^*\tilde{v} = 0$ must satisfy $\tilde{u} = cu + s\hat{u}$, $\tilde{v} = cv + s\hat{v}$ to high precision, for some $|c|^2 + |s|^2 = 1$. Define the 2×2 matrix

$$W = [u \ \ \hat{u}]^* \, [v \ \ \hat{v}] \, .$$

We wish to compute complex numbers c and s with $|c|^2 + |s|^2 = 1$ satisfying

$$\begin{bmatrix} c \\ s \end{bmatrix}^* W \begin{bmatrix} c \\ s \end{bmatrix} = 0.$$

Note that, letting w_{jk} denote the entries of W, the equation $\mu w_{11} + (1 - \mu)w_{22} = 0$ is valid to high precision because it is enforced by Method k.

Table 1 The computed residual for the two matrix families $A(\delta)$ which are block diagonal when $\delta = 0$. The second and fourth columns show the residual computed using only the singular vectors corresponding to the smallest singular value. The third and fifth columns show the residual computed using the singular vectors corresponding to the smallest two singular values, via Algorithm 1. The first residual is smaller when δ is small and the second is smaller when δ is large, motivating the use of Algorithm 2 to construct the nearest defective matrix.

δ	$r(\tilde{u}, \tilde{v})$ $(n = 4)$	$r(\hat{u}, \hat{v})$ $(n = 4)$	$r(\tilde{u}, \tilde{v})$ $(n = 100)$	$r(\hat{u}, \hat{v})$ $(n = 100)$
1.0e-015	5.5e-015	1.2e-001	1.2e-013	4.6e-001
1.0e-012	3.2e-015	5.1e-002	4.3e-014	2.5e-001
1.0e-009	2.5e-012	7.9e-006	1.2e-011	4.2e-005
1.0e-006	2.5e-009	8.3e-009	1.2e-008	1.9e-007
1.0e-003	2.5e-006	4.6e-012	1.2e-005	9.6e-011
1.0e+000	3.5e-014	3.5e-014	1.2e-002	2.1e-013

The following algorithm computes c and s, given the computed W.

Algorithm 1

1. *If $w_{11} = 0$, return $c = 1, s = 0$; if $w_{22} = 0$, return $c = 0, s = 1$.*
2. *If $w_{12} + w_{21} = 0$, return $c = \sqrt{\mu}, s = \pm\sqrt{1-\mu}$.*
3. *Set $\psi = w_{jj}/|w_{jj}|$, where $|w_{jj}| = \max(|w_{11}|, |w_{22}|)$, and replace W by $\bar{\psi}W$. This does not change the solution set for c, s, but it makes w_{jj} real. Furthermore, since the ratio w_{11}/w_{22} is the real number $(\mu - 1)/\mu$ before the scaling by ψ, this remains true after the scaling, and therefore both scaled diagonal entries are real.*
4. *Set $\xi = (\alpha - i\beta)/\sqrt{\alpha^2 + \beta^2}$, where $\alpha = \Re(w_{12} - w_{21})$ and $\beta = \Im(w_{12} + w_{21})$, and replace W by D^*WD, where $D = \operatorname{diag}([1, \xi])$. The new W satisfies $\Im(w_{11}) = \Im(w_{22}) = \Im(w_{12} + w_{21}) = 0$, and hence $[\gamma; \delta]^*\Im(W)[\gamma; \delta] = 0$ for any real γ, δ.*
5. *We now need to find real γ, δ with $\gamma^2 + \delta^2 = 1$ satisfying $[\gamma; \delta]^* Re(W)[\gamma; \delta] = 0$. This is achieved by $\gamma = 1/\sqrt{1+t^2}, \delta = t/\sqrt{1+t^2}$, where t is a solution of the quadratic equation $w_{11} + \Re(w_{12} + w_{21})t + w_{22}t^2 = 0$.*
6. *Return $[c; s] = D[\gamma; \delta] = [\gamma; \delta\xi]$.*

Let us define a "singular vector residual" mapping $\mathbb{C}^n \times \mathbb{C}^n$ to $\mathbb{R}$ by

$$r(p, q) = |p^*q| + \|(A - \tilde{z}I)q - \tilde{\epsilon}p\| + \|(A - \tilde{z}I)^*p - \tilde{\epsilon}q\|.$$

If $r(p, q) = 0$, then $B = A - \tilde{\epsilon}pq^*$ is defective by Lemma 1. Table 1 compares $r(\tilde{u}, \tilde{v}) = r(cu + s\hat{u}, cv + s\hat{v})$ (where c, s are computed by Algorithm 1) with $r(\hat{u}, \hat{v})$ (using only the computed singular vectors corresponding to the smallest singular value) for the families $A(\delta)$ from Section 4. We see that, as expected, for nearly nongeneric matrices (δ small), $r(\tilde{u}, \tilde{v})$ may be much smaller than $r(\hat{u}, \hat{v})$. On the other hand, for matrices that are *not* nearly nongeneric (δ large), $r(\tilde{u}, \tilde{v})$ may actually be larger than $r(\hat{u}, \hat{v})$. We therefore advocate the following algorithm for constructing the nearest defective matrix. Recall that $u, v, \hat{u}, \hat{v}$ were all defined at the beginning of this section.

Algorithm 2

1. *Set $\tilde{u} = cu + s\hat{u}, \tilde{v} = cv + s\hat{v}$, where c, s are computed by Algorithm 1.*
2. *If $r(\tilde{u}, \tilde{v}) \leqslant r(\hat{u}, \hat{v})$, set $B = A - \tilde{\epsilon}\tilde{u}\tilde{v}^*$; otherwise, set $B = A - \tilde{\epsilon}\hat{u}\hat{v}^*$.*

The following simple backward error analysis tells us that if $r(p, q)$ is small, B is close to a defective matrix.

Theorem 8 (Backwards error)**.** *For sufficiently small positive $\eta = r(p, q)$,*

$$B = A - \tilde{\epsilon}pq^* = SCS^{-1} + E,$$

where C is defective, $\|S^*S - I\| = O(\eta)$, and $\|E\| = O(\eta)$.

Proof. Let $S = [q\ p\ Y]$, where $Y \in \mathbb{C}^{n\times(n-2)}$ has orthonormal columns and $Y^*q = Y^*p = 0$. Since the first two columns of S are not exactly orthogonal, we may use Gram-Schmidt to orthogonalize S, resulting in $S = Q + be_2^T$, where Q is unitary, $b = p - d/\|d\|$ and $d = p - (q^*p)q$. Note that $\|d\|^2 = 1 - \theta^2$, where $\theta = |q^*p| \leqslant \eta$. Thus $\|b\| = \|p - d(1 + \theta^2/2 + O(\theta^4))\| = \theta + O(\theta^2)$. It follows that $S = Q + O(\theta), S^*S = I + O(\theta)$, and $S^{-1} = S^* + O(\theta)$. But immediately from the definition of S, we have $S^*BS = C + E$, where $\|E\| = O(\eta)$ and

$$C = \begin{bmatrix} \tilde{z} & * & * & \cdots & * \\ 0 & \tilde{z} & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix}.$$

The eigenvalue $\tilde{z}$ of C has right eigenvector e_1 and left eigenvector e_2, so it has algebraic multiplicity at least two and geometric multiplicity at least one. Furthermore, we know from Lemma 1 that for $\eta = 0$, the geometric multiplicity is one, so it is clear that when η is sufficiently small, the geometric multiplicity is one, and hence C is defective. □

Our MATLAB code implementing the methods and algorithms discussed in this paper is publicly available.[7]

6. The angle of pseudospectral wedges

Section 2 is primarily concerned with the local geometry of the pseudospectrum $\Lambda_{w(A)}$ near first-coalescence points $\tilde{z}$ when $\sigma_n(A - \tilde{z}I)$ has multiplicity two, as illustrated in the example shown on the right side of Fig. 1. The following, perhaps surprising, result addresses this issue in the generic case, that is when $\sigma_n(A - \tilde{z}I)$ is simple, as illustrated on the left side of Fig. 1. This theorem applies to *any* smooth saddle point of f, not only lowest saddle points, so we state it in that generality.

Theorem 9 (Angle of wedges)**.** *Suppose $\tilde{z}$ satisfies $\sigma_{n-1}(A - \tilde{z}I) > \sigma_n(A - \tilde{z}I) = \epsilon$, with corresponding left and right singular vectors u and v satisfying $u^*v = 0$, and so, by virtue of Lemma 1, $A - \epsilon uv^*$ has an eigenvalue $\tilde{z}$ with geometric multiplicity one and algebraic multiplicity at least two. Suppose that the algebraic multiplicity is two. Then, for all $\omega < \pi/2$, Λ_ϵ contains 2 wedges of angle at least ω emanating from $\tilde{z}$, that is*

$$\tilde{z} \pm \left\{\rho e^{i\theta/2} \,\middle|\, \theta \in [\tilde{\theta} - \omega, \tilde{\theta} + \omega], \rho \leqslant \tilde{\rho}\right\},$$

for some real $\tilde{\theta}$ and positive real $\tilde{\rho}$.

Proof. We have $A - \epsilon uv^* = MJM^{-1}$ where, without loss of generality, v is the first column of M, αu^* is the second row of M^{-1} for some nonzero α, and the leading block in the Jordan form J is[8]

$$\begin{bmatrix} \tilde{z} & 1 \\ 0 & \tilde{z} \end{bmatrix}.$$

[7] www.cs.nyu.edu/overton/software/neardefmat/.

[8] The second column of M and first row of M^{-1} are not unique, even up to multiples, but the scalar α is unique given the decision to normalize the first column of M to equal v.

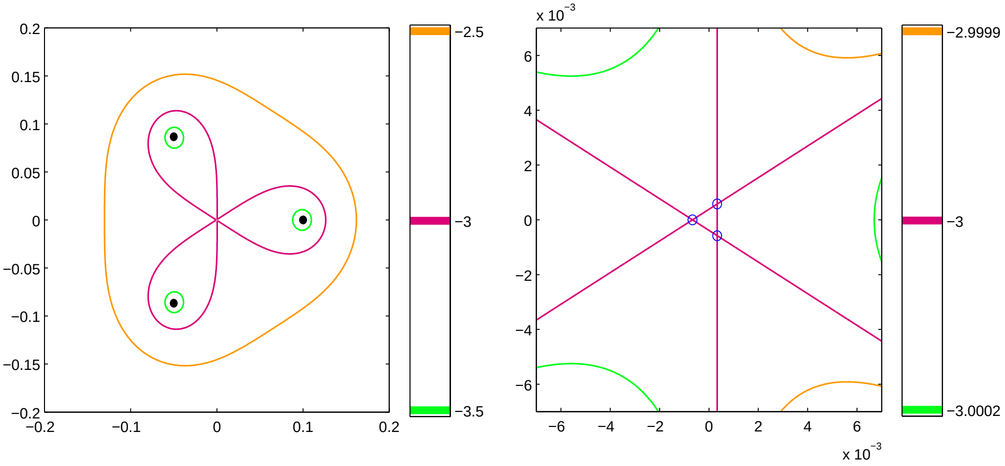

Fig. 3. Pseudospectra of an upper Jordan block of dimension 3 perturbed by setting the bottom left entry to 10^{-3}. Left: a first glance at the pseudospectra indicates that the nearest defective matrix is the Jordan block itself, with 3 wedges of angle $\pi/3$ meeting at zero. Right: a closer look indicates that there are 3 first-coalescence points (marked by circles), each corresponding to a nearest defective matrix with a double eigenvalue.

Let y^* be the first row of M^{-1} and let x be the second column of M. Thus, $y^*v = \alpha u^*x = 1$ and $y^*x = 0$. Define $E(\delta) = \epsilon uv^* - \delta xy^*$. Then $M^{-1}(A - E(\delta))M$ is block diagonal with leading block

$$\begin{bmatrix} \tilde{z} & 1 \\ \delta & \tilde{z} \end{bmatrix},$$

which has eigenvalues $\tilde{z} \pm \sqrt{\delta}$. On the other hand, for small complex numbers δ, we have

$$\|E(\delta)\|_2 = \epsilon - \Re(u^*(\delta xy^*)v) + o(\delta) = \epsilon - \Re(\delta/\alpha) + o(\delta).$$

Thus, $\|E(\delta)\|_2 < \epsilon$ when $\Re(\delta/\alpha) > 0$ and δ is sufficiently small. It follows that, for such δ, the eigenvalues of $A - E(\delta)$ that are close to $\tilde{z}$, namely $\tilde{z} \pm \sqrt{\delta}$, lie in the pseudospectrum Λ_ϵ. The result thus holds, with $\tilde{\theta}$ the complex argument of α and $\tilde{\rho}$ sufficiently small. □

This theorem says that, at coalescence points $\tilde{z}$ corresponding to geometric multiplicity one and algebraic multiplicity two, the angle of the wedges in the ϵ-pseudospectrum emanating from $\tilde{z}$ cannot be less than $\pi/2$. We may think of the tangent disks geometry described by Theorem 2 as being the limit case for a sequence of matrices for which the angle of the wedges is arbitrarily close to π.

Theorem 9 is easily extended to the case that the algebraic multiplicity is $m > 2$, stating that for all $\omega < \pi/m$, the pseudospectrum Λ_ϵ contains m wedges of angle at least ω emanating from $\tilde{z}$. Such geometry seems quite counterintuitive and we are almost certain that this case cannot occur, but in the absence of a proof it cannot yet be ruled out. As a final example, let A be an upper Jordan block of dimension three with the perturbation 10^{-3} in the bottom left corner. Clearly, $w(A) \leqslant 10^{-3}$. The left side of Fig. 3 shows a pseudospectral plot which seems, at first glance, to indicate that $w(A) = 10^{-3}$, with the nearest defective matrix having a triple eigenvalue and with $\Lambda_{w(A)}$ containing three wedges of angle arbitrarily close to $\pi/3$ meeting at a first-coalescence point. In fact, however, $w(A) = 0.99999985 \times 10^{-3}$, and a closer look at the pseudospectra on the right side of Fig. 3 indicates the presence of three first-coalescence points $\tilde{z}$, each corresponding to a nearest defective matrix with a double eigenvalue. Furthermore, for each such $\tilde{z}$, the pseudospectrum $\Lambda_{w(A)}$ contains two wedges of angle arbitrarily close to $2\pi/3$ emanating from $\tilde{z}$, which is consistent with Theorem 9.

7. Conclusions

We have presented several new results on the geometry of pseudospectra near coalescence points, specifically Thereoms 2 and 9 for the nongeneric and generic cases, respectively, and we have established that $w(A)$, the distance to the nearest defective matrix, equals $c(A)$, the smallest singular value of $A - zI$ at a first-coalescence point, for the Frobenius norm as it is for the 2-norm. We also showed that the distance to the nearest defective matrix is always attained when A has distinct eigenvalues, and that first-coalescence points are lowest generalized saddle points of $f(z) = \sigma_n(A - zI)$. We presented a local method to approximate a lowest saddle point that is applicable to generic, nongeneric and the (especially difficult) nearly nongeneric cases, and explained how to avoid rounding pitfalls in constructing the nearest defective matrix. We also presented a simple backward error analysis, allowing one to conclude that if a certain residual is small, a computed approximation to the nearest defective matrix is close to an exactly defective matrix.

We believe that these ideas will, if combined with a global technique to find the *lowest* saddle point, finally result in a reliable algorithm to accurately compute the nearest defective matrix. On the theoretical side, several interesting points remain open, particularly Conjecture 1 and Question 1.

Acknowledgments

This work began in 2004 when all four authors were visiting the Technische Universität Berlin, partially supported by the Deutsche Forschungsgemeinschaft Research Center, Mathematics for Key Technologies. This support, as well as the enthusiasm of Volker Mehrmann, is gratefully acknowledged. Work continued during Ralph Byers' visits to India in 2005 and New York in 2006. Sadly, Ralph died on December 15, 2007, but he will long be remembered for his outstanding research in the theory and practice of numerical linear algebra as well as his modesty, humor and kindness.

References

[1] R. Alam, S. Bora, On sensitivity of eigenvalues and eigendecompositions of matrices,Linear Algebra Appl. 396 (2005) 273–301.
[2] R. Alam, Wilkinson's problem revisited, J. Anal. 4 (2006) 176–205.
[3] A. Bunse-Gerstner, R. Byers, V. Mehrmann, N.K. Nichols, Numerical computation of an analytic singular value decomposition of a matrix valued function, Numer. Math. 60 (1) (1991) 1–39.
[4] J.M. Borwein, A.S. Lewis, Convex Analysis and Nonlinear Optimization: Theory and Examples, Springer, New York, 2000.
[5] J.V. Burke, A.S. Lewis, M.L. Overton, Optimization and pseudospectra, with applications to robust stability, SIAM J. Matrix Anal. Appl. 25 (2003) 80–104, Corrigendum: <http://www.cs.nyu.edu/overton/papers/pseudo_corrigendum.html>.
[6] J.V. Burke, A.S. Lewis, M.L. Overton, Spectral conditioning and pseudospectral growth, Numer. Math. 107 (2007) 27–37.
[7] L. Boulton, P. Lancaster, P. Psarrakos, On pseudospectra of matrix polynomials and their boundaries, Math. Comput. 77 (2007) 313–334.
[8] F.H. Clarke, Optimization and Nonsmooth Analysis, John Wiley, New York, 1983. Reprinted by SIAM, Philadelphia, 1990.
[9] J.W. Demmel, A Numerical Analyst's Jordan Canonical Form, Ph.D. thesis, University of California at Berkeley, 1983.
[10] J.W. Demmel, Computing stable eigendecompositions of matrices, Linear Algebra Appl. 79 (1986) 163–193.
[11] J.W. Demmel, Nearest defective matrices and the geometry of ill-conditioning, Reliab. Numer. Comput., Oxford Univ. Press, 1990, pp. 35–55.
[12] S. Friedland, J. Nocedal, M.L. Overton, Four quadratically convergent methods for solving inverse eigenvalue problems, in: D.F. Griffiths (Ed.), Numerical Analysis, Pitman Research Note in Mathematics, vol. 140, John Wiley, New York, 1986, pp. 47–65.
[13] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, UK, 1985.
[14] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.
[15] T. Kato, A Short Introduction to Perturbation Theory for Linear Operators, Springer-Verlag, New York, 1982.
[16] R.A. Lippert, A. Edelman, The computation and sensitivity of double eigenvalues, in: Z. Chen, Y. Li, C.A. Micchelli, Y. Xu (Eds.), Advances in Computational Mathematics: Proceedings of the Guangzhou International Symposium, Dekker, New York, 1999, pp. 353–393.
[17] A.S. Lewis, C.H.J. Pang, Variational analysis of pseudospectra, SIAM J. Optim. 19 (2008) 1048–1072.
[18] A.S. Lewis, C.H.J. Pang, Private communication, 2009.
[19] A.N. Malyshev, A formula for the 2-norm distance from a matrix to the set of matrices with multiple eigenvalues, Numer. Math. 83 (3) (1999) 443–454.
[20] E. Mengi, Private communication, 2009.
[21] J.J. Moré, T.S. Munson, Computing mountain passes and transition states, Math. Program. 100 (2004) 151–182.
[22] J. Nocedal, M.L. Overton, Numerical methods for solving inverse eigenvalue problems, in: V. Pereyra, A. Reinoza (Eds.), Lecture Notes in Mathematics, vol. 1005, Springer-Verlag, New York, 1983.

[23] G.W. Stewart, Matrix Algorithms. II: Eigensystems, SIAM, 2001.
[24] J.-G. Sun, A note on simple nonzero singular values, J. Comput. Math. 6 (3) (1988) 258–266.
[25] L.N. Trefethen, M. Embree, Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators, Princeton University Press, 2005.
[26] J.M. Varah, On the separation of two matrices, SIAM J. Numer. Anal. 16 (1979) 216–222.
[27] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.
[28] J.H. Wilkinson, On neighbouring matrices with quadratic elementary divisors, Numer. Math. 44 (1984) 1–21.
[29] J.H. Wilkinson, Sensitivity of eigenvalues, Util. Math. 25 (1984) 5–76.
[30] J.H. Wilkinson, Sensitivity of Eigenvalues. II, Util. Math. 30 (1986) 243–286.
[31] T.G. Wright, EigTool: A Graphical Tool for Nonsymmetric Eigenproblems, Oxford University Computer Laboratory, 2002. Available from: <http://web.comlab.ox.ac.uk/pseudospectra/eigtool/>.