A Newton's Iteration Converges Quadratically to Nonisolated Solutions Too

Zhonggang Zeng *

October 25, 2019

Abstract

An extension of Newton's iteration maintains its local quadratic convergence to nonisolated solutions of nonlinear systems assuming the solutions are regular as properly defined. Even if the given system is perturbed and the nonisolated solution disappears, the iteration still locally converges to a stationary point that approximates a nonisolated solution of the underlying system with an error in the same order of the data accuracy. Furthermore, this paper provides a geometric interpretation of the convergence tendency, elaborates the modeling and applications involving nonisolated solutions and demonstrates a software implementation with computing examples.

1 Introduction

Newton's iteration as we know it loses its quadratic rate of convergence, if it applies and converges at all, to *nonisolated* solutions of nonlinear systems of equations with large errors in numerical computation. A subtle tweak of its formulation restores the fast rate of convergence and the optimal accuracy as we shall elaborate in this paper.

Perhaps there is no need to explain or even mention the importance of Newton's iteration in scientific computing, applied mathematics and numerical analysis. In its most common and well-known formulation, Newton's iteration (see, e.g. [21] and most textbooks in numerical analysis)

$$\mathbf{x}_{k+1} \;=\; \mathbf{x}_k - J(\mathbf{x}_k)^{-1}\,\mathbf{f}(\mathbf{x}_k) \quad \text{for} \quad k = 0, 1, \ldots \tag{1}$$

is the standard method for solving systems of nonlinear equations in the form of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ where $J(\mathbf{x})$ is the Jacobian of the mapping $\mathbf{f}$ at $\mathbf{x}$. It is well

*Department of Mathematics, Northeastern Illinois University, Chicago, Illinois 60625, USA. email: zzeng@neiu.edu. Research is supported in part by NSF under grant DMS-1620337.

documented that Newton's iteration quadratically converges to any isolated solution under natural conditions: The mapping is smooth and the initial iterate is near a solution at which the Jacobian is invertible.

Solving systems of nonlinear equations is a standard topic in textbooks of numerical analysis but the discussions have always been limited to isolated solutions. Models with nonisolated solutions frequently arise in scientific computing as we shall elaborate in §8 with case studies. However, the version (1) is formulated under the assumption that the Jacobian is invertible at the solution and not intended for nonisolated solutions at which the inverse of the Jacobian is either undefined or nonexistent. Even if it converges, Newton's iteration (1) is known to approach nonisolated solutions slowly at linear rate with an attainable accuracy being limited and often dismal. To circumvent those difficulties, scientific computing practitioners go to great lengths to isolate solutions with auxiliary equations and variables. There have been attempts and results in the literature extending Newton's iteration on nonisolated solutions directly. Those works appears to be under disseminated, scarcely applied and needing further development. Filling the analytical and algorithmic gap is long overdue in direct computation of nonisolated solutions of nonlinear systems.

In this paper, we formulate a notion of *regular* nonisolated solutions, establish an extension of Newton's iteration for such solutions and prove its local quadratic convergence on exact equations along with local linear convergence on perturbed equations with empirical data. Furthermore, we provide a geometric interpretation of the convergence tendency, elaborate the modeling and applications involving nonisolated solutions and demonstrate our software implementation with computing examples.

An isolated zero is regular if the Jacobian is invertible, namely the nullity of the Jacobian and the dimension of the zero are both zero. Nonisolated zeros of a smooth mapping can form a smooth submanifold of a positive dimension such as curves and surfaces. We generalize the regularity of isolated zeros to nonisolated cases as the dimension being identical to the nullity of the Jacobian at the zero. Regular zeros of a positive dimension form branches with locally invariant dimensions (c.f. Lemma 3) and, near a regular zero, we prove a crucial property that every stationary point is a regular zero (c.f. Lemma 4).

We extended Newton's iteration to the form of

$$\mathbf{x}_{k+1} \quad = \quad \mathbf{x}_k - J_{\text{rank-}r}(\mathbf{x}_k)^\dagger \, \mathbf{f}(\mathbf{x}_k) \quad \text{for} \quad k = 0, 1, \ldots \tag{2}$$

for a smooth mapping $\mathbf{f}$ from an open domain in $\mathbb{R}^m$ or $\mathbb{C}^m$ to $\mathbb{R}^n$ or $\mathbb{C}^n$ where $J_{\text{rank-}r}(\mathbf{x}_k)^\dagger$ is the Moore-Penrose inverse of $J_{\text{rank-}r}(\mathbf{x}_k)$ that is the rank-r projection of the Jacobian $J(\mathbf{x}_k)$. We establish its local quadratic convergence (Theorem 1) under minimal natural assumptions: The mapping $\mathbf{f}$ is smooth and the initial iterate is near a regular zero at which the Jacobian is of rank r.

Nonisolated solutions can be highly sensitive to data perturbations. When the system of equations is perturbed, represented with empirical data or solved using floating

point arithmetic, the nonisolated solution can be significantly altered or even disappears altogether. We prove that the proposed Newton's iteration still converges to a stationary point that approximates an exact solution of the underlying system with an accuracy in the same order of the data error (c.f. Theorem 2). In other words, the proposed extension of Newton's iteration also serves as a regularization mechanism for such an ill-posed zero-finding problem. A condition number can also be derived from Theorem 2 for nonisolated zeros with respect to data perturbations.

As a geometric interpretation, we shall illustrate the behavior of the Newton's iteration (2) for converging to roughly the point on the solution manifold nearest to the initial iterate. We also elaborate case studies on mathematical modeling with nonisolated solutions, demonstrate our software implementation in solving for such solutions with step-by-step calling sequences and computing results.

Extending Newton's iteration beyond (1) by replacing the inverse with a certain kind of generalized inverse traces back to Gauss for solving least squares solutions of overdetermined systems by the so-called Gauss-Newton iteration with an assumption that the Jacobian is injective. For systems with rank-deficient Jacobians, Ben-Israel [3] is the first to propose using Moore-Penrose inverses to generalize Newton's iteration as

$$\mathbf{x}_{k+1} \;=\; \mathbf{x}_k - J(\mathbf{x}_k)^\dagger\, \mathbf{f}(\mathbf{x}_k) \quad \text{for} \quad k \,=\, 0, 1, \ldots \tag{3}$$

"but the conditions for the [convergence] theorem are somewhat restrictive and unnatural" [4]. Chu is the first to prove the local convergence of Newton's iteration (3) with essentially minimal assumptions for underdetermined systems with surjective Jacobians [6]. Applying the alpha theory, Dedieu and Kim [9] prove Newton's iteration (3) locally quadratically converges to a nonisolated solution under the assumption that the Jacobian has a constant deficient rank in a neighborhood of the initial iterate.

Nashed and Chen [20] extend Newton's iteration further as

$$\mathbf{x}_{k+1} \;=\; \mathbf{x}_k - J(\mathbf{x}_k)^\#\, \mathbf{f}(\mathbf{x}_k) \quad \text{for} \quad k \,=\, 0, 1, \ldots \tag{4}$$

where, for any matrix A, the notation $A^\#$ stands for one of the outer inverses of A satisfying the identity $A^\# A A^\# \equiv A^\#$, and prove its local quadratic convergence to a stationary point $\hat{\mathbf{x}}$ at which $J(\mathbf{x}_0)^\# \mathbf{f}(\hat{\mathbf{x}}) = \mathbf{0}$ using a specific outer inverse

$$J(\mathbf{x}_k)^\# \;=\; \big(I + J_{\text{rank-}r}(\mathbf{x}_0)^\dagger \big(J(\mathbf{x}_k) - J(\mathbf{x}_0)\big)\big)^{-1} J_{\text{rank-}r}(\mathbf{x}_0)^\dagger \tag{5}$$

for a proper r but it is unknown whether $\hat{\mathbf{x}}$ is a zero of $\mathbf{f}$. Chen, Nashed and Qi [5] along with Levin and Ben-Israel [17] follow up with similar convergence results toward stationary points using outer inverses.

In comparison to those pioneer works, our extension (2) is suitable for any rank of the Jacobian at the zero. Furthermore, our convergence theorems require minimal assumptions and the iteration quadratically converges to a stationary point that is

guarenteed to be a zero if the initial iterate is near a regular zero. The iteration also permits data perturbations and floating point arithmetic while still converges to an approximate zero at linear rate.

We loosely refer to our extension (2) as the *rank-r Newton's iteration* or simply *Newton's iteration* following a long-standing practice. The terminology is actually debatable as it is fair to ask "Is Newton's method really Newton's method?" [10], and the term may even be considered "an enduring myth" [15]. Heavily influenced by François Viète, Isaac Newton's original method is not even an iteration and can be considered a special case of Joseph Raphson's later formulation restricted to univariate polynomial equations. The version (1) of Newton's iteration can deservedly be credit to Thomas Simpson as well. As a myth or not, however, the term "Newton's iteration" has been used for all extensions of Newton's original method and will likely to used as a convention in the future.

2 Preliminaries

Column vectors are denoted by boldface lower case letters such as $\mathbf{b}$, $\mathbf{x}$, $\mathbf{y}$ etc. with $\mathbf{0}$ being a zero vector whose dimension can be derived from the context. The vector spaces of n-dimensional real and complex column vectors are denoted by $\mathbb{R}^n$ and $\mathbb{C}^n$ respectively. The vector space of $m \times n$ complex matrices including real matrices is denoted by $\mathbb{C}^{m\times n}$. Matrices are denoted by upper case letters such as A, B, X, etc. with O and I denoting a zero matrix and an identity matrix respectively. The range, kernel, rank and Hermitian transpose of a matrix A are denoted by $\mathcal{Range}(A)$, $\mathcal{Kernel}(A)$, $\mathit{rank}(A)$ and A^{H} respectively. For any matrix A, its *Moore-Penrose inverse* [12, §5.5.2, p. 290] $A^\dagger$ is the unique matrix satisfying

$$A\,A^\dagger\,A \;=\; A,\;\; A^\dagger\,A\,A^\dagger \;=\; A^\dagger,\;\; (A\,A^\dagger)^{\mathsf{H}} \;=\; A\,A^\dagger,\;\; (A^\dagger\,A)^{\mathsf{H}} \;=\; A^\dagger\,A. \tag{6}$$

The j-th largest singular value of A is denoted by $\sigma_j(A)$. Let $U\,\Sigma\,V^{\mathsf{H}}$ be the singular value decomposition of A where $U = [\mathbf{u}_1, \cdots, \mathbf{u}_m]$ and $V = [\mathbf{v}_1, \cdots, \mathbf{v}_n]$ are unitary matrices formed by the left singular vectors and the right singular vectors respectively. The *rank-r projection* $A_{\text{rank-}r}$ of A, also known as rank-r truncated singular value decomposition (TSVD) and rank-r approximation of A, is defined as

$$A_{\text{rank-}r} \;\;:=\;\; \sigma_1(A)\,\mathbf{u}_1\,\mathbf{v}_1^{\mathsf{H}} + \cdots + \sigma_r(A)\,\mathbf{u}_r\,\mathbf{v}_r^{\mathsf{H}}$$

Using singular values and singular vectors, the identity [12, §5.5.2]

$$A^\dagger \;\equiv\; \sum_{\sigma_j(A)>0} \frac{1}{\sigma_j(A)}\,\mathbf{v}_j\,\mathbf{u}_j^{\mathsf{H}}$$

holds and it is straightforward to verify

$$A\,A^\dagger_{\text{rank-}r} \;=\; A_{\text{rank-}r}\,A^\dagger_{\text{rank-}r} \;=\; [\mathbf{u}_1, \cdots, \mathbf{u}_r]\,[\mathbf{u}_1, \cdots, \mathbf{u}_r]^{\mathsf{H}} \tag{7}$$

$$A^\dagger_{\text{rank-}r}\,A \;=\; A^\dagger_{\text{rank-}r}\,A_{\text{rank-}r} \;=\; [\mathbf{v}_1, \cdots, \mathbf{v}_r]\,[\mathbf{v}_1, \cdots, \mathbf{v}_r]^{\mathsf{H}} \tag{8}$$

that are orthogonal projections onto the subspaces spanned by $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ respectively. For any matrix A, denote $A^\dagger_{\text{rank-}r}$ as $(A_{\text{rank-}r})^\dagger$.

We say $\mathbf{f}$ is a *smooth mapping* if $\mathbf{f} : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ has continuous derivatives of second order, or $\mathbf{f} : \Omega \subset \mathbb{C}^n \to \mathbb{C}^m$ is holomorphic, where the domain Ω is an open subset of $\mathbb{C}^n$ or $\mathbb{R}^n$. We may designate a variable name, say $\mathbf{x}$, for $\mathbf{f}$ and denote $\mathbf{f}$ as $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$. In that case the Jacobian of $\mathbf{f}$ at any particular $\mathbf{x}_0 \in \Omega$ is denoted by $\mathbf{f}_\mathbf{x}(\mathbf{x}_0)$ or $J(\mathbf{x}_0)$ while $J_{\text{rank-}r}(\mathbf{x}_0)$ or equivalently $\mathbf{f}_\mathbf{x}(\mathbf{x}_0)_{\text{rank-}r}$ is its rank-r projection. For a smooth mapping $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y})$ at $(\mathbf{x}_0, \mathbf{y}_0)$, the notation $\mathbf{f}_{\mathbf{xy}}(\mathbf{x}_0, \mathbf{y}_0)$ represents its Jacobian (with respect to *both* $\mathbf{x}$ and $\mathbf{y}$) while $\mathbf{f}_\mathbf{x}(\mathbf{x}_0, \mathbf{y}_0)$ and $\mathbf{f}_\mathbf{y}(\mathbf{x}_0, \mathbf{y}_0)$ denote the (partial) Jacobians with respect to $\mathbf{x}$ and $\mathbf{y}$ respectively at $(\mathbf{x}_0, \mathbf{y}_0)$.

Lemma 1 *Let $J(\mathbf{x})$ be the Jacobian of a smooth mapping $\mathbf{f}$ at any $\mathbf{x}$ in its open domain Ω in $\mathbb{C}^n$ or $\mathbb{R}^n$. Assume $\mathit{rank}(J(\mathbf{x}_*)) = r$ at $\mathbf{x}_* \in \Omega$. Then there is an open bounded convex subset $\Omega_* \ni \mathbf{x}_*$ of Ω and constants $\zeta, \mu, \eta > 0$ such that, for every $\mathbf{x}, \mathbf{y} \in \Omega_*$, the inequality $\mathit{rank}(J(\mathbf{x})) \geq r$ holds along with*

$$\begin{aligned}
\left\| J_{\mathit{rank\text{-}r}}(\mathbf{x})\, J_{\mathit{rank\text{-}r}}(\mathbf{x})^\dagger - J_{\mathit{rank\text{-}r}}(\mathbf{y})\, J_{\mathit{rank\text{-}r}}(\mathbf{y})^\dagger \right\|_2 &\leq \zeta \|\mathbf{x} - \mathbf{y}\|_2 && (9)\\
\left\| J_{\mathit{rank\text{-}r}}(\mathbf{x}) - J_{\mathit{rank\text{-}r}}(\mathbf{y}) \right\|_2 &\leq \mu \|\mathbf{x} - \mathbf{y}\|_2 && (10)\\
\left\| J_{\mathit{rank\text{-}r}}(\mathbf{x})^\dagger - J_{\mathit{rank\text{-}r}}(\mathbf{y})^\dagger \right\|_2 &\leq \eta \|\mathbf{x} - \mathbf{y}\|_2. && (11)
\end{aligned}$$

Proof. From $\mathit{rank}(J(\mathbf{x}_*)) = r$, we have $\sigma_r(J(\mathbf{x}_*)) > \sigma_{r+1}(J(\mathbf{x}_*)) = 0$. Weyl's Theorem [25, Corollary 4.31, p. 69] ensures the singular values to be continuous with respect to the matrix entries. By the continuity of $J(\mathbf{x})$ with respect to $\mathbf{x}$ in Ω, there is an open bounded convex neighborhood Ω_* of $\mathbf{x}_*$ with $\overline{\Omega}_* \subset \Omega$ such that $\sigma_r(J(\mathbf{x})) > 2\,\sigma_{r+1}(J(\mathbf{x}))$ for every $\mathbf{x} \in \Omega_*$. We can further assume Ω_* to be sufficiently small so that

$$\left\| J(\mathbf{x}) - J(\mathbf{y}) \right\|_2 < \tfrac{1}{2}\big(\sigma_r(J(\mathbf{x})) - \sigma_{r+1}(J(\mathbf{x}))\big) \tag{12}$$

for all $\mathbf{x}, \mathbf{y} \in \Omega_*$ and

$$\max_{\mathbf{x} \in \overline{\Omega}_*} \left\| J_{\text{rank-}r}(\mathbf{x})^\dagger \right\|_2 = \max_{\mathbf{x} \in \overline{\Omega}_*} \frac{1}{\sigma_r(J(\mathbf{x}))} \leq 2 \left\| J(\mathbf{x}_*)^\dagger \right\|_2. \tag{13}$$

The left-hand side of (9) is the distance between the subspaces spanned by the first r left singular vectors of $J(\mathbf{x})$ and $J(\mathbf{y})$ respectively and, by Wedin's Theorem [29] (also see [26, Theorem 4.4]) and (12), there is a constant $\zeta > 0$ such that

$$\begin{aligned}
\left\| J_{\text{rank-}r}(\mathbf{x})\, J_{\text{rank-}r}(\mathbf{x})^\dagger - J_{\text{rank-}r}(\mathbf{y})\, J_{\text{rank-}r}(\mathbf{y})^\dagger \right\|_2 &\leq 4 \left\| J_{\text{rank-}r}(\mathbf{x})^\dagger \right\|_2 \left\| J(\mathbf{x}) - J(\mathbf{y}) \right\|_2 \\
&\leq \zeta \|\mathbf{x} - \mathbf{y}\|_2
\end{aligned}$$

for all $\mathbf{x}, \mathbf{y} \in \Omega_*$. As a result, the inequality (10) follows from (9) and

$$\begin{aligned}
&\left\| J_{\text{rank-}r}(\mathbf{x}) - J_{\text{rank-}r}(\mathbf{y}) \right\|_2 \\
&\quad = \left\| J_{\text{rank-}r}(\mathbf{x})\, J_{\text{rank-}r}(\mathbf{x})^\dagger\, J(\mathbf{x}) - J_{\text{rank-}r}(\mathbf{y})\, J_{\text{rank-}r}(\mathbf{y})^\dagger\, J(\mathbf{y}) \right\|_2 \quad \text{(by (6) and (8))} \\
&\quad \leq \left\| J_{\text{rank-}r}(\mathbf{x})\, J_{\text{rank-}r}(\mathbf{x})^\dagger \right\|_2 \|J(\mathbf{x}) - J(\mathbf{y})\|_2 \\
&\qquad + \left\| J_{\text{rank-}r}(\mathbf{x})\, J_{\text{rank-}r}(\mathbf{x})^\dagger - J_{\text{rank-}r}(\mathbf{y})\, J_{\text{rank-}r}(\mathbf{y})^\dagger \right\|_2 \|J(\mathbf{y})\|_2.
\end{aligned}$$

since $\left\|J_{\text{rank-}r}(\mathbf{x})\, J_{\text{rank-}r}(\mathbf{x})^\dagger\right\|_2 = 1$ and $\|J(\mathbf{y})\|_2$ is bounded on the compact set $\overline{\Omega}_*$. By [24, Theorem 3.3], there is a constant $\alpha > 0$ such that

$$\left\|J_{\text{rank-}r}(\mathbf{x}) - J_{\text{rank-}r}(\mathbf{y})\right\|_2 \;\leq\; \alpha \left\|J_{\text{rank-}r}(\mathbf{x})^\dagger\right\|_2 \left\|J_{\text{rank-}r}(\mathbf{y})^\dagger\right\|_2 \left\|J_{\text{rank-}r}(\mathbf{x}) - J_{\text{rank-}r}(\mathbf{y})\right\|_2$$

leading to (11). □

Lemma 2 *For $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, let $\mathbf{z} \mapsto \phi(\mathbf{z})$ be a continuous injective mapping from an open set Ω in $\mathbb{F}^n$ to $\mathbb{F}^m$. At any $\mathbf{z}_0 \in \Omega$, there is an open neighborhood Δ of $\phi(\mathbf{z}_0)$ in $\mathbb{F}^m$ such that, for every $\mathbf{b} \in \Delta$, there exists a $\mathbf{z}_\mathbf{b}$ in Ω and an open neighborhood Σ_0 of $\mathbf{z}_\mathbf{b}$ with*

$$\|\mathbf{b} - \phi(\mathbf{z}_\mathbf{b})\|_2 \;=\; \min_{\mathbf{z}\in\Sigma_0} \|\mathbf{b} - \phi(\mathbf{z})\|_2. \tag{14}$$

Further assume ϕ is differentiable in Ω. Then

$$\phi_\mathbf{z}(\mathbf{z}_\mathbf{b})^\dagger \left(\mathbf{b} - \phi(\mathbf{z}_\mathbf{b})\right) \,=\, \mathbf{0}. \tag{15}$$

Proof. Let Σ_0 be an open bounded neighborhood of $\mathbf{z}_0$ such than $\overline{\Sigma}_0 \subset \Omega$. Since ϕ is one-to-one and continuous, we have

$$\delta \;=\; \min_{\mathbf{z}\in\overline{\Sigma}_0\setminus\Sigma_0} \|\phi(\mathbf{z}) - \phi(\mathbf{z}_0)\|_2 \;>\; 0.$$

Let $\Delta = \left\{\mathbf{y} \in \mathbb{F}^m \;\middle|\; \|\mathbf{y} - \phi(\mathbf{z}_0)\|_2 < \frac{1}{2}\,\delta\right\}$. Then, for every $\mathbf{b} \in \Delta$, there exists a $\mathbf{z}_\mathbf{b} \in \overline{\Sigma}_0$ such that $\|\phi(\mathbf{z}_\mathbf{b}) - \mathbf{b}\|_2 = \min_{\mathbf{z}\in\overline{\Sigma}_0} \|\phi(\mathbf{z}) - \mathbf{b}\|_2$. For every $\mathbf{z} \in \overline{\Sigma}_0 \setminus \Sigma_0$, however,

$$\begin{aligned} \|\phi(\mathbf{z}) - \mathbf{b}\|_2 \;&\geq\; \|\phi(\mathbf{z}) - \phi(\mathbf{z}_0)\|_2 - \|\phi(\mathbf{z}_0) - \mathbf{b}\|_2 \\ &>\; \tfrac{1}{2}\,\delta \;>\; \|\phi(\mathbf{z}_0) - \mathbf{b}\|_2 \;\geq\; \|\phi(\mathbf{z}_\mathbf{b}) - \mathbf{b}\|_2 \end{aligned}$$

implying $\mathbf{z}_\mathbf{b} \in \Sigma_0$ and (14). Since a local minimum of

$$\|\phi(\mathbf{z}) - \mathbf{b}\|_2^2 \;=\; (\phi(\mathbf{z}) - \mathbf{b})^\mathsf{H}\,(\phi(\mathbf{z}) - \mathbf{b})$$

occurs at the interior point $\mathbf{z}_\mathbf{b} \in \Sigma_0$, it is straightforward to verify the equation $\phi_\mathbf{z}(\mathbf{z}_\mathbf{b})^\mathsf{H}\,(\phi(\mathbf{z}) - \mathbf{b}) = \mathbf{0}$ and thus (15) from $\mathit{Range}(\phi_\mathbf{z}(\mathbf{z}_\mathbf{b})^\mathsf{H}) = \mathit{Range}(\phi_\mathbf{z}(\mathbf{z}_\mathbf{b})^\dagger)$. □

3 Regular zeros of smooth mappings

A point $\mathbf{x}_*$ is a *zero* of a mapping $\mathbf{f}$ if $\mathbf{x} = \mathbf{x}_*$ is a *solution* of the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. A common notation $\mathbf{f}^{-1}(\mathbf{0})$ stands for the set of all zeros of $\mathbf{f}$. A zero $\mathbf{x}_*$ of $\mathbf{f}$ is *isolated* if there is an open neighborhood Λ of $\mathbf{x}_*$ in the domain of $\mathbf{f}$ such that $\mathbf{f}^{-1}(\mathbf{0}) \cap \Lambda = \{\mathbf{x}_*\}$ or $\mathbf{x}_*$ is *nonisolated* otherwise. A nonisolated zero $\mathbf{x}_*$ of a smooth mapping $\mathbf{f}$ may belong to a curve, a surface or a higher dimensional subset of $\mathbf{f}^{-1}(\mathbf{0})$. We adopt a simple definition of the dimension of a zero as follows. For more in-depth elaboration on the dimension of zero sets, see [2, p. 17].

Definition 1 (Dimension of a Zero) *For* $\mathbb{F} = \mathbb{C}$ *or* $\mathbb{R}$, *let* $\mathbf{x}_*$ *be a zero of a smooth mapping* $\mathbf{f} : \Omega \subset \mathbb{F}^m \to \mathbb{F}^n$. *If there is an open neighborhood* $\Delta \subset \Omega$ *of* $\mathbf{x}_*$ *in* $\mathbb{F}^m$ *such that* $\Delta \cap \mathbf{f}^{-1}(\mathbf{0}) = \phi(\Lambda)$ *where* $\mathbf{z} \mapsto \phi(\mathbf{z})$ *is a differentiable mapping defined in a connected open set* Λ *in* $\mathbb{F}^k$ *for a certain* $k > 0$ *with* $\phi(\mathbf{z}_*) = \mathbf{x}_*$ *and* $\mathit{rank}(\,\phi_{\mathbf{z}}(\mathbf{z}_*)\,) = k$, *then the* dimension *of* $\mathbf{x}_*$ *as a zero of* $\mathbf{f}$ *is defined as*

$$\mathit{dim}_{\mathbf{f}}(\mathbf{x}_*) \;:=\; \mathit{dim}(\mathcal{R}\mathit{ange}(\phi_{\mathbf{z}}(\mathbf{z}_*))) \;\equiv\; \mathit{rank}(\,\phi_{\mathbf{z}}(\mathbf{z}_*)\,) \;=\; k.$$

As a special case, the dimension of an isolated zero is zero.

A so-defined k-dimensional zero $\mathbf{x}_*$ is on a smooth submanifold of dimension k in $\mathbb{F}^m$. If the dimension $\mathit{dim}_{\mathbf{f}}(\mathbf{x}_*)$ is well-defined, then $\phi_{\mathbf{z}}(\mathbf{z}_*)$ in Definition 1 is of full column rank and there is an open neighborhood $\Lambda_* \subset \Lambda$ of $\mathbf{z}_*$ such that $\mathit{rank}(\,\phi_{\mathbf{z}}(\hat{\mathbf{z}})\,) \equiv k$ for all $\hat{\mathbf{z}} \in \Lambda_*$. Namely *the dimension of a zero is locally invariant*. We shall also say every $\mathbf{x} \in \phi(\Lambda_*)$ is *in the same branch* of zeros as $\mathbf{x}_*$ and, if a zero $\tilde{\mathbf{x}}$ is in the same branch of $\mathbf{x}_*$, every zero $\hat{\mathbf{x}}$ in the same branch of $\tilde{\mathbf{x}}$ is in the same branch of $\mathbf{x}_*$.

An isolated zero $\mathbf{x}_*$ of $\mathbf{f}$ is regular if its dimension 0 is identical to the nullity of the Jacobian $\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*)$. This notion of regularity can naturally be generalized to zeros of higher dimensions in the definition below. There are tremendous advantages in computing regular zeros as we shall elaborate throughout this paper.

Definition 2 (Regular Zero) *A zero* $\mathbf{x}_*$ *of a smooth mapping* $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$ *is* regular *if* $\mathit{dim}_{\mathbf{f}}(\mathbf{x}_*)$ *is well-defined and identical to* $\mathit{nullity}(\,\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*)\,)$. *Namely*

$$\mathit{dim}_{\mathbf{f}}(\mathbf{x}_*) + \mathit{rank}(\,\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*)\,) \;=\; \textit{the dimension of the domain of } \mathbf{f}. \tag{16}$$

A zero is ultrasingular *if it is not regular.*

A system of equations $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ is said to be *underdetermined* if $\mathbf{f}$ is a mapping from $\Omega \subset \mathbb{F}^m$ to $\mathbb{F}^n$ with $m > n$ where $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. A solution of an underdetermined system is always regular if the Jacobian is surjective or, equivalently, of full row rank. For instance, let $(\mathbf{u}_*, \mathbf{v}_*)$ be a zero of a smooth mapping $(\mathbf{u}, \mathbf{v}) \mapsto \mathbf{f}(\mathbf{u}, \mathbf{v})$ from $\mathbb{R}^k \times \mathbb{R}^m$ to $\mathbb{R}^m$ and the partial Jacobian $\mathbf{f}_{\mathbf{v}}(\mathbf{u}_*, \mathbf{v}_*)$ is invertible. By the Implicit Mapping Theorem, there is a differentiable mapping $\mathbf{u} \mapsto \mathbf{g}(\mathbf{u})$ from a neighborhood Λ of $\mathbf{u}_*$ in $\mathbb{R}^k$ to $\mathbb{R}^m$ with $\mathbf{g}(\mathbf{u}_*) = \mathbf{v}_*$ and there is an open neighborhood Δ of $(\mathbf{u}_*, \mathbf{v}_*)$ in $\mathbb{R}^k \times \mathbb{R}^m$ such that $\Delta \cap \mathbf{f}^{-1}(\mathbf{0}) = \phi(\Lambda)$ where $\phi(\mathbf{u}) = (\mathbf{u}, \mathbf{g}(\mathbf{u}))$ for $\mathbf{u} \in \Lambda$. Furthermore, the Jacobian $\phi_{\mathbf{u}}(\mathbf{u}_*)$ is obviously of full column rank k. As a result, the dimension of the zero $(\mathbf{u}_*, \mathbf{v}_*)$ is k that is identical to the nullity of Jacobian of $\mathbf{f}$ at $(\mathbf{u}_*, \mathbf{v}_*)$, implying $(\mathbf{u}_*, \mathbf{v}_*)$ is regular.

Ultrasingular zeros can be isolated multiple zeros [7], isolated ultrasingularity embedded in nonisolated zero set (c.f. Example 8) or can form an entire branch of zeros (c.f. Example 9). Like the dimension, regularity is also invariant on a branch of nonisolated zeros as asserted in the following lemma.

Lemma 3 (Local Invariance of Regularity) *Let* $\mathbf{x}_*$ *be a regular zero of a smooth mapping* $\mathbf{f}$. *Then there is an open neighborhood* Δ_* *of* $\mathbf{x}_*$ *such that every* $\hat{\mathbf{x}} \in \Delta_* \cap \mathbf{f}^{-1}(\mathbf{0})$ *is a regular zero of* $\mathbf{f}$ *in the same branch of* $\mathbf{x}_*$.

Proof. The assertion is obviously true for isolated regular zeros. Let $\mathbb{F}$ be either $\mathbb{C}$ or $\mathbb{R}$ and the domain of $\mathbf{f}$ is an open subset in $\mathbb{F}^n$. Assume $\mathbf{x}_*$ is a regular k-dimensional zero of $\mathbf{f}$. There is an open neighborhoods Δ of $\mathbf{x}_*$ in $\mathbb{F}^n$ and there is an open connected set Λ_1 of a certain $\mathbf{z}_*$ in $\mathbb{F}^k$ along with a differentiable mapping $\phi : \Lambda_1 \to \mathbb{F}^n$ such that $\phi(\mathbf{z}_*) = \mathbf{x}_*$, $\Delta \cap \mathbf{f}^{-1}(\mathbf{0}) = \phi(\Lambda_1)$, $\mathit{rank}(\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*)) = n - k$ and $\mathit{rank}(\phi_{\mathbf{z}}(\mathbf{z}_*)) = k$. By the continuity of singular values we can assume $\mathit{rank}(\phi_{\mathbf{z}}(\mathbf{z})) \equiv k$ for all $\mathbf{z} \in \Lambda_1$. From $\mathbf{f}(\phi(\mathbf{z})) \equiv \mathbf{0}$ and $\mathbf{f}_{\mathbf{x}}(\phi(\mathbf{z}))\,\phi_{\mathbf{z}}(\mathbf{z}) \equiv O$ for all $\mathbf{z} \in \Lambda_1$, we have $\mathit{nullity}(\mathbf{f}_{\mathbf{x}}(\phi(\mathbf{z}))) \geq k$ for all $\mathbf{z} \in \Lambda_1$. By the continuity of singular values again, there is an open neighborhood $\Delta_* \subset \Delta$ of $\mathbf{x}_*$ such that $\mathit{rank}(\mathbf{f}_{\mathbf{x}}(\mathbf{x})) \geq n - k$ for all $\mathbf{x} \in \Delta_*$. Consequently, every $\hat{\mathbf{x}} \in \Lambda \cap \mathbf{f}^{-1}(\mathbf{0})$ is a regular zero where $\Lambda = \phi^{-1}(\Delta_*)$. □

We shall propose a new version of Newton's iteration that, under proper conditions, converges to a *stationary point* $\hat{\mathbf{x}}$ at which $J_{\text{rank-}r}(\hat{\mathbf{x}})^{\dagger}\,\mathbf{f}(\hat{\mathbf{x}}) = \mathbf{0}$. The following stationary point property of regular zeros ensures that, in a neighborhood of a regular zero, all stationary points are regular zeros in the same branch.

Lemma 4 (Stationary Point Property) *Let* $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$ *be a smooth mapping with a regular zero* $\mathbf{x}_*$ *and* $r = \mathit{rank}(\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*))$. *Then there is an open neighborhood* Ω_* *of* $\mathbf{x}_*$ *such that, for any* $\hat{\mathbf{x}} \in \Omega_*$, *the equality* $\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}})^{\dagger}_{\mathit{rank\text{-}r}}\,\mathbf{f}(\hat{\mathbf{x}}) = \mathbf{0}$ *holds if and only if* $\hat{\mathbf{x}}$ *is a regular zero of* $\mathbf{f}$ *in the same branch of* $\mathbf{x}_*$.

Proof. We first prove there is a neighborhood Ω_1 of $\mathbf{x}_*$ such that $\mathbf{f}(\hat{\mathbf{x}}) = \mathbf{0}$ for every $\hat{\mathbf{x}} \in \Omega_1$ with $\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}})^{\dagger}_{\text{rank-}r}\,\mathbf{f}(\hat{\mathbf{x}}) = \mathbf{0}$. Assume the assertion is false, namely there is a sequence $\{\mathbf{x}_j\}_{j=1}^{\infty}$ converging to $\mathbf{x}_*$ such that $\mathbf{f}_{\mathbf{x}}(\mathbf{x}_j)^{\dagger}_{\text{rank-}r}\,\mathbf{f}(\mathbf{x}_j) = \mathbf{0}$ but $\mathbf{f}(\mathbf{x}_j) \neq \mathbf{0}$ for all $j = 1, 2, \ldots$. Let $\mathbf{z} \mapsto \phi(\mathbf{z})$ be the parameterization of the solution branch containing $\mathbf{x}_*$ as in Definition 1 with $\phi(\mathbf{z}_*) = \mathbf{x}_*$. Since $\phi_{\mathbf{z}}(\mathbf{z}_*)$ is injective, there is a neighborhood of $\mathbf{z}_*$ in which ϕ is one-to-one by the Inverse Mapping Theorem. From Lemma 2 with any sufficiently large j, there is a $\check{\mathbf{x}}_j \in \Omega_* \cap \mathbf{f}^{-1}(\mathbf{0}) = \phi(\Delta)$ such that

$$\|\mathbf{x}_j - \check{\mathbf{x}}_j\|_2 \;=\; \min_{\mathbf{z}\in\Delta} \|\mathbf{x}_j - \phi(\mathbf{z})\|_2 \;=\; \|\mathbf{x}_j - \phi(\mathbf{z}_j)\|_2 \tag{17}$$

at a certain $\mathbf{z}_j$ with $\phi(\mathbf{z}_j) = \check{\mathbf{x}}_j$, implying

$$\phi_{\mathbf{z}}(\mathbf{z}_j)\,\phi_{\mathbf{z}}(\mathbf{z}_j)^{\dagger}\,\frac{\mathbf{x}_j - \phi(\mathbf{z}_j)}{\|\mathbf{x}_j - \phi(\mathbf{z}_j)\|_2} \;=\; \frac{\phi_{\mathbf{z}}(\mathbf{z}_j)}{\|\mathbf{x}_j - \phi(\mathbf{z}_j)\|_2}\Big(\phi_{\mathbf{z}}(\mathbf{z}_j)^{\dagger}\,\big(\mathbf{x}_j - \phi(\mathbf{z}_j)\big)\Big) \;=\; \mathbf{0}. \tag{18}$$

We claim $\check{\mathbf{x}}_j \to \mathbf{x}_*$ as well when $j \to \infty$. Assume otherwise. Namely there is an $\varepsilon > 0$ such that, for any $N > 0$, there is a $j > N$ with $\|\check{\mathbf{x}}_j - \mathbf{x}_*\|_2 \geq 2\,\varepsilon$. However, we have $\|\mathbf{x}_j - \mathbf{x}_*\|_2 < \varepsilon$ for all j larger than a certain N, implying

$$\|\check{\mathbf{x}}_j - \mathbf{x}_j\| \;\geq\; \|\check{\mathbf{x}}_j - \mathbf{x}_*\| - \|\mathbf{x}_* - \mathbf{x}_j\| \;>\; \varepsilon \;>\; \|\mathbf{x}_j - \mathbf{x}_*\|_2$$

that is a contradiction to (17).

Since $\mathbf{f}(\mathbf{x}_j) \neq \mathbf{0}$, we have $\mathbf{x}_j \neq \check{\mathbf{x}}_j$ and we can assume $(\mathbf{x}_j - \check{\mathbf{x}}_j)/\|\mathbf{x}_j - \check{\mathbf{x}}_j\|_2$ converges to a unit vector $\mathbf{v}$ for $j \to \infty$ due to compactness. Then

$$\begin{aligned} 0 &= \lim_{j\to\infty} \frac{\mathbf{f}_{\mathbf{x}}(\mathbf{x}_j)^{\dagger}_{\text{rank-}r} \left(\mathbf{f}(\check{\mathbf{x}}_j) - \mathbf{f}(\mathbf{x}_j)\right)}{\|\mathbf{x}_j - \check{\mathbf{x}}_j\|_2} \\ &= \lim_{j\to\infty} \frac{\mathbf{f}_{\mathbf{x}}(\mathbf{x}_j)^{\dagger}_{\text{rank-}r} \mathbf{f}_{\mathbf{x}}(\mathbf{x}_j)\,(\mathbf{x}_j - \check{\mathbf{x}}_j)}{\|\mathbf{x}_j - \check{\mathbf{x}}_j\|_2} \quad = \quad \mathbf{f}_{\mathbf{x}}(\mathbf{x}_*)^{\dagger}\, \mathbf{f}_{\mathbf{x}}(\mathbf{x}_*)\, \mathbf{v} \end{aligned}$$

by (8) and (9), implying $\mathbf{v} \in \mathit{Kernel}(\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*))$. As a result,

$$\mathit{span}\{\mathbf{v}\} \oplus \mathit{Range}(\phi_{\mathbf{z}}(\mathbf{z}_*)) \quad \subset \quad \mathit{Kernel}\,(\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*))$$

since $\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*)\phi_{\mathbf{z}}(\mathbf{z}_*) = O$ due to $\mathbf{f}(\phi(\mathbf{z})) \equiv \mathbf{0}$ in a neighborhood of $\mathbf{z}_*$. From the limit of (18) for $j \to \infty$, we have $\mathbf{v} \in \mathit{Range}(\phi_{\mathbf{z}}(\mathbf{z}_*))^{\perp}$ and thus

$$\mathit{nullity}\,(\,\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*)\,) \quad \geq \quad \mathit{rank}\,(\,\phi_{\mathbf{z}}(\mathbf{z}_*)\,) + 1$$

which is a contradiction to the regularity of $\mathbf{x}_*$.

By Lemma 3, there is a neighborhood Ω_2 of $\mathbf{x}_*$ such that every $\mathbf{x} \in \Omega_2 \cap \mathbf{f}^{-1}(\mathbf{0})$ is a regular zero of $\mathbf{f}$ in the same branch of $\mathbf{x}_*$. Thus the lemma holds for $\Omega_* = \Omega_1 \cap \Omega_2$. □

Remark 1 (A note on terminology) Keller [14] argues that the terminology "is somewhat unfortunate" on *isolated* and *nonisolated* zeros as he defines the later as zeros at which the Jacobian is not injective. The isolated zeros here are referred to as *geometrically isolated* in [14]. *Nonisolated* zeros defined by Keller including multiple 0-dimensional zeros and zeros on positive dimensional branches. The term *singular zero* is broadly accepted to include multiple isolated zero and nonisolated zero (c.f. [2, §1.2.2, p. 20]). We propose the term *ultrasingular zero* to distinguish those singular zeros from regular ones as defined in Definition 2.

4 Convergence theorem on exact equations

Consider the system of equations in the form of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ where $\mathbf{f} : \Omega \subset \mathbb{F}^m \to \mathbb{F}^n$ is a smooth mapping with $\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$. The system can be square ($m = n$), underdetermined ($m > n$) or overdetermined ($m < n$). We propose an iteration

$$\mathbf{x}_{k+1} \quad = \quad \mathbf{x}_k - J_{\text{rank-}r}(\mathbf{x}_k)^{\dagger}\, \mathbf{f}(\mathbf{x}_k) \quad \text{for} \quad k = 0, 1, \ldots \tag{19}$$

for computing a zero $\mathbf{x}_*$ of $\mathbf{f}$ at which the Jacobian $J(\mathbf{x}_*)$ is of rank r particularly when $\mathbf{x}_*$ is on a branch of regular nonisolated zeros. We loosely refer to (19) as the *rank-r Newton's iteration* or simply *Newton's iteration* since it is identical to the commonly-known Newton's iteration when $r = m = n$ with an invertible Jacobian. Assume the mapping $\mathbf{f}$ is given with exact data. The following theorem establishes the local quadratic convergence of the iteration (19). We shall consider the equation with empirical data in §5.

Theorem 1 (Convergence Theorem) *Let* $\mathbf{f}$ *be a smooth mapping in an open domain with a rank* r *Jacobian* $J(\mathbf{x}_*)$ *at a regular zero* $\mathbf{x}_*$. *For every open neighborhood* Ω_1 *of* $\mathbf{x}_*$, *there is a neighborhood* Ω_0 *of* $\mathbf{x}_*$ *such that, from every initial iterate* $\mathbf{x}_0 \in \Omega_0$, *the rank-r Newton's iteration* (19) *converges quadratically to a regular zero* $\hat{\mathbf{x}} \in \Omega_1$ *of* $\mathbf{f}$ *in the same branch as* $\mathbf{x}_*$.

Proof. Let $\Omega_* \ni \mathbf{x}_*$ be an open convex subset of the domain for $\mathbf{f}$ as specified in Lemma 1 so that (9) and (13) hold in Ω_*. From Lemma 4, we can further assume $J_{\text{rank-}r}(\mathbf{x})^\dagger \mathbf{f}(\mathbf{x}) = \mathbf{0}$ implies $\mathbf{x}$ is a regular zero of $\mathbf{f}$ in the same branch of $\mathbf{x}_*$ for every $\mathbf{x} \in \Omega_*$. Since $\mathbf{f}$ is smooth, there is a constant $\gamma > 0$ such that

$$\begin{aligned} \big\|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})\big\|_2 &\le \mu \|\mathbf{y} - \mathbf{x}\|_2 \\ \big\|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) - J(\mathbf{x})(\mathbf{y} - \mathbf{x})\big\|_2 &\le \gamma \|\mathbf{y} - \mathbf{x}\|_2^2 \end{aligned}$$

for all $\mathbf{x}, \mathbf{y} \in \Omega_*$. Denote $S_\varepsilon(\mathbf{x}_*) := \big\{\mathbf{x} \;\big|\; \|\mathbf{x} - \mathbf{x}_*\|_2 < \varepsilon\big\}$ for any $\varepsilon > 0$. For any given open neighborhood Ω_1 of $\mathbf{x}_*$, there is a δ with $0 < \delta < 2$ such that $\overline{S_\delta(\mathbf{x}_*)} \subset \Omega_* \cap \Omega_1$ with

$$\big\|J_{\text{rank-}r}(\mathbf{x})^\dagger\big\|_2 \big(\gamma \|\mathbf{x} - \mathbf{y}\|_2 + \zeta \|\mathbf{f}(\mathbf{y})\|_2\big) < h \tag{20}$$

for all $\mathbf{x}, \mathbf{y} \in S_\delta(\mathbf{x}_*)$ and, there is a τ with $0 < \tau < \frac{1}{2}$ such that

$$\big\|J_{\text{rank-}r}(\mathbf{z})^\dagger\big\|_2 \|\mathbf{f}(\mathbf{z})\|_2 \le \tfrac{1}{2}(1-h)\,\delta < \tfrac{1}{2}\,\delta < 1 \tag{21}$$

for all $\mathbf{z} \in S_\tau(\mathbf{x}_*)$. Then, for every $\mathbf{x}_0 \in S_\tau(\mathbf{x}_*)$, we have

$$\|\mathbf{x}_1 - \mathbf{x}_*\| \le \|\mathbf{x}_1 - \mathbf{x}_0\|_2 + \|\mathbf{x}_0 - \mathbf{x}_*\|_2 \le \|J_{\text{rank-}r}(\mathbf{x}_0)^\dagger\|_2 \|\mathbf{f}(\mathbf{x}_0)\|_2 + \tau < \delta.$$

Namely $\mathbf{x}_1 \in S_\delta(\mathbf{x}_*)$. Assume $\mathbf{x}_i \in S_\delta(\mathbf{x}_*)$ for all $i \in \{0, 1, \ldots, k\}$. Since

$$\begin{aligned} \mathbf{x}_j - \mathbf{x}_{j-1} - J_{\text{rank-}r}(\mathbf{x}_{j-1})^\dagger \mathbf{f}(\mathbf{x}_{j-1}) &= \mathbf{0} && \text{(by (19))} \\ J_{\text{rank-}r}(\mathbf{x}_j)^\dagger J_{\text{rank-}r}(\mathbf{x}_j) J_{\text{rank-}r}(\mathbf{x}_j)^\dagger &= J_{\text{rank-}r}(\mathbf{x}_j)^\dagger && \text{(by (6))} \\ J(\mathbf{x}_{j-1}) J_{\text{rank-}r}(\mathbf{x}_{j-1})^\dagger &= J_{\text{rank-}r}(\mathbf{x}_{j-1}) J_{\text{rank-}r}(\mathbf{x}_{j-1})^\dagger && \text{(by (7))} \end{aligned}$$

for all $j \in \{1, 2, \ldots, k\}$, we have

$$\begin{aligned} \big\|\mathbf{x}_{j+1} - \mathbf{x}_j\big\|_2 &= \big\|J_{\text{rank-}r}(\mathbf{x}_j)^\dagger \mathbf{f}(\mathbf{x}_j)\big\|_2 \\ &= \Big\|J_{\text{rank-}r}(\mathbf{x}_j)^\dagger \Big(\mathbf{f}(\mathbf{x}_j) - J(\mathbf{x}_{j-1})\big(\mathbf{x}_j - \mathbf{x}_{j-1} - J_{\text{rank-}r}(\mathbf{x}_{j-1})^\dagger \mathbf{f}(\mathbf{x}_{j-1})\big)\Big)\Big\|_2 \\ &\le \big\|J_{\text{rank-}r}(\mathbf{x}_j)^\dagger \big(\mathbf{f}(\mathbf{x}_j) - \mathbf{f}(\mathbf{x}_{j-1}) - J(\mathbf{x}_{j-1})(\mathbf{x}_j - \mathbf{x}_{j-1})\big)\big\|_2 \\ &\quad + \big\|J_{\text{rank-}r}(\mathbf{x}_j)^\dagger \big(J_{\text{rank-}r}(\mathbf{x}_j) J_{\text{rank-}r}(\mathbf{x}_j)^\dagger - J_{\text{rank-}r}(\mathbf{x}_{j-1}) J_{\text{rank-}r}(\mathbf{x}_{j-1})^\dagger\big) \mathbf{f}(\mathbf{x}_{j-1})\big\|_2 \\ &\le \|J_{\text{rank-}r}(\mathbf{x}_j)^\dagger\|_2 \big(\gamma \|\mathbf{x}_j - \mathbf{x}_{j-1}\|_2 + \zeta \|\mathbf{f}(\mathbf{x}_{j-1})\|_2\big) \|\mathbf{x}_j - \mathbf{x}_{j-1}\|_2 \qquad (22) \\ &\le h \|\mathbf{x}_j - \mathbf{x}_{j-1}\|_2 \end{aligned}$$

leading to

$$\big\|\mathbf{x}_{j+1} - \mathbf{x}_j\big\|_2 \le h^j \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \le h^j \tfrac{1-h}{2}\,\delta$$

for all $j \in \{1, \ldots, k\}$. Thus

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_0\|_2 &\leq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 + \cdots + \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \\ &\leq \left(h^k + h^{k-1} + \cdots + 1\right) \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \\ &< \tfrac{1}{1-h}\, \tfrac{1-h}{2}\, \delta \;=\; \tfrac{1}{2}\, \delta \end{aligned}$$

and $\mathbf{x}_{k+1} \in S_\delta(\mathbf{x}_*)$ since

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\|_2 \;\leq\; \|\mathbf{x}_{k+1} - \mathbf{x}_0\|_2 + \|\mathbf{x}_0 - \mathbf{x}_*\|_2 \;<\; \left(\tfrac{1}{2} + \tfrac{1}{2}\right) \delta \;=\; \delta,$$

completing the induction so all iterates $\left\{\mathbf{x}_j\right\}_{j=0}^{\infty}$ of (19) are in $S_\delta(\mathbf{x}_*)$ from any initial iterate $\mathbf{x}_0 \in S_\tau(\mathbf{x}_*)$. Furthermore, the iterates $\left\{\mathbf{x}_j\right\}_{j=0}^{\infty}$ form a Cauchy sequence since, for any $k,\, j \geq 0$,

$$\begin{aligned} \|\mathbf{x}_{k+j} - \mathbf{x}_k\|_2 &\leq \|\mathbf{x}_{k+j} - \mathbf{x}_{k+j-1}\|_2 + \cdots + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \\ &\leq \left(h^{j-1} + \cdots + h + 1\right) \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \\ &< \tfrac{1}{1-h}\, \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \qquad (23) \\ &< \tfrac{1}{1-h}\, h^k \cdot \tfrac{1-h}{2}\, \delta \;=\; \tfrac{1}{2}\, \delta\, h^k \qquad (24) \end{aligned}$$

can be as small as needed when k is sufficient large. Consequently, the sequence $\left\{\mathbf{x}_k\right\}_{k=0}^{\infty}$ generated by the iteration (19) converges to a certain $\hat{\mathbf{x}} \in \overline{S_\delta(\mathbf{x}_*)} \subset \Omega_1$ at which $J_{\text{rank-}r}(\hat{\mathbf{x}})^\dagger\, \mathbf{f}(\hat{\mathbf{x}}) = \mathbf{0}$ and thus, by Lemma 4, the limit $\hat{\mathbf{x}}$ is a regular zero in the same branch of $\mathbf{x}_*$ with the identical dimension.

We now have $\|\mathbf{x}_k - \hat{\mathbf{x}}\|_2 \leq \frac{1}{1-h} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \leq \frac{1}{2}\, \delta\, h^k$ for all $k \geq 1$ by setting $j \to \infty$ in (23) and (24). Substituting

$$\begin{aligned} \|\mathbf{f}(\mathbf{x}_{j-1})\|_2 \;&=\; \|\mathbf{f}(\mathbf{x}_{j-1}) - \mathbf{f}(\hat{\mathbf{x}})\|_2 \;\leq\; \mu\, \|\mathbf{x}_{j-1} - \hat{\mathbf{x}}\|_2 \\ &\leq\; \mu \left(\|\mathbf{x}_{j-1} - \mathbf{x}_j\|_2 + \|\mathbf{x}_j - \mathbf{x}_{j+1}\|_2 + \|\mathbf{x}_{j+1} - \mathbf{x}_{j+2}\|_2 + \cdots\right) \\ &\leq\; \frac{\mu}{1-h}\, \|\mathbf{x}_j - \mathbf{x}_{j-1}\|_2 \end{aligned}$$

for a certain $\mu > 0$, the inequality (22) yields

$$\|\mathbf{x}_{j+1} - \mathbf{x}_j\|_2 \;\leq\; \beta\, \|\mathbf{x}_j - \mathbf{x}_{j-1}\|_2^2 \;\leq\; \beta\, \|\mathbf{x}_1 - \mathbf{x}_0\|_2^{2^j} \;\leq\; \beta \left(\tfrac{\delta}{2}\right)^{2^j}$$

for all $j = 1, 2, \ldots,$ and thus

$$\begin{aligned} \|\mathbf{x}_k - \hat{\mathbf{x}}\|_2 &= \lim_{j \to \infty} \|\mathbf{x}_{k+j} - \mathbf{x}_k\|_2 \\ &= \lim_{j \to \infty} \left(\|\mathbf{x}_{k+j} - \mathbf{x}_{k+j-1}\|_2 + \cdots + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2\right) \\ &\leq \tfrac{1}{1-h}\, \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \;\leq\; \tfrac{\beta}{1-h} \left(\tfrac{\delta}{2}\right)^{2^k} \qquad (25) \end{aligned}$$

with $\frac{1}{2}\, \delta < 1$. Consequently the convergence to $\hat{\mathbf{x}}$ is at quadratic rate. $\square$

Theorem 1 can serve as the universal convergence theorem of Newton's iteration including the conventional version (1) as a special case. The sufficient conditions for Theorem 1 consist of smoothness of $\mathbf{f}$, regularity of $\mathbf{x}_*$ and the initial iterate $\mathbf{x}_0$ being near $\mathbf{x}_*$. These assumptions are minimal and indispensable for the convergence of the rank-r Newton's iteration (19). The iteration would need to be adapted if the mapping $\mathbf{f}$ is not smooth. Without the regularity of $\mathbf{x}_*$, the limit $\hat{\mathbf{x}}$ as a stationary point is seldom a zero from our experiment. The non-global convergence is an accepted imperfection of Newton's iteration rather than a drawback. Theoretical criteria on the initial iterate are of little practical meaning as trial-and-error would be easier than verifying those conditions.

The common definition of quadratic convergence of a sequence $\{\mathbf{x}_k\}$ to its limit $\hat{\mathbf{x}}$ would require that there is a constant λ such that $\|\mathbf{x}_k - \hat{\mathbf{x}}\|_2 \leq \lambda \|\mathbf{x}_{k-1} - \hat{\mathbf{x}}\|_2^2$ for k larger than a certain k_0 and imply $\|\mathbf{x}_k - \hat{\mathbf{x}}\|_2 \leq \lambda \|\mathbf{x}_{k_0} - \hat{\mathbf{x}}\|_2^{2^{k-k_0}}$. The inequality (25) ensures essentially the same error bound for the iterate $\mathbf{x}_k$ toward $\hat{\mathbf{x}}$ and can be used as an alternative definition of quadratic convergence.

Remark 2 (Convergence near a ultrasingular zero) If $\hat{\mathbf{x}}$ is an ultrasingular zero of $\mathbf{f}$ where $r = \mathit{rank}(\, J(\hat{\mathbf{x}})\,)$, the iteration (19) still locally converges to a certain stationary point $\check{\mathbf{x}}$ at which $J_{\text{rank-}r}(\check{\mathbf{x}})^\dagger \mathbf{f}(\check{\mathbf{x}}) = \mathbf{0}$ but $\check{\mathbf{x}}$ is not necessarily a zero of $\mathbf{f}$. Such an $\check{\mathbf{x}}$ satisfies the necessary condition for $\|\mathbf{f}(\mathbf{x})\|_2$ to reach a local minimum if $\mathit{rank}(\, J(\check{\mathbf{x}})\,) = r$. From our computational experiment, a stationary point to which the iteration (19) converges is rarely a zero of $\mathbf{f}$ when the initial iterate is near a ultrasingular zero.

Remark 3 (On the projection rank r) Application of the iteration (19) requires identification of the rank r of the Jacobian at a zero. There are various approaches for determining r such as analytical methods on the application model (c.f. §8.2 and §8.3), numerical matrix rank-revealing [12, 16, 19], and even trial-and-error. For any positive error tolerance $\theta < \left\|J(\mathbf{x}_*)^\dagger\right\|_2^{-1}$, the numerical rank $\mathit{rank}_\theta(\, J(\mathbf{x}_0)\,)$ within θ is identical to $r = \mathit{rank}(\, J(\mathbf{x}_*)\,)$ if $\mathbf{x}_0$ is sufficiently close to $\mathbf{x}_*$ [33]. Furthermore, the projection rank needs to be identified or computed only once for a solution branch and the same rank can be used repeatedly in calculating other witness points in the same branch.

5 Convergence theorem on perturbed equations

Practical applications in scientific computing often involve equations that are given through empirical data with limited accuracy. While the underlying exact equation can have solution sets of positive dimensions, the perturbed equation may not. Such applications can be modeled as solving an equation

$$\mathbf{f}(\mathbf{x},\, \mathbf{y}) \;=\; \mathbf{0} \quad \text{for} \quad \mathbf{x} \in \Omega \subset \mathbb{C}^m \;\text{ or }\; \mathbb{R}^m \tag{26}$$

at a particular parameter value $\mathbf{y} \in \Sigma \subset \mathbb{C}^n$ or $\mathbb{R}^n$ representing the data. The equation (26) may have a nonisolated solution set only at a particular isolated parameter value $\mathbf{y} = \mathbf{y}_*$. A typical example is given as Example 3 later in §8.1 where the 1-dimensional solutions in $\mathbf{x}$ exist for an equation $\mathbf{f}(\mathbf{x}, t) = \mathbf{0}$ only when $t = 1$ exactly. The question becomes: *Assuming the equation* (26) *has a regular solution* $\mathbf{x} = \mathbf{x}_*$ *at an underlying parameter* $\mathbf{y} = \mathbf{y}_*$ *but* $\mathbf{y}_*$ *is known only through empirical data in* $\tilde{\mathbf{y}} \approx \mathbf{y}_*$, *does the iteration*

$$\mathbf{x}_{k+1} \;=\; \mathbf{x}_k - \mathbf{f}_{\mathbf{x}}(\mathbf{x}_k,\, \tilde{\mathbf{y}})^{\dagger}_{\textit{rank-}r}\, \mathbf{f}(\mathbf{x}_k,\, \tilde{\mathbf{y}}), \qquad k \,=\, 0, 1, \ldots \tag{27}$$

converge and, if so, does the limit $\tilde{\mathbf{x}}$ *approximate a zero* $\mathbf{x} = \hat{\mathbf{x}}$ *of the mapping* $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y}_*)$ *with an accuracy* $\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|_2 = O(\|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2)$ *in the same order of the data?* The following theorem attempts to answer that question.

Theorem 2 (Convergnece Theorem on Perturbed Equations) *Let a smooth mapping* $(\mathbf{x},\, \mathbf{y}) \mapsto \mathbf{f}(\mathbf{x},\, \mathbf{y})$ *be defined in an open domain. Assume* $\mathbf{x}_*$ *is a regular zero of* $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y}_*)$ *at a fixed* $\mathbf{y}_*$ *with* $\mathit{rank}\big(\, \mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)\,\big) = r$. *Then there exist a neighborhood* $\Omega_* \times \Sigma_*$ *of* $(\mathbf{x}_*, \mathbf{y}_*)$, *a neighborhood* Ω_0 *of* $\mathbf{x}_*$ *and a constant* h *with* $0 < h < 1$ *such that, at every fixed* $\tilde{\mathbf{y}} \in \Sigma_*$ *serving as empirical data for* $\mathbf{y}_*$ *and from any initial iterate* $\mathbf{x}_0 \in \Omega_0$, *the iteration* (27) *linearly converges to a stationary point* $\tilde{\mathbf{x}} \in \Omega_*$ *at which* $\mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})^{\dagger}_{\textit{rank-}r}\, \mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \mathbf{0}$ *with an error bound*

$$\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|_2 \;\;\leq\;\; \tfrac{8}{1-h} \left\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)^{\dagger}\right\|_2 \left\|\mathbf{f}_{\mathbf{y}}(\mathbf{x}_*, \mathbf{y}_*)\right\|_2 \|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2 + O\big(\|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2^2\big) \tag{28}$$

toward a regular zero $\hat{\mathbf{x}}$ *of* $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y}_*)$ *in the same branch of* $\mathbf{x}_*$.

Proof. Following almost the same proof of Lemma 1 using the smoothness of the mapping $(\mathbf{x},\, \mathbf{y}) \mapsto \mathbf{f}(\mathbf{x},\, \mathbf{y})$, there is an open bounded convex neighborhood $\Omega_1 \times \Sigma_1$ of $(\mathbf{x}_*, \mathbf{y}_*)$ along with constants $\zeta,\, \gamma,\, \eta > 0$ such that

$$\begin{aligned}
\left\|\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})_{\text{rank-}r}\, \mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})^{\dagger}_{\text{rank-}r} - \mathbf{f}_{\mathbf{x}}(\check{\mathbf{x}}, \hat{\mathbf{y}})_{\text{rank-}r}\, \mathbf{f}_{\mathbf{x}}(\check{\mathbf{x}}, \hat{\mathbf{y}})^{\dagger}_{\text{rank-}r}\right\|_2 &\;\leq\; \zeta \left\|\hat{\mathbf{x}} - \check{\mathbf{x}}\right\|_2 \\
\left\|\mathbf{f}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) - \mathbf{f}(\check{\mathbf{x}}, \hat{\mathbf{y}}) - \mathbf{f}_{\mathbf{x}}(\check{\mathbf{x}}, \hat{\mathbf{y}})\,(\hat{\mathbf{x}} - \check{\mathbf{x}})\right\|_2 &\;\leq\; \gamma \left\|\hat{\mathbf{x}} - \check{\mathbf{x}}\right\|_2^2 \\
\left\|\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})^{\dagger}_{\text{rank-}r} - \mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}}, \check{\mathbf{y}})^{\dagger}_{\text{rank-}r}\right\|_2 &\;\leq\; \eta \left\|\hat{\mathbf{y}} - \check{\mathbf{y}}\right\|_2 \\
\left\|\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})\right\|_2 \;<\; 2 \left\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)\right\|_2 \quad \text{and} \quad \left\|\mathbf{f}_{\mathbf{y}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})\right\|_2 &\;<\; 2 \left\|\mathbf{f}_{\mathbf{y}}(\mathbf{x}_*, \mathbf{y}_*)\right\|_2
\end{aligned}$$

for all $\hat{\mathbf{x}},\, \check{\mathbf{x}} \in \Omega_1$ and $\hat{\mathbf{y}},\, \check{\mathbf{y}} \in \Sigma_1$ with

$$\max_{(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \overline{\Omega}_1 \times \overline{\Sigma}_1} \left\|\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})^{\dagger}_{\text{rank-}r}\right\|_2 \;\leq\; 2 \left\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)^{\dagger}\right\|_2$$

Let $S_\varepsilon(\mathbf{x}_*) := \big\{\mathbf{x} \in \Omega_1 \;\big|\; \|\mathbf{x} - \mathbf{x}_*\|_2 < \varepsilon\big\}$ for any $\varepsilon > 0$ and h be any fixed constant with $0 < h < 1$. There are constants $\delta,\, \tau,\, \tau' > 0$ with $6\,\tau' < 2\,\tau < \delta$, $S_\tau(\mathbf{x}_*) \subset S_\delta(\mathbf{x}_*) \subset \Omega_1$ and an open neighborhood $\Sigma_0 \subset \Sigma_1$ of $\mathbf{y}_*$ such that

$$\begin{aligned}
\left\|\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})^{\dagger}_{\text{rank-}r}\right\|_2 \left(\gamma \|\hat{\mathbf{x}} - \check{\mathbf{x}}\|_2 + \zeta \|\mathbf{f}(\hat{\mathbf{x}}, \hat{\mathbf{y}})\|_2\right) &\;<\; h \\
\left\|\mathbf{f}_{\mathbf{x}}(\mathbf{z}, \hat{\mathbf{y}})^{\dagger}_{\text{rank-}r}\right\|_2 \|\mathbf{f}(\mathbf{z}, \hat{\mathbf{y}})\|_2 &\;\leq\; \tfrac{1}{2}\,(1-h)\,\delta \;<\; \tfrac{1}{2}\,\delta \\
\left\|\mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{z}}, \hat{\mathbf{y}})^{\dagger}_{\text{rank-}r}\right\|_2 \|\mathbf{f}(\tilde{\mathbf{z}}, \hat{\mathbf{y}})\|_2 &\;\leq\; \tfrac{1}{3}\,(1-h)\,\tau \;<\; \tfrac{1}{3}\,\tau
\end{aligned}$$

for all $\hat{\mathbf{x}}, \check{\mathbf{x}} \in S_\delta(\mathbf{x}_*)$, $\mathbf{z} \in S_\tau(\mathbf{x}_*)$, $\tilde{\mathbf{z}} \in S_{\tau'}(\mathbf{x}_*)$ and $\hat{\mathbf{y}} \in \Sigma_0$. Using the same argument in the proof of Theorem 1, the sequence $\{\mathbf{x}_k\}$ generated by the iteration (27) starting from any $\mathbf{x}_0 \in S_{\tau'}(\mathbf{x}_*)$ satisfies $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 < h \|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2$ for all $k \geq 1$ and is a Cauchy sequence staying in $S_\tau(\mathbf{x}_*)$ for every fixed $\tilde{\mathbf{y}} \in \Sigma_0$ and converges to an $\tilde{\mathbf{x}} \in \overline{S_{\frac{2}{3}\tau}(\mathbf{x}_*)} \subset S_\tau(\mathbf{x}_*)$ that depends on the choices of $\mathbf{x}_0$ and $\tilde{\mathbf{y}}$. Resetting $\mathbf{x}_0 = \tilde{\mathbf{x}} \in S_\tau(\mathbf{x}_*)$, Theorem 1 ensures the iteration (27) with $\mathbf{y} = \mathbf{y}_* \in \Sigma_0$ stays in $S_\delta(\mathbf{x}_*)$ and converges to a certain regular zero $\hat{\mathbf{x}}$ of the mapping $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y}_*)$. We can assume

$$\big\|\hat{\mathbf{y}} - \check{\mathbf{y}}\big\|_2 \;<\; \frac{1-h}{4\,\eta\,\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)\|_2}$$

for all $\hat{\mathbf{y}}, \check{\mathbf{y}} \in \Sigma_0$ by shrinking Σ_0 if necessary. Subtracting both sides of

$$\begin{aligned} \tilde{\mathbf{x}} &= \tilde{\mathbf{x}} - \mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})^\dagger_{\text{rank-}r}\, \mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \\ \mathbf{x}_1 &= \tilde{\mathbf{x}} - \mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \mathbf{y}_*)^\dagger_{\text{rank-}r}\, \mathbf{f}(\tilde{\mathbf{x}}, \mathbf{y}_*) \end{aligned}$$

yields

$$\begin{aligned} \|\tilde{\mathbf{x}} - \mathbf{x}_1\|_2 &= \big\|\mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})^\dagger_{\text{rank-}r}\, \mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - \mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \mathbf{y}_*)^\dagger_{\text{rank-}r}\, \mathbf{f}(\tilde{\mathbf{x}}, \mathbf{y}_*)\big\|_2 \\ &\leq \big\|\mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})^\dagger_{\text{rank-}r} - \mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \mathbf{y}_*)^\dagger_{\text{rank-}r}\big\|_2\, \|\mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - \mathbf{f}(\hat{\mathbf{x}}, \mathbf{y}_*)\|_2 \\ &\quad + \big\|\mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \mathbf{y}_*)^\dagger_{\text{rank-}r}\big\|_2\, \|\mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - \mathbf{f}(\tilde{\mathbf{x}}, \mathbf{y}_*)\|_2. \end{aligned}$$

From

$$\begin{aligned} &\big\|\mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \mathbf{y}_*)^\dagger_{\text{rank-}r}\big\|_2\, \|\mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - \mathbf{f}(\tilde{\mathbf{x}}, \mathbf{y}_*)\|_2 \\ &\quad\leq\; 4\,\big\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)^\dagger\big\|_2\, \big\|\mathbf{f}_{\mathbf{y}}(\mathbf{x}_*, \mathbf{y}_*)\big\|_2\, \|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2 \end{aligned}$$

and, since $\mathbf{f}(\hat{\mathbf{x}}, \mathbf{y}_*) = \mathbf{0}$,

$$\begin{aligned} &\big\|\mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})^\dagger_{\text{rank-}r} - \mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, \mathbf{y}_*)^\dagger_{\text{rank-}r}\big\|_2\, \|\mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - \mathbf{f}(\hat{\mathbf{x}}, \mathbf{y}_*)\|_2 \\ &\quad<\; \eta\,\big\|\tilde{\mathbf{y}} - \mathbf{y}_*\big\|_2\, \big(2\,\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)\|_2\, \|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|_2 + O(\|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2)\big) \\ &\quad\leq\; \tfrac{1-h}{2}\, \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\| + O\big(\|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2^2\big), \end{aligned}$$

we have

$$\|\tilde{\mathbf{x}} - \mathbf{x}_1\|_2 \;\leq\; 4\,\big\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)^\dagger\big\|_2\, \big\|\mathbf{f}_{\mathbf{y}}(\mathbf{x}_*, \mathbf{y}_*)\big\|_2\, \|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2 + \tfrac{1-h}{2}\, \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\| + O\big(\|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2^2\big).$$

As a result,

$$\begin{aligned} \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|_2 &= \lim_{k\to\infty} \|\mathbf{x}_k - \tilde{\mathbf{x}}\|_2 \\ &\leq \lim_{k\to\infty} \big(\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2 + \cdots + \|\mathbf{x}_1 - \tilde{\mathbf{x}}\|_2\big) \\ &\leq \lim_{k\to\infty} (h^k + h^{k-1} + \cdots + 1)\, \|\mathbf{x}_1 - \tilde{\mathbf{x}}\|_2 \\ &\leq \tfrac{4}{1-h}\, \big\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*)^\dagger\big\|_2\, \big\|\mathbf{f}_{\mathbf{y}}(\mathbf{x}_*, \mathbf{y}_*)\big\|_2\, \|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2 + \tfrac{1}{2}\, \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\| + O\big(\|\tilde{\mathbf{y}} - \mathbf{y}_*\|_2^2\big), \end{aligned}$$

leading to (28). The theorem is proved by setting $\Omega_* \times \Sigma_* = S_\delta(\mathbf{x}_*) \times \Sigma_0$ and $\Omega_0 = S_{\tau'}(\mathbf{x}_*)$. □

Theorem 2 leads to some intriguing implications. At the exact data parameter value $\mathbf{y} = \mathbf{y}_*$, solving the equation $\mathbf{f}(\mathbf{x}, \mathbf{y}_*) = \mathbf{0}$ for a nonisolated solution in $\mathbf{x}$ can be an ill-posed problem as the structure of the solution can be infinitely sensitive to perturbations on the parameter $\mathbf{y}$. However, Theorem 2 implies that solving the stationary equation

$$\mathbf{f}_\mathbf{x}(\mathbf{x}, \mathbf{y}_*)^\dagger_{\text{rank-r}}\, \mathbf{f}(\mathbf{x}, \mathbf{y}_*) \;=\; \mathbf{0} \quad \text{for} \quad \mathbf{x} \in \Omega$$

in a neighborhood of a regular zero $\mathbf{x}_*$ is a well-posed problem in the sense that the solution is Lipschitz continuous with respect to the perturbations on the parameter $\mathbf{y}$ from $\mathbf{y}_*$ with $r = \mathit{rank}(\,\mathbf{f}_\mathbf{x}(\mathbf{x}_*, \mathbf{y}_*)\,)$. When the parameter $\mathbf{y}_*$ is only known through empirical data $\tilde{\mathbf{y}}$, the mapping $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, \tilde{\mathbf{y}})$ may not have a positive-dimensional zero of its own near $\mathbf{x}_*$ and, in some cases, does not have a zero at all. As a result, solving the equation $\mathbf{f}(\mathbf{x}, \tilde{\mathbf{y}}) = \mathbf{0}$ in exact sense for $\mathbf{x}$ is futile even if we extend the machine precision. It may be a pleasant surprise that a zero $\tilde{\mathbf{x}}$ of the mapping $\mathbf{x} \mapsto \mathbf{f}_\mathbf{x}(\mathbf{x}, \tilde{\mathbf{y}})^\dagger_{\text{rank-r}}\, \mathbf{f}(\mathbf{x}, \tilde{\mathbf{y}})$ exists near $\mathbf{x}_*$. Furthermore that $\tilde{\mathbf{x}}$ approximates a desired zero $\hat{\mathbf{x}}$ of the underlying mapping $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y}_*)$ with an accuracy in the same order as accuracy of the data. More importantly in practice, this numerical zero $\tilde{\mathbf{x}}$ is attainable as the rank-r Newton's iteration (27) locally converges to it.

The iteration (27) is more than an algorithm as it is, at the same time, a natural regularization of a hypersensitive zero-finding problem. Since $\mathbf{x}_*$ is a nonisolated zero of the mapping $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y}_*)$, the iteration does not necessarily converge to $\mathbf{x}_*$ but to some $\hat{\mathbf{x}}$ in the same branch of the solution set.

Remark 4 (Condition number of a nonisolated zero) Another implication of Theorem 2 lies in the sensitivity measure derived from the error estimate (28), from which we can naturally define a *condition number* of a zero $\mathbf{x}_*$ of the mapping $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y}_*)$ with respect to the parameter value $\mathbf{y} = \mathbf{y}_*$ as

$$\kappa_{\mathbf{f},\mathbf{x}}(\mathbf{x}_*, \mathbf{y}_*) \;:=\; \begin{cases} \left\|\mathbf{f}_\mathbf{x}(\mathbf{x}_*, \mathbf{y}_*)^\dagger\right\|_2 \left\|\mathbf{f}_\mathbf{y}(\mathbf{x}_*, \mathbf{y}_*)\right\|_2 & \text{if } \mathbf{x}_* \text{ is regular} \\ \infty & \text{otherwise.} \end{cases} \tag{29}$$

The condition number of a regular zero is finite and, if it is not large, a small perturbation in the data parameter $\mathbf{y}$ results in accurate approximation of the zero with an error estimate (28). On the other hand, errors of a ultrasingular zero has no known bound and the condition number is thus infinity.

Remark 5 (Convergence rate on perturbed systems) As established in Theorem 2, the convergence rate is generally linear on perturbed systems. The Newton's shift $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 = \|\mathbf{f}_\mathbf{x}(\mathbf{x}_k, \tilde{\mathbf{y}})^\dagger_{\text{rank-}r}\, \mathbf{f}(\mathbf{x}_k, \tilde{\mathbf{y}})\|_2$ approaches zero but the residual

$\|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_k, \tilde{\mathbf{y}})\|_2$ may not. However, the ratio of convergence is in the same order of the eventual residual $\|\mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})\|_2 = O(\|\mathbf{y}_* - \tilde{\mathbf{y}}\|_2)$. If this residual is as small as, say 10^{-4}, the convergence rate reduction may be negligible in practical computation since it takes only about four iteration steps to reach the hardware precision.

6 A geometric interpretation

We consider a special case first: Solving a consistent linear system $A\,\mathbf{x} = \mathbf{b}$ with a rank r matrix $A \in \mathbb{C}^{m\times n}$ using the rank-r Newton's iteration (19). Since the system is consistent, namely $\mathbf{b} \in \mathit{Range}(A)$, the solution set is an $(n-r)$-dimensional affine subspace

$$\begin{aligned} A^\dagger\,\mathbf{b} + \mathit{Kernel}(A) &:= \left\{ A^\dagger\,\mathbf{b} + \mathbf{y} \;\middle|\; \mathbf{y} \in \mathit{Kernel}(A) \right\} \\ &= \left\{ A^\dagger\,\mathbf{b} + N\,\mathbf{z} \;\middle|\; \mathbf{z} \in \mathbb{C}^{n-r} \right\} \end{aligned}$$

in which every particular solution is regular as defined in Definition 2 where columns of $N \in C^{n\times(n-r)}$ form an orthonormal basis for $\mathit{Kernel}(A)$. From any initial iterate $\mathbf{x}_0 \in \mathbb{C}^n$, the nearest point in the solution set $A^\dagger\,\mathbf{b} + \mathit{Kernel}(A)$ is $A^\dagger\,\mathbf{b} + N\,\mathbf{z}_0$ where $\mathbf{z} = \mathbf{z}_0$ is the least squares solution to the linear system

$$A^\dagger\,\mathbf{b} + N\,\mathbf{z} \;=\; \mathbf{x}_0.$$

Namely $\mathbf{z}_0 = N^{\mathsf{H}}\,(\mathbf{x}_0 - A^\dagger\,\mathbf{b}) = N^{\mathsf{H}}\,\mathbf{x}_0$ since $N^{\mathsf{H}}\,A^\dagger = O$, and the nearest solution

$$A^\dagger\,\mathbf{b} + N\,\mathbf{z}_0 \;=\; A^\dagger\,\mathbf{b} + N\,N^{\mathsf{H}}\,\mathbf{x}_0 \;=\; A^\dagger\,\mathbf{b} + (I - A^\dagger\,A)\,\mathbf{x}_0$$

since $(I - A^\dagger\,A) = N\,N^{\mathsf{H}}$ are the same orthogonal projection onto $\mathit{Kernel}(A)$. On the other hand, let the mapping $\mathbf{f} : \mathbb{C}^n \to \mathbb{C}^m$ be defined as $\mathbf{f}(\mathbf{x}) = A\,\mathbf{x} - \mathbf{b}$ with the Jacobian $J(\mathbf{x}) \equiv A \equiv J_{\text{rank-}r}(\mathbf{x})$. From $\mathbf{x}_0 \in \mathbb{C}^n$, the rank-$r$ Newton's iteration (19) requires only one step

$$\mathbf{x}_1 \;=\; \mathbf{x}_0 - A^\dagger\,(A\,\mathbf{x}_0 - \mathbf{b}) \;=\; A^\dagger\,\mathbf{b} + (I - A^\dagger\,A)\,\mathbf{x}_0$$

In other words, the rank-r Newton's iteration converges to the nearest solution on the $(n-r)$-dimensional solution set from the initial iterate.

We now consider a general nonlinear mapping $\mathbf{f} : \Omega \subset \mathbb{C}^m \to \mathbb{C}^n$ with the Jacobian $J(\mathbf{x})$ at any $\mathbf{x} \in \Omega$. At an iterate $\mathbf{x}_j$ near a regular $(n-r)$-dimensional solution $\hat{\mathbf{x}}$ of the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, we have

$$\begin{aligned} \mathbf{f}(\mathbf{x}) &= \mathbf{f}(\mathbf{x}_j) + J(\mathbf{x}_j)\,(\mathbf{x} - \mathbf{x}_j) + O(\|\mathbf{x} - \mathbf{x}_j\|_2^2) \\ &\approx \mathbf{f}(\mathbf{x}_j) + J_{\text{rank-}r}(\mathbf{x}_j)\,(\mathbf{x} - \mathbf{x}_j). \end{aligned}$$

On the other hand, from $\mathbf{f}(\mathbf{x}_j) = J(\hat{\mathbf{x}})\,(\mathbf{x}_j - \hat{\mathbf{x}}) + O\big(\|\mathbf{x}_j - \hat{\mathbf{x}}\|_2^2\big)$ we have

$$\big(I - J(\hat{\mathbf{x}})\,J(\hat{\mathbf{x}})^\dagger\big)\,\big(\mathbf{f}(\mathbf{x}_j) - \mathbf{f}(\hat{\mathbf{x}})\big) \;=\; O\big(\|\mathbf{x}_j - \hat{\mathbf{x}}\|_2^2\big)$$

and

$$\begin{aligned}
\mathbf{f}(\mathbf{x}_j) &= J_{\text{rank-}r}(\mathbf{x}_j)\, J_{\text{rank-}r}(\mathbf{x}_j)^\dagger\, \mathbf{f}(\mathbf{x}_j) + \big(I - J_{\text{rank-}r}(\mathbf{x}_j)\, J_{\text{rank-}r}(\mathbf{x}_j)^\dagger\big)\, \mathbf{f}(\mathbf{x}_j) \\
&= J_{\text{rank-}r}(\mathbf{x}_j)\, J_{\text{rank-}r}(\mathbf{x}_j)^\dagger\, \mathbf{f}(\mathbf{x}_j) + \big(I - J(\hat{\mathbf{x}})\, J(\hat{\mathbf{x}})^\dagger\big)\, \big(\mathbf{f}(\mathbf{x}_j) - \mathbf{f}(\hat{\mathbf{x}})\big) \\
&\qquad + \big(J(\hat{\mathbf{x}})\, J(\hat{\mathbf{x}})^\dagger - J_{\text{rank-}r}(\mathbf{x}_j)\, J_{\text{rank-}r}(\mathbf{x}_j)^\dagger\big)\, \big(\mathbf{f}(\mathbf{x}_j) - \mathbf{f}(\hat{\mathbf{x}})\big) \\
&= J_{\text{rank-}r}(\mathbf{x}_j)\, J_{\text{rank-}r}(\mathbf{x}_j)^\dagger\, \mathbf{f}(\mathbf{x}_j) + O\big(\|\mathbf{x}_j - \hat{\mathbf{x}}\|_2^2\big)
\end{aligned}$$

The basic principle of Newton's iteration is to designate the next iterate $\mathbf{x} = \mathbf{x}_{j+1}$ as a numerical solution of the linear system

$$\mathbf{f}(\mathbf{x}_j) + J(\mathbf{x}_j)\,(\mathbf{x} - \mathbf{x}_j) \;=\; \mathbf{0}. \tag{30}$$

Since $J(\hat{\mathbf{x}})$ is rank-deficient at the nonisolated zero $\hat{\mathbf{x}}$ and $\mathbf{x}_j$ is near $\hat{\mathbf{x}}$, the coefficient matrix $J(\mathbf{x}_j)$ of the system (30) is expected to be highly ill-conditioned for being nearly rank-deficient. Consequently, solving the system (30) as it is may not be reliable. As elaborated in [33, §8], an alternatives is to solve the nearby linear system

$$J_{\text{rank-}r}(\mathbf{x}_j)\, J_{\text{rank-}r}(\mathbf{x}_j)^\dagger\, \mathbf{f}(\mathbf{x}_j) + J_{\text{rank-}r}(\mathbf{x}_j)\,(\mathbf{x} - \mathbf{x}_j) \;=\; \mathbf{0} \tag{31}$$

and pick a proper vector $\mathbf{x} = \mathbf{x}_{j+1}$ from the solution in the affine Grassmannian. The linear system (31) is well-conditioned if the actual sensitivity measure $\|J(\hat{\mathbf{x}})\|_2\, \|J(\hat{\mathbf{x}})^\dagger\|_2 \approx \|J_{\text{rank-}r}(\mathbf{x}_j)\|_2\, \|J_{\text{rank-}r}(\mathbf{x}_j)^\dagger\|_2$ is not large and (31) is an approximation to the system (30). The iterate $\mathbf{x} = \mathbf{x}_{j+1}$ from (19) is the minimum norm solution of the approximate system (31) and

$$\mathbf{x}_{j+1} - \mathbf{x}_j \;\in\; \mathit{Kernel}\big(J_{\text{rank-}r}(\mathbf{x}_j)\big)^\perp \;\approx\; \mathit{Kernel}\big(J(\hat{\mathbf{x}})\big)^\perp \;=\; \mathit{Range}\big(\phi_{\mathbf{z}}(\hat{\mathbf{z}})\big)^\perp$$

where $\mathbf{z} \mapsto \phi(\mathbf{z})$ is the parametrization of the $(n-r)$-dimensional zero of $\mathbf{f}$ with $\phi(\hat{\mathbf{z}}) = \hat{\mathbf{x}}$, implying the geometric interpretation of the rank-r Newton's iteration:

> From the initial iterate, the rank-r Newton's iteration (19) aims at the nearest point on the solution branch following a normal line toward the solution set.

How accurate the iteration hitting the nearest point depending on several factors including how far the initial iterate is. Similar geometric interpretations are also observed in [6, 27] in the cases where the Jacobians are surjective.

Example 1 (Normal line direction of convergence) Consider $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ as follows

$$\mathbf{f}(x, y) \;=\; \begin{pmatrix} x^3 + x\,y^2 - x + 2\,x^2 + 2\,y^2 - 2 \\ x^2\,y + y^3 - y - 3\,x^2 - 3\,y^2 + 3 \end{pmatrix} \tag{32}$$

whose zeros consist of a regular 1-dimensional unit circle $x^2 + y^2 = 1$ and a 0-dimensional isolated point $(-2, 3)$. Starting from $(x_0, y_0) = (1.8, 0.6)$, the

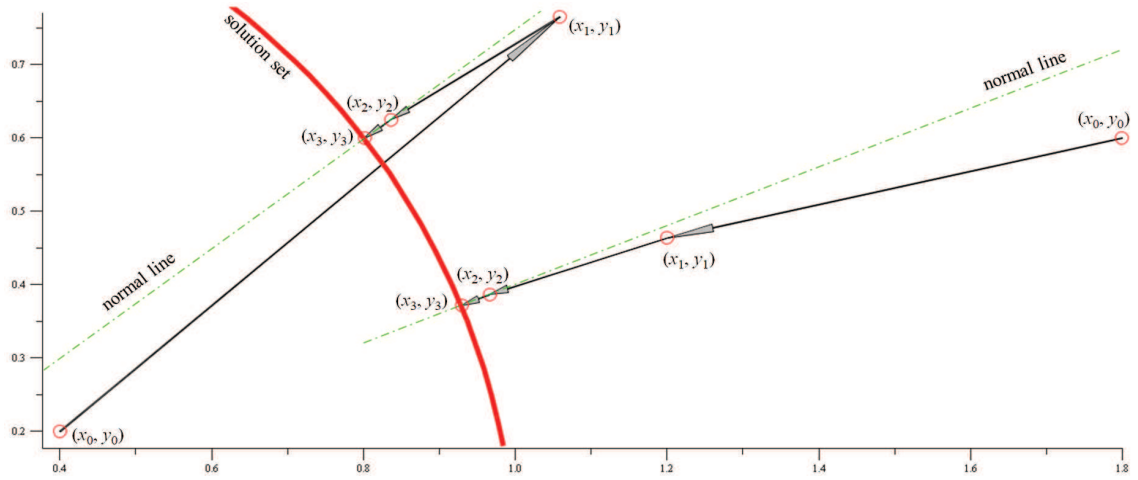

Figure 1: Each sequence of iteration (19) asymptotically follows a normal line toward the solution set. Illustration is plotted using actual data in Example 1 from two initial iterates.

rank-1 Newton's iteration (19) converges to $(\hat{x}, \hat{y}) \approx (0.928428592, 0.3715109)$ on the unit circle. Starting from $(x_0, y_0) = (0.4, 0.2)$, the iteration converges to $(0.8007609\ldots, 0.5989721\ldots)$. Both sequences of iterates asymptotically follow corresponding normal lines of the solution set toward the particular solutions as shown in Figure 1.

7 Algorithms for computing $J_{\text{rank-}r}(\mathbf{x}_k)^\dagger \mathbf{f}(\mathbf{x}_k)$

Computing the shift $J_{\text{rank-}r}(\mathbf{x}_k)^\dagger \mathbf{f}(\mathbf{x}_k)$ at the iterate $\mathbf{x}_k$ is the problem of calculating the *minimum norm solution*, or *minimum norm least squares solution* if $\mathbf{b} \notin \mathcal{Range}(A_{\text{rank-}r})$, of the linear system

$$A_{\text{rank-}r}\,\mathbf{z} \;=\; \mathbf{b} \quad \text{where} \quad A \in \mathbb{C}^{m\times n} \quad \text{and} \quad r \le \min\{m, n\}. \tag{33}$$

The most reliable method is based on the singular value decomposition in the following Algorithm 1.

Algorithm 1: Computing $A^\dagger_{\text{rank-}r}\,\mathbf{b}$ by SVD

Input: matrix $A \in \mathbb{C}^{m\times n}$, vector $\mathbf{b} \in \mathbb{C}^m$, integer $r \le \min\{m, n\}$.

- By a full or partial SVD, calculate singular values and corresponding left and right singular vectors σ_j, $\mathbf{u}_j$, $\mathbf{v}_j$, for $j = 1, \ldots, r$.
- calculate $\mathbf{y} = \left[\frac{\mathbf{u}_1^{\mathsf{H}}\mathbf{b}}{\sigma_r}, \ldots, \frac{\mathbf{u}_r^{\mathsf{H}}\mathbf{b}}{\sigma_1}\right]^\top$
- calculate $\mathbf{z} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]\,\mathbf{y}$

output $A^\dagger_{\text{rank-}r}\,\mathbf{b} = \mathbf{z}$

If the Jacobian $J(\mathbf{x}_*)$ at the zero $\mathbf{x}_*$ is of low rank such that $r \ll \min\{m, n\}$, a partial SVD such as the USV-plus decomposition [16] can be more efficient. For the cases of $r \lessapprox \min\{m, n\}$, we need the following lemma [33, Lemma 13].

Lemma 5 *Let A be an $m \times n$ matrix with right singular vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $A_{\textit{rank-}r}$ be its rank-r projection. Assume $\sigma_r(A) > \sigma_{r+1}(A)$ and N is a matrix whose columns form an orthonormal basis for $\mathit{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$. Then, for every dimension-m vector $\mathbf{b}$, the following identity hold:*

$$A^{\dagger}_{\textit{rank-}r}\,\mathbf{b} \;=\; (I - N\,N^{\mathsf{H}}) \begin{bmatrix} \mu\,N^{\mathsf{H}} \\ A \end{bmatrix}^{\dagger} \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}. \tag{34}$$

For matrices $A \in \mathbb{C}^{m\times n}$ of large sizes with $r \approx \min\{m, n\}$, the SVD can be unnecessarily expensive. There are reliable algorithms that can be substantially more efficient. We briefly elaborate some of those algorithms in this section.

The simplest case of computing $A^{\dagger}_{\text{rank-}r}\,\mathbf{b}$ is when r equals the row dimension n. This case arises when solving an underdetermined system $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ when the Jacobian is surjective at the nonisolated solution. The following algorithm is a well established numerical methods for computing the minimum norm solution of a full rank underdetermined linear system [12, §5.6.2].

Algorithm 2: Solving (33) when $r = m < n$
Input: matrix $A \in \mathbb{C}^{m\times n}$ with $m < n$, vector $\mathbf{b} \in \mathbb{C}^m$ (integer $r = m$).
- calculate the thin QR decomposition [12, p. 248] $A^{\mathsf{H}} = Q\,R$
- solve the lower-triangular system $R^{\mathsf{H}}\mathbf{y} = \mathbf{b}$ for $\mathbf{y}$
- set $\mathbf{z} = Q\,\mathbf{y}$

output $A^{\dagger}_{\text{rank-}r}\,\mathbf{b} = \mathbf{z}$

On the case $r < m < n$, the following Algorithm 3 is modified from the rank-revealing method in [19].

Algorithm 3: Solving (33) when $r < m < n$
Input: matrix $A \in \mathbb{C}^{m\times n}$ with $m < n$, vector $\mathbf{b} \in \mathbb{C}^m$, integer $r < m$.
- calculate the thin QR decomposition [12, p. 248] $A^{\mathsf{H}} = Q_0\,R_0$
- set $G_0 = R_0^{\mathsf{H}}$.
- for $k = 1, \ldots, m - r$ do
  - ∗ calculate the vector $\mathbf{u}_k$ as the terminating iterate of the iteration (see [19] for details) with $\tau = \|A\|_\infty$

$$\mathbf{y}_{j+1} = \mathbf{y}_j - \begin{bmatrix} 2\tau\,\mathbf{y}_j^{\mathsf{H}} \\ G_{k-1} \end{bmatrix}^{\dagger} \begin{bmatrix} \tau\,\mathbf{y}_j^{\mathsf{H}}\mathbf{y}_j - \tau \\ G_{k-1}\,\mathbf{y}_j \end{bmatrix}, \quad j = 0, 1, \ldots \tag{35}$$

from a random unit vector $\mathbf{y}_0 \in \mathbb{C}^m$

* As a by product of terminating (35), extract the thin QR decomposition $[2\,\tau\,\mathbf{u}_k,\, G_{k-1}^{\mathsf{H}}]^{\mathsf{H}} = Q_k\,G_k$

end do

– solve for $\mathbf{y} = \mathbf{y}_*$ of the triangular system

$$G_{m-r}^{\mathsf{H}}\,\mathbf{y} = Q_{m-r}^{\mathsf{H}} \begin{bmatrix} \mathbf{0}_{m-r} \\ \mathbf{b} \end{bmatrix}$$

– set $\mathbf{z}_* = Q_0\,(I - U\,U^{\mathsf{H}})\,\mathbf{y}_*$ where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_{m-r}]$

output $A^{\dagger}_{\text{rank-}r}\,\mathbf{b} = \mathbf{z}_*$

Notice that G_0 is lower triangular. It is a standard technique in numerical linear algebra to apply Given's rotation [12, §5.1.8] to obtain a QR decomposition $[2\,\tau\,\mathbf{y}_j,\, G_{k-1}^{\mathsf{H}}]^{\mathsf{H}} = Q\,R$ where R is in lower triangular form throughout the process. The algorithm can be easily explained as follows: Let $A^{\mathsf{H}} = [Q_0, \tilde{Q}] \begin{bmatrix} R_0 \\ O \end{bmatrix}$ be a full QR decomposition of A^{H} (while $Q_0\,R_0$ is the corresponding thin version). Then $\mathit{Range}(\tilde{Q})$ is a subspace of $\mathit{Kernel}\,(A)$ whose basis consists of $m-r$ additional vectors besides columns of $\tilde{Q}$. The vectors $Q_0\,\mathbf{u}_1, \ldots, Q_0\,\mathbf{u}_{m-r}$ approximately form an orthonormal basis for the vector space spanned by the $m-r$ right singular vectors of A. The solution $\mathbf{y}_*$ of the equation $G_{m-r}^{\mathsf{H}}\,\mathbf{y} = Q_{m-r}^{\mathsf{H}}\mathbf{b}$ is the least squares solution of the linear system

$$\begin{bmatrix} 2\,\tau\,U^{\mathsf{H}} \\ R_0^{\mathsf{H}} \end{bmatrix} \mathbf{y} \;=\; \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}.$$

Then it is a straightforward verification using Lemma 5 that $\mathbf{z}_* = Q_0\,(I - U\,U^{\mathsf{H}})\,\mathbf{y}_*$ is the least squares solution of $A_{\text{rank-}r}\,\mathbf{z} = \mathbf{b}$ that is orthogonal to columns of $[Q_0\,U,\, \tilde{Q}]$.

For the cases of solving (33) where $m \geq n$, the kernel or the partial kernel of A generally can not be computed from a QR decomposition of A^{H} like the cases for $m < n$. As a result, a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_{n-r}\}$ for $\mathit{Kernel}\,(A_{\text{rank-}r})$ needs to be computed as shown in Algrithm 4 below.

Algorithm 4: Solving (33) when $r \leq n \leq m$

Input: matrix $A \in \mathbb{C}^{m\times n}$ with $m \geq n$, vector $\mathbf{b} \in \mathbb{C}^m$, integer $r \leq n$.

– calculate the thin QR decomposition [12, p. 248] $A = Q_0\,R_0$

– for $k = 1, \ldots, n-r$ do

* calculate the vector $\mathbf{u}_k$ as the terminating iterate of the iteration (see [19] for details) with $\tau = \|A\|_\infty$

$$\mathbf{y}_{j+1} = \mathbf{y}_j - \begin{bmatrix} 2\tau\,\mathbf{y}_j^{\mathsf{H}} \\ R_{k-1} \end{bmatrix}^{\dagger} \begin{bmatrix} \tau\,\mathbf{y}_j^{\mathsf{H}}\,\mathbf{y}_j - \tau \\ R_{k-1}\,\mathbf{y}_j \end{bmatrix}, \quad j = 0, 1, \ldots \tag{36}$$

from a random unit vector $\mathbf{y}_0 \in \mathbb{C}^m$

* As a by product of terminating (35), extract the thin QR decomposition $[2\,\tau\,\mathbf{u}_k,\ R_{k-1}]^{\mathsf{H}} = Q_k\,R_k$

end do

– solve for $\mathbf{y} = \mathbf{y}_*$ of the triangular system

$$R_{n-r}\,\mathbf{y} = Q_{n-r}^{\mathsf{H}} \begin{bmatrix} \mathbf{0}_{n-r} \\ \mathbf{b} \end{bmatrix}$$

– set $\mathbf{z}_* = (I - U\,U^{\mathsf{H}})\,\mathbf{y}_*$ where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_{n-r}]$

output $A^{\dagger}_{\text{rank-}r}\,\mathbf{b} = \mathbf{z}_*$

8 Modeling with nonisolated solutions

Models with nonisolated solutions arise in many applications. Unaware of the capability of Newton's iteration in computing such solutions, scientific computing practionners go to great lengths to make solutions isolated by various techniques such as adding auxiliary equations and variables. We shall elaborate case studies in which nonisolated solutions can naturally be modeled into well-posed computational problems. The dimensions of the solution sets and their regularity may also be obtained analytically in the modeling process so that the projection rank r of Newton's iteration (19) becomes readily available. The rank-r Newton's iteration for any integer r is implemented in our Numerical Algebraic Computation toolbox NACLAB[1] for Matlab as a functionality `Newton`. We shall demonstrate the implementation and its effectiveness in computing nonisolated solutions with no need for auxiliary equations.

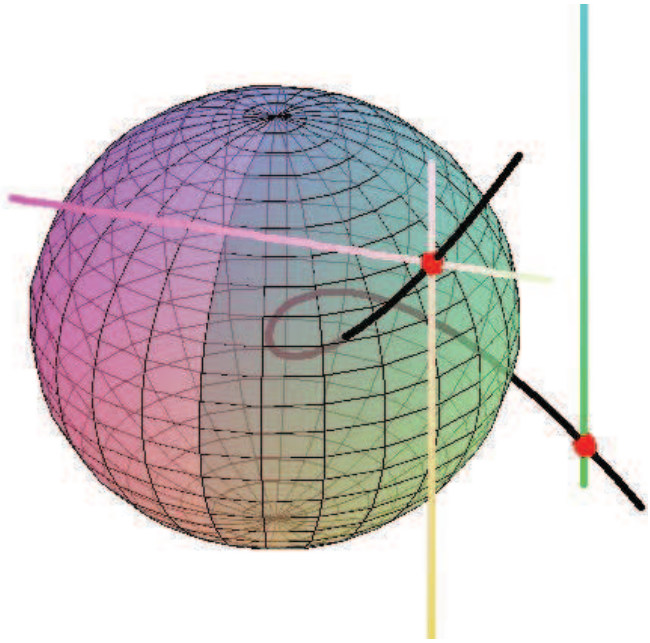

Figure 2: Solution sets of the system in (37)

8.1 Numerical Algebraic Geometry

One of the main subjects of numerical algebraic geometry and its application in algebraic kinematics is computing solutions of positive dimensions to polynomial sys-

[1] `http://homepages.neiu.edu/ zzeng/naclab.html`

tems, as elaborated extensively by Wampler and Sommese [28] and in the monographs [2, 23]. Various mechanisms have been developed in solving those polynomial systems for nonisolated solutions, including adding auxiliary equations and variables to isolate witness points on the solution sets. The rank-r Newton's iteration (19) can be applied to calculating regular witness points directly on the polynomial systems without needing extra equations and variables.

Example 2 (A polynomial system) An illustrative example for nonisoalted solutions of different dimensions is given in [2, p. 143] as follows:

$$\mathbf{f}(x,y,z) = \begin{pmatrix} (y-x^2)\,(x^2+y^2+z^2-1)\,(x-1) \\ (z-x^3)\,(x^2+y^2+z^2-1)\,(y-1) \\ (y-x^2)\,(z-x^3)\,(x^2+y^2+z^2-1)\,(z-1) \end{pmatrix}. \tag{37}$$

Among the solution sets, a point $(1,1,1)$, a curve $\{y = x^2,\ z = x^3\}$ along with three lines, and a surface $\{x^2+y^2+z^2 = 1\}$ are of dimensions 0, 1 and 2 respectively, as shown in Figure 2. Zeros are regular except at intersections of the solution sets. The iteration (19) converges at quadratic rate toward those solutions from proper initial iterates by setting the projection rank $r = 3,\ 2$ and 1 respectively.

Example 3 (Perturbed cyclic-4 system) Cyclic-n roots are among the benchmark problems in solving polynomial systems. Those systems arise in applications such as biunimodular vectors, a notion traces back to Gauss [11]. In general, every cyclic-n system possesses positive dimensional solution sets if n is a multiple of a perfect square [1].

We simulate practical computation with empirical data through the cyclic-4 system in $\mathbf{x} = (x_1, x_2, x_3, x_4) \in \mathbb{C}^4$ with a parameter $t \in \mathbb{C}$:

$$\mathbf{f}(\mathbf{x},\, t) \;:=\; \begin{pmatrix} x_1 + x_2 + x_3 + x_4 \\ t\,x_1\,x_2 + x_2\,x_3 + x_3\,x_4 + x_4\,x_1 \\ x_1\,x_2\,x_3 + x_2\,x_3\,x_4 + x_3\,x_4\,x_1 + x_4\,x_1\,x_2 \\ x_1\,x_2\,x_3\,x_4 - 1 \end{pmatrix} \tag{38}$$

With $t_* = 1$, the equation $\mathbf{f}(\mathbf{x},\, 1) = \mathbf{0}$ is the cyclic-4 system whose solution consists of two 1-dimensional sets

$$\{x_1 = -x_3,\ x_2 = -x_4,\ x_3\,x_4 = \pm 1,\ t = 1\}. \tag{39}$$

All zeros are regular except eight ultrasingular zeros in the form of $(\pm 1, \pm 1, \pm 1, \pm 1)$ and $(\pm i, \pm i, \pm i, \pm i)$ with proper choices of signs. When the parameter t is perturbed from $t_* = 1$ to any other nearby value $\tilde{t}$, the 1-dimensional solution sets dissipate into 16 isolated solutions. Consequently, the parameter value $t_* = 1$ is a bifurcation point at which the solution changes structure.

For instance, suppose the parameter t_* is known approximately, say $\tilde{t} = 0.9999$. Even though the solutions of $\mathbf{f}(\mathbf{x}, \tilde{t}) = \mathbf{0}$ are all isolated, Theorem 2 ensures the iteration

$$\mathbf{x}_{k+1} \;=\; \mathbf{x}_k - \mathbf{f}_{\mathbf{x}}\big(\mathbf{x}_k, \tilde{t}\,\big)^{\dagger}_{\text{rank-3}}\, \mathbf{f}(\mathbf{x}_k, \tilde{t}), \qquad k = 0, 1, \ldots$$

converges to a numerical solution $\tilde{\mathbf{x}}$ as a zero of the mapping $\mathbf{x} \mapsto \mathbf{f}_{\mathbf{x}}(\mathbf{x}, \tilde{t})^{\dagger}_{\text{rank-3}}\, \mathbf{f}(\mathbf{x}, \tilde{t})$ that approximates a point in the 1-dimensional solution set of the underlying system $\mathbf{f}(\mathbf{x}, 1) = \mathbf{0}$ with an error in the order of $|\tilde{t} - t_*| = 10^{-4}$ if $\mathbf{x}_0$ is sufficiently close to a regular point in the solution set (39), as demonstrated in the following calling sequence applying the NACLAB module `Newton`:

```
>> P = {'x1+x2+x3+x4','0.9999*x1*x2+x2*x3+x3*x4+x4*x1',...; % enter the perturbed cyclic-4 system as
       'x1*x2*x3+x2*x3*x4+x3*x4*x1+x4*x1*x2', 'x1*x2*x3*x4-1'};  % a cell array of character strings
>> v = {'x1','x2','x3','x4'};                                  % enter cell array of the variable names
>> J = PolynomialJacobian(P,v);                          % Jacobian of P w.r.t.  the variable names in v
>> f = @(x,P,J,v) PolynomialEvaluate(P,v,x);     % the function handle for evaluate the system P at x
>> fjac = @(x,x0,P,J,v) PolynomialEvaluate(J,v,x0)*x;              % function for evaluating J at x
>> domain = ones(4,1);  parameter = {P,J,v};          % domain (space of 4x1 vectors) and parameters
>> z0 = [0.8;1.2;-0.8;-1.2];                                                          % initial z0
>> [z,res,fcond] =  Newton({f,domain,parameter},{fjac,3},z0,1);  % call rank-3 Newton iteration on f
                                                        % projection from z0 using display type 1
 Step  0:  residual = 7.8e-02
 Step  1:  residual = 2.4e-03  shift = 2.4e-02
 Step  2:  residual = 1.0e-04  shift = 6.8e-04
 Step  3:  residual = 1.0e-04  shift = 5.8e-07
 Step  4:  residual = 1.0e-04  shift = 4.3e-13
 Step  5:  residual = 1.0e-04  shift = 3.6e-16
 Step  6:  residual = 1.0e-04  shift = 1.5e-16
```

The iteration does *not* converges to a solution to $\mathbf{f}(\mathbf{x}, 0.9999) = \mathbf{0}$ as the residual $\|\mathbf{f}(\mathbf{x}_j, 0.9999)\|_2$ can only be reduced to 10^{-4} but the shifts

$$\|\mathbf{x}_{j+1} - \mathbf{x}_j\|_2 \;=\; \|\mathbf{f}_{\mathbf{x}}(\mathbf{x}_j, 0.9999)^{\dagger}_{\text{rank-3}}\, \mathbf{f}(\mathbf{x}_j, 0.9999)\|_2 \;\longrightarrow\; 1.51 \times 10^{-16}$$

approaches to the unit roundoff, indicating the iteration converges to a stationary point $\tilde{\mathbf{x}}$ at which $\mathbf{f}_{\mathbf{x}}(\tilde{\mathbf{x}}, 0.9999)^{\dagger}_{\text{rank-3}}\, \mathbf{f}(\tilde{\mathbf{x}}, 0.9999) = \mathbf{0}$. The module `Newton` terminates at $\tilde{\mathbf{x}}$ that approximates the nearest solution $\hat{\mathbf{x}}$ of the underlying equation $\mathbf{f}(\mathbf{x}, 1) = \mathbf{0}$ where

$$\begin{aligned}
\tilde{\mathbf{x}} \;&=\; \big(\texttt{0.82287906 1867739},\ \texttt{1.21524540 1950727},\ \texttt{-0.82287906 2858240},\ \texttt{-1.215245403 413521}\big)\\
\hat{\mathbf{x}} \;&=\; \big(\texttt{0.82287906 3773473},\ \texttt{1.21524540 3637205},\ \texttt{-0.82287906 3773473},\ \texttt{-1.215245403 637205}\big)
\end{aligned}$$

with a forward error 2.71×10^{-9} much smaller than the data error $10^{-4} = |\tilde{t} - t_*|$ as asserted in Theorem 2. As a confirmation of the geometric interpretation of the iteration elaborated in §6, the point $\hat{\mathbf{x}}$ is also only about 6.3×10^{-4} away from the nearest point in the solution set to the initial iterate $\mathbf{x}_0 = (\texttt{0.8}, \texttt{1.2}, \texttt{-0.9}, \texttt{-1.2})$.

Example 4 (A bifurcation model) The parameterized cyclic-4 system (38) leads

to a bifurcation model in the form of the equation

$$\mathbf{f}(\mathbf{x},\, t) \;\; = \;\; \mathbf{0} \quad \text{for} \quad (\mathbf{x}, t) \,\in\, \mathbb{R}^4 \times \mathbb{R}$$

depending on the parameter t. The solution of the equation $\mathbf{f}(\mathbf{x}, t) = \mathbf{0}$ for $(\mathbf{x}, t) \in \mathbb{R}^4 \times \mathbb{R}$ consists of 16 branches $\mathbf{x} = \psi_j(t)$ for $j = 1, \ldots, 16$ and $t \in \mathbb{R}$. On the plane $t = 1$, eight pairs of those 16 branches intersect at 8 bifurcation points that are embeded in two additional branches of solution curves (39). All solutions are regular except at the intersections. The parameter value $t_* = 1$ is the bifurcation point at which the solution changes the structure.

After finding $\tilde{\mathbf{x}}$ in Example 2 as an approximate zero of $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x},\, t_*)$ from the empirical data $\tilde{t} = 0.9999$, the point $(\tilde{\mathbf{x}}, \tilde{t})$ is very close to a branch in (39). The point $(\tilde{\mathbf{x}}, \tilde{t})$ is likely to be closer to a solution branch in (39) than to other 17 branches. If so, the iteration

$$(\mathbf{x}_{j+1},\, t_{j+1}) \;\; = \;\; (\mathbf{x}_j,\, t_j) - \mathbf{f}_{\mathbf{x}t}(\mathbf{x}_j, t_j)^{\dagger}_{\text{rank-4}}\, \mathbf{f}(\mathbf{x}_j, t_j), \qquad j \,=\, 0, 1, \ldots$$

starting from the initial iterate $(\mathbf{x}_0,\, t_0) = (\tilde{\mathbf{x}}, \tilde{t})$ converges roughly to the nearest point in the branch in (39) on the plane $t = 1$ in $\mathbb{R}^4 \times \mathbb{R}$ by the geometric interpretation elaborated in §6. This is significant because the sequence $\{t_j\}$ converges to the important bifurcation point $t_* = 1$. This observation can easily be confirmed by NACLAB `Newton` and obtains a solution $(x_1, x_2, x_3, x_4,\, t)$ as

```
(0.822879063773473, 1.215245403637205, -0.822879063773474, -1.215245403637204, 1.000000000000000)
```

that is accurate with at least 15 digits, particularly the bifurcation point at $t = 1.0$. The residual $\left\|\mathbf{f}(\mathbf{x}_j, t_j)\right\|_2$ reduces from 10^{-4} to 4.44×10^{-16}, indicating convergence to a solution of $\mathbf{f}(\mathbf{x}, t) = \mathbf{0}$. Also notice the $\mathbf{x}$ component is identical to nearest point $\hat{\mathbf{x}}$ for at least 15 digits, confirming the geometric interpretation in §6 again.

8.2 Numerical greatest common divisor

For a polynomial pair p and q of degrees m and n respectively with a greatest common divisor (GCD) u of degree k along with cofactors v and w, the GCD problem can be modeled as a zero-finding problem

$$\begin{aligned} \mathbf{f}(u, v, w) \;\; &= \;\; (0, 0) \;\in\; \mathbb{P}_m \times \mathbb{P}_n \\ &\qquad \text{for} \quad (u, v, w) \;\in\; \mathbb{P}_k \times \mathbb{P}_{m-k} \times \mathbb{P}_{n-k} \end{aligned} \tag{40}$$

with the holomorphic mapping

$$\begin{aligned} \mathbf{f} \;\; : \;\; \mathbb{P}_k \times \mathbb{P}_{m-k} \times \mathbb{P}_{n-k} \;\; &\longrightarrow \;\; \mathbb{P}_m \times \mathbb{P}_n \\ (u, v, w) \;\; &\longmapsto \;\; (u\, v - p,\; u\, w - q) \end{aligned} \tag{41}$$

where $\mathbb{P}_m$ is the vector space of polynomials with degrees up to m, etc. Through standard bases of monomials, the mapping $\mathbf{f}$ in (41) can be represented by a mapping

from $\mathbb{C}^{m+n-k+3}$ to $\mathbb{C}^{m+n+2}$. Let $(u_*,\, v_*,\, w_*)$ denote a particular solution of (40). The general solution of (40) is a 1-dimensional set

$$\left\{ \left(t\, u_*,\, \tfrac{1}{t}\, v_*,\, \tfrac{1}{t}\, w_*\right) \;\middle|\; t \in \mathbb{C} \setminus \{0\} \right\} \tag{42}$$

on which the Jacobian is rank-deficient. Adding one auxiliary equation, however, the Jacobian of the modified mapping becomes injective [31], implying the Jacobian of $\mathbf{f}$ in (41) is of rank deficient by one columnwise. As a result, the set (42) consists of regular zeros of the mapping $\mathbf{f}$ in (41) and the iteration (19) is locally quadratically convergent by setting the projection rank

$$r \;=\; (k+1) + (m-k+1) + (n-k+1) - 1 \;=\; m+n-k+2.$$

Extra equations are not needed. Although there are infinitely many solutions forming a 1-dimensional set, there is practically no difference in finding anyone over the other.

Example 5 (Numerical greatest common divisor) For a simple demo of the iteration (19) and its NACLAB implementation `Newton`, let

$$p \;=\; -1.3333 - 2.3333\, x - 4\, x^2 - 3.6667\, x^3 - 2.6667\, x^4 - x^5, \quad \text{and} \quad q \;=\; -1.9999 + x + x^2 + 3\, x^3 \tag{43}$$

that are considered perturbed data of an underlying polynomial pair with a GCD $1 + x + x^2$. The process of identifying the GCD degree and computing the initial approximations of the GCD along with the cofactors can be found in [31]. With NACLAB installed, the following sequence of Matlab statements carries out the rank-8 Newton's iteration

$$(u_{j+1},\, v_{j+1},\, w_{j+1}) \;=\; (u_j,\, v_j,\, w_j) - J_{\text{rank-8}}(u_j,\, v_j,\, w_j)^{\dagger}\, \mathbf{f}(u_j,\, v_j,\, w_j)$$

for $j \,=\, 0, 1, \ldots$.

```
>> p = '-1.3333-2.3333*x-4*x^2-3.6667*x^3-2.6667*x^4-x^5';          % enter p as a character string
>> q = '-1.9999+x+x^2+3*x^3';                                        % enter q similarly
>> f = ...                     % enter the mapping as function handle for (u,v,w) -> (u*v-p, u*w-q)
      @(u,v,w,p,q) {pminus(ptimes(u,v),p),pminus(ptimes(u,w),q)};
>> J = ...    % enter the Jacobian J at (u0,v0,w0) as the mapping (u,v,w) -> (u0*v+u*v0, u0*w+u*w0)
       @(u,v,w,u0,v0,w0,p,q){pplus(ptimes(u0,v),ptimes(u,v0)),pplus(ptimes(u0,w),ptimes(u,w0))};
>> domain = {'1+x+x^2','1+x+x^2+x^3', '1+x'};      % representation of the domain for the mapping f
>> parameter = {p,q};                                              % parameters for the mapping f
>> u0 = 'x^2+1.4*x+1.6';  v0 = '-1.5-x-1.6*x^2-x^3'; w0 = '-2+2.8*x';          % initial (u0,v0,w0)
>> [z,res,fcond] = Newton({f,domain,parameter},{J,8},{u0,v0,w0},1);         % Newton on f with J in
                    % rank-8 projection from the initial iterate (u0,v0,w0) with display setting 1
Step  0:  residual = 1.5e+00                  1.6+14.*x+x^2
Step  1:  residual = 1.1e-01  shift = 4.9e-01  1.114771189108 + 1.114523702576*x + 1.088690420621*x^2
Step  2:  residual = 1.2e-03  shift = 5.9e-02  1.089839864820 + 1.090023690710*x + 1.089788605779*x^2
Step  3:  residual = 8.4e-06  shift = 1.0e-03  1.089756319215 + 1.089767203330*x + 1.089783432095*x^2
Step  4:  residual = 8.3e-06  shift = 1.4e-07  1.089756333892 + 1.089767171466*x + 1.089783428226*x^2
Step  5:  residual = 8.3e-06  shift = 5.1e-13  1.089756333892 + 1.089767171466*x + 1.089783428226*x^2
Step  6:  residual = 8.3e-06  shift = 7.0e-15  1.089756333892 + 1.089767171469*x + 1.089783428226*x^2
```

The computed GCD $1.08976 + 1.08977\, x + 1.08978\, x^2$ is of the scaling independent distance 1.02×10^{-5} that is in the same order of the data error 2.41×10^{-5}.

Similar to Example 4, the system with inexact data for (p, q) does not have a solution as the residual $\|\mathbf{f}(u_j, v_j, w_j)\|_2$ can not be reduced below 8.0×10^{-6} but the shift $\|(u_{j+1}, v_{j+1}, w_{j+1}) - (u_j, v_j, w_j)\|_2$ reduces to near unit roundoff, implying $J_{\text{rank-8}}(u_j, v_j, w_j)^\dagger \, \mathbf{f}(u_j, v_j, w_j)$ approaches zero.

8.3 Accurate computation of defective eigenvalues

Accurate computation of defective eigenvalues requires regularization due to hypersensitivities to data perturbations [32]. Let $\hat{\lambda} \in \mathbb{C}$ be an eigenvalue of $A \in \mathbb{C}^{n\times n}$ with what we call a multiplicity support $m \times k$, namely the geometric multiplicity and the smallest Jordan block size are m and k respectively. The problem of computing $\hat{\lambda}$ can be naturally modeled as solving the equation

$$\begin{aligned} \mathbf{g}(\lambda,\, X,\, A) \;&=\; O \;\in\; \mathbb{C}^{n\times k} \\ &\text{for} \quad (\lambda, X) \;\in\; \mathbb{C} \times \mathbb{C}^{n\times k} \end{aligned} \tag{44}$$

where

$$\begin{aligned} \mathbf{g} \;:\; \mathbb{C} \times \mathbb{C}^{n\times k} \times \mathbb{C}^{n\times n} \;&\longrightarrow\; \mathbb{C}^{n\times k} \\ (\lambda,\, X,\, G) \;&\longmapsto\; (G - \lambda\, I)\, X - X\, S \end{aligned} \tag{45}$$

with a constant nilpotent upper triangular matrix parameter $S \in \mathbb{C}^{k\times k}$ of rank $k-1$. Let $(\hat{\lambda}, \hat{X})$ be a particular solution of (44) and $\mathcal{L} : X \mapsto (A - \hat{\lambda} I)\, X - X\, S$. The kernel $\mathit{Kernel}(\mathcal{L})$ is of dimension $m\, k$ and can be spanned by $X_1, \ldots, X_{mk} \in \mathbb{C}^{n\times k}$. Then, in a neighborhood of $(\hat{\lambda}, \hat{X})$ in $\mathbb{C} \times \mathbb{C}^{n\times k}$, the solution of (44) is an $m\, k$-dimensional algebraic variety in the form of

$$\left\{ \left(\hat{\lambda}, X\right) \;\middle|\; X \;=\; \hat{X} + \alpha_1\, X_1 + \alpha_{mk}\, X_{mk}, \quad \alpha_1, \ldots, \alpha_{mk} \in \mathbb{C} \right\} \tag{46}$$

By [32, Lemma 2(ii)], the Jacobian $\mathbf{g}_{\lambda X}(\hat{\lambda}, \hat{X}, A)$ is of nullity $m\, k$ that is identical to the dimension of the solution set (46), implying the solution is regular.

Similar to the case study in §8.2, only a representative in the solution set (46) is needed. In [32], tedious auxiliary equations are imposed to isolate a solution point in the solution set (46). Those arbitrary extra equations that complicate the analysis and computation are now unnecessary in light of Theorem 1 and Theorem 2 that ensure a defective eigenvalue can be accurately computed by the rank-$(n\, k - m\, k + 1)$ Newton's iteration (19).

Example 6 (Defective eigenvalue computation) We demonstrate the effectiveness with a simple example where the matrix A and a perturbed version $\tilde{A} = A + E$ are as follows.

$$A \;=\; \begin{bmatrix} -1 & 0 & 3 & 0 & 2 & 1 \\ 1 & 1 & -1 & 1 & 0 & 0 \\ -2 & -1 & 4 & 1 & 1 & 0 \\ 3 & -3 & -3 & 5 & -1 & -1 \\ -3 & 1 & 3 & -1 & 5 & 2 \\ 1 & 0 & -1 & 0 & -1 & 2 \end{bmatrix} \quad \text{and} \quad E \;=\; 10^{-6} \begin{bmatrix} .1 & -.7 & -.4 & -1.0 & .2 & .6 \\ -.2 & .1 & -.1 & -.5 & .5 & .0 \\ .3 & -.8 & -.6 & -.1 & .4 & .1 \\ -.5 & .0 & .1 & .7 & -.2 & .5 \\ -.2 & -.2 & -.8 & -.7 & -.4 & -.5 \\ -.2 & -.1 & .8 & -.5 & -.7 & -.6 \end{bmatrix}$$

Matrix $A \in \mathbb{C}^{6\times 6}$ possesses an exact eigenvalue $\lambda_* = 3$ with the multiplicity 2×2 (c.f. [32] for identifying multiplicity supports). At the fixed $\lambda_0 = 2.9$, we can solve the linear system $\mathbf{g}(\lambda_0, X, A) = O$ for the least squares solution $X = X_0$ and obtain the initial iterate (λ_0, X_0). With A and E above entered in Matlab, we apply the rank-r Newton's iteration (19) with $r = (n-m)\,k+1 = 9$ by executing the NAClab module `Newton` in the following calling sequence.

```
>> g = @(lambda,X,G,S) G*X-lambda*X-X*S;              % enter the mapping g as a function handle
>> J = @(lambda,X,lambda0,X0,G,S) G*X-lambda*X0-lambda0*X-X*S;              % enter the Jacobian
>> domain = {1,ones(n,k)};                      % representation of the domain for the mapping g
>> parameter = {A,S};                                                % parameters of the mapping g
>> [z,res,fcond] = ...           % Call Newton on the mapping g with J in rank-9 projection from
   Newton({g,domain,parameter},{J,9},{lambda0,X0},2);        % (lambda0,X0) using display type 2

 Step  0:  residual = 2.5e-02                    2.900000000000000
 Step  1:  residual = 4.1e-03  shift = 1.0e-01   3.000493218848026
 Step  2:  residual = 3.0e-06  shift = 6.5e-03   2.999999313752690
 Step  3:  residual = 2.6e-12  shift = 4.1e-06   3.000000000001909
 Step  4:  residual = 4.4e-16  shift = 3.2e-12   3.000000000000000
 Step  5:  residual = 2.2e-16  shift = 5.1e-16   3.000000000000000
```

The λ components of the iteration accurately converges to the defective eigenvalue $\lambda_* = 3$ at roughly the quadratic rate with an accuracy at the order of the unit roundoff.

We simulate practical computation with imperfect empirical data by using the perturbed matrix $\tilde{A} = A + E$ in the iteration

$$\big(\tilde{\lambda}_{j+1},\, \tilde{X}_{j+1}\big) \;=\; \big(\tilde{\lambda}_j,\, \tilde{X}_j\big) - J_{\text{rank-9}}\big(\tilde{\lambda}_j,\, \tilde{X}_j,\, A+E\big)^{\dagger}\, \mathbf{g}\big(\tilde{\lambda}_j,\, \tilde{X}_j,\, A+E\big).$$

The iteration reaches the numerical solution $(\tilde{\lambda}, \tilde{X})$ where $\tilde{\lambda} = 3.00000102$, with a forward accuracy 1.02×10^{-6} about the same as the data perturbation $\|E\|_2 \approx 1.94\times 10^{-6}$.

Example 7 (Nearest matrix with the defective eigenvalue) Furthermore, the iteration (19) can again be applied to refine the eigenvalue computation by solving the equation

$$\mathbf{g}(\lambda,\, X,\, G) \;=\; O \quad \text{for} \quad (\lambda,\, X,\, G) \,\in\, \mathbb{C}\times\mathbb{C}^{n\times k}\times\mathbb{C}^{n\times n} \tag{47}$$

in the specific iteration

$$\big(\hat{\lambda}_{j+1},\, \hat{X}_{j+1},\, \hat{A}_{j+1}\big) \;=\; \big(\hat{\lambda}_j,\, \hat{X}_j,\, \hat{A}_j\big) - J_{\text{rank-12}}\big(\hat{\lambda}_j,\, \hat{X}_j,\, \hat{A}_j\big)^{\dagger}\, \mathbf{g}\big(\hat{\lambda}_j,\, \hat{X}_j,\, \hat{A}_j\big) \tag{48}$$

for $j = 0, 1, \ldots,$ starting from $\big(\tilde{\lambda}_0,\, \tilde{X}_0,\, \hat{A}_0\big) = (\tilde{\lambda},\, \tilde{X},\, A+E)$. The projection rank $r = n\,k = 12$ because, by [32, Lemma 2(ii)], the (full) Jacobian of $\mathbf{g}$ is surjective, implying the solution is regular with dimension $(1 + n\,k + n^2) - n\,k \;=\; n^2 + 1$.

In this example, the refinement (48) needs only 1 step to reduce the residual of the

equation (47) from 2.99×10^{-7} to 1.17×10^{-15}. The iterate terminates at $(\hat{\lambda},\, \hat{X},\, \hat{A})$ with $\hat{\lambda} \approx 3.000000103$, and the third component $\hat{A}$ is the matrix with an exact defective eigenvalue $\hat{\lambda}$ of multiplicity support $m \times k$, implying the backward error is $\|\tilde{A} - \hat{A}\| \approx 7.59 \times 10^{-7}$. From the geometric interpretation elaborated in §6, the iteration (48) converges to $(\hat{\lambda},\, \hat{X},\, \hat{A})$ that is roughly the nearest point in the solution set of (47) from the initial iterate $(\tilde{\lambda},\, \tilde{X},\, A + E)$.

9 A note on computing ultrasingular zeros

As formulated in §3, ultrasingularity occurs if the nullity of the Jacobian is higher at a zero than the dimension of the zero. Difficulties arise in computing ultrasingular zeros including slow convergence rate of iterative methods (c.f. [8]) and, more importantly, barriers of low attainable accuracy [22, 30]. As pointed out in [14], "*direct* attempts at such computation may easily fail or give inconclusive results". Quadratic convergence rate of Newton's iteration at regular zeros is not expected at ultrasingular zeros. At an isolated ultrasingular zero $\mathbf{x}_*$ of a smooth mapping $\mathbf{f} : \Omega \subset \mathbb{F}^m \to \mathbb{F}^n$ where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, the Jacobian $J(\mathbf{x}_*)$ is of rank $r < m$. At k-th step of conventional Newton's iteration (1) or the Gauss-Newton iteration when $n > m$, the Jacobian $J(\mathbf{x}_k)$ is usually highly ill-conditioned so that the computation of the iterate $\mathbf{x}_{k+1}$ is generally inaccurate, substantially limiting the attainable accuracy of the computed zero even if the iteration converges.

A *depth-deflation* strategy [7] deflates the ultrasingularity and transforms the zero $\mathbf{x}_*$ into a component of a regular zero $(\mathbf{x}_*, \mathbf{y}_*)$ of an expanded mapping

$$\mathbf{g} \;:\; (\mathbf{x}, \mathbf{y}) \;\mapsto\; \big(\; \mathbf{f}(\mathbf{x}),\, J(\mathbf{x})\,\mathbf{y},\, R\,\mathbf{y} - \mathbf{e} \;\big) \tag{49}$$

where R is a random $(m - r) \times m$ matrix and $\mathbf{e} \neq \mathbf{0}$. If $(\mathbf{x}_*, \mathbf{y}_*)$ is still an ultrasingular zero of $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{g}(\mathbf{x}, \mathbf{y})$, the deflation process can be continued recursively. It is proved in [7] that the number of deflation steps is bounded by the so-called *depth* of an ultrasingular isolated zero. In practice, however, one deflation step is likely to be enough except for the cases where the *breadth* $\mathit{nullity}\,(\,J(\mathbf{x}_*)\,) = 1$. At the terminating step of depth deflation, the ultrasingular zero $\mathbf{x}_*$ of $\mathbf{f}$ is a component of the regular zero of the final expanded mapping. As a result, the Gauss-Newton iteration locally converges at quadratic rate. More importantly, the zero $\mathbf{x}_*$ can be computed with an accuracy proportional to the data precision or unit round-off, circumventing the barrier of the perceived attainable accuracy at ultrasingular zeros. An earlier deflation strategy in [18] is proven to terminate with the number of steps bounded by the multiplicity. A so-called strong deflation method in symbolic-numerical computation is proposed in [13] and also proved to terminate in finitely many steps.

The deflation strategy applies to systems at ultrasingular nonisolated solutions as well. We illustrate the deflation process in the following examples.

Example 8 (Isolated ultrasingularity in a nonisolated zero set) In the cyclic-4 system in Example 3, the mapping $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, 1)$ as in (38) possesses 8 ultrasingular zeros embedded in the two solution curves (39). Those 8 points are nonisolated zeros with isolated ultrasingularity since each point is a unique ultrasingular zero in a small open neighborhood. For instance, the point $\mathbf{x}_* = (1, -1, -1, 1)$ is such an ultrasingular zero at which $\mathit{rank}\,(\,\mathbf{f}_{\mathbf{x}}(\mathbf{x}_*, 1)\,) = 2$. Interestingly, for almost all $R \in \mathbb{R}^{2\times 4}$, there is a unique $\mathbf{y}_* \in \mathbb{R}^4$ such that $(\mathbf{x}_*, \mathbf{y}_*)$ is a *regular* isolated zero of the expanded system $(\mathbf{f}(\mathbf{x}, 1), \mathbf{f}_{\mathbf{x}}(\mathbf{x}, 1)\,\mathbf{y}, R\,\mathbf{y} - (1, 0))$. In other words, the deflation method applies to isolated ultrasingularity at nonisolated zeros as well and the Gauss-Newton iteration converge quadratically to $(\mathbf{x}_*, \mathbf{y}_*)$. Even if the system is given with imperfect empirical data, the Gauss-Newton iteration on the expanded system still converges linearly to a stationary point that approximates $(\mathbf{x}_*, \mathbf{y}_*)$ with an accuracy at the same order of the data. The depth deflation method being applicable to nonisolated solutions of isolated ultrasingularities has apparently not been observed before. The theoretical termination and the bound on the number of deflation steps are still unknown.

Example 9 (An ultrasingular branch of zeros) There are nonisolated zeros where the entire branch is ultrasingular. Let $\mathbf{x} = (x_1, x_2, x_3, x_4)$ and the mapping

$$\mathbf{f}(\mathbf{x}) \;=\; \begin{pmatrix} x_1^3 + x_2^2 + x_3^2\,x_4^2 - 1 \\ x_1^2 + x_2^3 + x_3^2\,x_4^2 - 1 \\ x_1^2 + x_2^2 + x_3^3\,x_4^3 - 1 \end{pmatrix}$$

with a 1-dimensional solution curve $S = \{\mathbf{x} = (0, 0, s, 1/s) \mid s \neq 0\}$ on which the Jacobian $J(\mathbf{x})$ is of rank 1. All the solutions in S are ultrasingular due to $\mathit{nullity}\,(\,J(\mathbf{x})\,) = 3 \neq 1 = \mathit{dim}_{\mathbf{f}}(\mathbf{x})$ on S. As as result, the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ is underdetermined since, for every $\mathbf{x} \in S$ and almost all matrix $R \in \mathbb{R}^{3\times 4}$, there is a unique $\mathbf{y} \in \mathbb{R}^4$ such that $(\mathbf{x}, \mathbf{y})$ is a 1-dimensional zero of the expanded mapping $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{g}(\mathbf{x}, \mathbf{y})$ in (49) where $\mathbf{e}$ can be any nonzero vector, say $(1, 0, 0)$. The actual rank-7 Newton's iteration

$$(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \;=\; (\mathbf{x}_k, \mathbf{y}_k) - \mathbf{g}_{\mathbf{xy}}(\mathbf{x}_k, \mathbf{y}_k)^{\dagger}_{\text{rank-7}}\,\mathbf{g}(\mathbf{x}_k, \mathbf{y}_k), \quad k = 0, 1, \ldots$$

can be carried out using the following NACLAB calling sequence.

```
>> P = {'x1^3+x2^2+x3^2*x4^2-1'; 'x1^2+x2^3+x3^2*x4^2-1'; 'x1^2+x2^2+x3^3*x4^3-1'};   % enter system
>> x = {'x1';'x2';'x3';'x4'}; J = pjac(P,x);        % variable name array x, Jacobian of P w.r.t.  x
>> y = ['y1';'y2';'y3';'y4']; R = Srand(3,4);         % expanded variable array y, random 3x4 matrix
>> F = [P; ptimes(J,y); pminus(ptimes(R,y),[1;0;0])]; K = pjac(F,[x;y]);   % new system and Jacobian
>> g = @(u,v,F,K,x,y) peval(F,[x;y],[u;v]);            % function handle for expanded mapping g(x,y)
>> gjac = @(u,v,u0,v0,F,K,x,y) peval(K,[x;y],[u0;v0])*[u;v];  % function handle for expanded Jacobian
>> u0 = [0.001; 0.003; 0.499; 2.002];                    % an initial estimate of the solution
>> [~,~,V] = svd(peval(J,v,u0)); v0 = V(:,2:4)*((R*V(:,2:4))\[1;0;0]);          % new initial iterate
>> [Z,res,fcd] = Newton({g,{ones(4,1),ones(4,1)},{F,K,x,y}}, {gjac,7},{u0,v0}, 1)     % rank-7 Newton

 Step  0:  residual = 6.2e-03
 Step  1:  residual = 1.5e-05  shift = 3.0e-03
```

```
Step  2:  residual = 1.2e-09  shift = 2.4e-05
Step  3:  residual = 2.2e-16  shift = 1.6e-09
Step  4:  residual = 4.4e-16  shift = 8.2e-16
```

The step-by-step `shift` and `residual` show both $\|\mathbf{g}_{\mathbf{xy}}(\mathbf{x}_k, \mathbf{y}_k)^{\dagger}_{\text{rank-7}}\, \mathbf{g}(\mathbf{x}_k, \mathbf{y}_k)\|_2$ and $\|\mathbf{g}(\mathbf{x}_k, \mathbf{y}_k\|_2$ approach zero, implying the limit $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is both a stationary point and a zero of $\mathbf{g}$. The first component of the output 1×2 cell array `Z` is the computed zero

$$\hat{\mathbf{x}} \;=\; (\texttt{0.000000000000000},\ \texttt{0.000000000000000},\ \texttt{0.499435807628269},\ \texttt{2.002259318867864})$$

that is accurate to the 16th digits with residual near the unit roundoff. The Jacobian $\mathbf{g}_{\mathbf{xy}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is of nullity 1 that is identical to the dimension of the zero $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ to the expanded mapping $\mathbf{g}$. Namely, the ultrasingularity of the mapping $\mathbf{f}$ is deflated by expanding it to $\mathbf{g}$, resulting in a regular zero curve in $\mathbb{R}^4 \times \mathbb{R}^4$ whose first component is the zero curve of $\mathbf{f}$.

At the current stage, theories of the deflation approach for ultrasingular nonisolated zeros are still in development. It is proved in [13] that the Hauenstein-Wampler strong-deflation method terminates in finite number of steps. We propose a conjecture: *Assume $\mathbf{x}_*$ is a k-dimensional ultrasingular zero of an analytic mapping $\mathbf{f}$ whose Jacobian $J(\mathbf{x})$ maintains a constant nullity $n > k$ for all $\mathbf{x} \in \Omega \cap \mathbf{f}^{-1}(\mathbf{0})$ where Ω is an open neighborhood of $\mathbf{x}_*$. Then the recursive deflation process* (49) *terminates in finitely many steps so that $\mathbf{x}_*$ is a component of a regular k-dimensional zero of the final expanded system.*

References

[1] J. Backelin. Square multiples n give infinitely many cyclic n-roots. Reports, Matematiska Institutionen 8, Stockholms universitet, 1989.

[2] D. J. Bates, A. J. Sommese, J. D. Hauenstein, and C. W. Wampler. *Numerical Solving Polynomial Systems with Bertini*. SIAM, Philadelphia, 2013.

[3] A. Ben-Israel. A Newton-Raphson method for the solution of systems of equations. *J. Math. Anal. Appl.*, 15:243–252, 1966.

[4] P. T. Boggs. The convergence of the Ben-Israel iteration for nonlinear least squares problems. *Math. Comp.*, 30:512–522, 1976.

[5] X. Chen, M. Z. Nashed, and L. Qi. Convergence of Newton's method for singular smooth and nonsmooth equations using adaptive outer inverses. *SIAM J. Optim.*, 7:445–462, 1997.

[6] M. T. Chu. On a numerical treatment for the curve-tracing of the homotopy method. *Numer. Math.*, 42:323–329, 1983.

[7] B. Dayton, T.-Y. Li, and Z. Zeng. Multiple zeros of nonlinear systems. *Mathematics of Computation*, 80:2143–2168, 2011. DOI. 10.1090/S0025-5718-2011-02462-2.

[8] D. W. Decker, H. B. Keller, and C. T. Kelley. Convergence rate for Newton's method at singular points. *SIAM J. Numer. Anal*, pages 296–314, 1983. DOI. 10.1137/0720020.

[9] J.-P. Dedieu and M.-H. Kim. Newton's method for analytic systems of equations with constant rank derivatives. *J. Complexity*, 18:187–209, 2002.

[10] P. Deuflhard. A short history of Newton's method. *Documenta Mathematica*, Extra Volume:25–30, 2012.

[11] H. Führ and Z. Rzeszotnik. On biunimodular vectors for unitary matrices. *Linear Algebra and its Applications*, 484:86–129, 2015.

[12] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore and London, 4th edition, 2013.

[13] J. D. Hauenstein and C. W. Wampler. Isosingular sets and deflation. *Found. of Comput. Math.*, 13:371–403, 2013. DOI: 10.1007/s10208-013-9147-y.

[14] H. B. Keller. Geometrically isolated nonisolated solutions and their approximation. *SIAM J. Numer. Anal.*, 18:822–838, 1981. DOI. 10.1137/0718056.

[15] N. Kollerstrom. Thomas Simpson and 'Newton's method of approximation': an enduring myth. *Brit. J. Hist. Sci.*, 25:347–354, 2012.

[16] T.-L. Lee, T.-Y. Li, and Z. Zeng. A rank-revealing method with updating, downdating and applications, Part II. *SIAM J. Matrix Anal. Appl.*, 31:503–525, 2009. DOI. 10.1137/07068179X.

[17] Y. Levin and A. Ben-Israel. A Newton's method for systems of m equations in n variables. *Nonlinear Anal.*, 47:1961–1971, 2001.

[18] A. Leykin, J. Verschelde, and A. Zhao. Newton's method with deflation for isolated singularities of polynomial systems. *Theoretical Computer Science*, pages 111–122, 2006.

[19] T.-Y. Li and Z. Zeng. A rank-revealing method with updating, downdating and applications. *SIAM J. Matrix Anal. Appl.*, 26:918–946, 2005. DOI. 10.1137/S0895479803435282.

[20] M. Z. Nashed and X. Chen. Convergence of Newton-like methods for singular operator equations using outer inverses. *Numer. Math.*, 66:235–257, 1993.

[21] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. SIAM Publ., Philadelphia, 2000. First published by Academic Press, New York and London, 1977.

[22] V. Y. Pan. Solving polynomial equations: Some history and recent progress. *SIAM Review*, 39:187–220, 1997.

[23] A. J. Sommese and C. W. Wampler. *The Numerical Solution of Systems of Polynomials*. World Scientific Pub., Hackensack, NJ, 2005.

[24] G. W. Stewart. On the perturbation of pseudo-inverses, projections, and linear least squares problems. *SIAM Review*, 19:634–662, 1977.

[25] G. W. Stewart. *Matrix Algorithms, Volume I, Basic Decompositions.* SIAM, Philadelphia, 1998.

[26] G. W. Stewart and J. Sun. *Matrix Perturbation Theory.* Academic Press, Inc, Boston, San Diego, New York, London, Sydney, Tokyo, Toronto, 1990.

[27] K. Tenabe. Continuous Newton-Raphson method for solving an underdetermined system of nonlinear equations. *Nonlinear Analysis, Theory, Methods and Applications*, 3:495–503, 1979.

[28] C. W. Wampler and A. J. Sommese. Numerical algebraic geometry and algebraic kinematics. *Acta Numerica*, 20:469–567, 2011. DOI. 10.1017/S0962492911000067.

[29] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12:99–111, 1972.

[30] T. J. Ypma. Finding a multiple zero by transformations and Newton-like methods. *SIAM Review*, 25:365–378, 1983.

[31] Z. Zeng. The numerical greatest common divisor of univariate polynomials. In J. R. L. Gurvits, P. Pébay and D. Thompson, editors, *Contemporary Mathematics Vol. 556, American Mathematical Society,* Randomization, Relaxation and Complexity in Polynomial Equation Solving, pages 187–217, Providence, RI, 2011.

[32] Z. Zeng. Sensitivity and computation of a defective eigenvalue. *SIAM J. Matrix Analysis and Applications*, 37(2):798–817, 2016. DOI. 10.1137/15M1016266.

[33] Z. Zeng. On the sensitivity of singular and ill-conditioned linear systems. *SIAM J. Matrix Anal. Appl.*, 40(3):918–942, 2019. DOI. 10.1137/18M1197990.