October 1987

UILU-ENG-87-2262
ACT-82

COORDINATED SCIENCE LABORATORY
College of Engineering
Applied Computation Theory

ON THE CONVERGENCE OF NEWTON'S METHOD FOR SOLVING SYSTEMS OF LINEAR EQUATIONS

Scot W. Hornick
Franco P. Preparata
Prasoon Tiwari

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS None
2a. SECURITY CLASSIFICATION AUTHORITY N/A	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) UILU-ENG-87-2262 (ACT-82)	5. MONITORING ORGANIZATION REPORT NUMBER(S) N/A
6a. NAME OF PERFORMING ORGANIZATION Coordinated Science Lab University of Illinois 6b. OFFICE SYMBOL (If applicable) N/A	7a. NAME OF MONITORING ORGANIZATION Semiconductor Research Corporation (IBM & RCA Doctoral Fellowships)
6c. ADDRESS (City, State, and ZIP Code) 1101 W. Springfield Avenue Urbana, IL 61801	7b. ADDRESS (City, State, and ZIP Code) P.O. Box 12053 Research Triangle Park, NC 27709
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Semiconductor Research Corporation, IBM, RCA 8b. OFFICE SYMBOL (If applicable) N/A	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER SRC 86-12-109
8c. ADDRESS (City, State, and ZIP Code) P.O. Box 12053 Research Triangle Park, NC 27709	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

On the Convergence of Newton's Method for Solving Systems of Linear Equations

12. PERSONAL AUTHOR(S)
Hornick, Scot W., Preparata, Franco P., and Tiwari, Prasoon

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	October 1987	22

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	attunement number, condition number, minimum-Euclidean-norm, least-squared-error solution of linear systems, Newton's method, parallel algorithms, roundoff error

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The notion of attunement of a linear system of equations is introduced and formalized. By defining the attunement number of a vector b to a matrix A for a given relative error ε, we relate the number of Newton iterations required to solve the linear system $Ax = b$ (within the prescribed relative error) to joint properties of the matrix and the known vector. Previous analyses of the convergence properties of Newton's method referred exclusively to the condition number, a property of the matrix alone. Using the attunement number, we show that Newton's method provides an efficient, polylog-time parallel algorithm for solving a much broader class of linear systems than those for which A is well-conditioned. An analysis of roundoff error reveals that these results remain valid in the context of finite-precision arithmetic provided that a reasonable number of bits are used.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) / 22c. OFFICE SYMBOL

ON THE CONVERGENCE OF NEWTON'S METHOD FOR SOLVING SYSTEMS OF LINEAR EQUATIONS

Scot W. Hornick, Franco P. Preparata, and Prasoon Tiwari*

Coordinated Science Laboratory and

Department of Electrical and Computer Engineering

University of Illinois

Urbana, IL 61801

Abstract: The notion of *attunement* of a linear system of equations is introduced and formalized. By defining the *attunement number* of a vector b to a matrix A for a given relative error ε, we relate the number of Newton iterations required to solve the linear system $Ax = b$ (within the prescribed relative error) to *joint* properties of the matrix and the known vector. Previous analyses of the convergence properties of Newton's method referred exclusively to the condition number, a property of the matrix alone. Using the attunement number, we show that Newton's method provides an efficient, polylog-time parallel algorithm for solving a much broader class of linear systems than those for which A is well-conditioned. An analysis of roundoff error reveals that these results remain valid in the context of finite-precision arithmetic provided that a reasonable number of bits are used.

Key words: attunement number; condition number; minimum-Euclidean-norm, least-squared-error solution of linear systems; Newton's method; parallel algorithms; roundoff error

---

* This author is presently at the IBM T. J. Watson Research Center, Yorktown Heights, NY.

This research was supported in part by the Semiconductor Research Corporation under contract 86-12-109, by an RCA Doctoral Fellowship, and by an IBM Doctoral Fellowship.

1. Introduction

The solution of systems of linear equations is a central problem in numerical mathematics with a number of extremely significant applications. Algorithms for this important problem are generally classified as either *direct* (e.g., Gauss-Jordan elimination, LU decomposition, and Cramer's rule) or *iterative* (e.g., Jacobi's method, the Gauss-Seidel method, and simultaneous overrelaxation). These algorithms can be further classified according to their suitability for solving systems with multiple right-hand sides (known vectors). In particular, algorithms that attempt to invert the system matrix can be used to solve systems with multiple known vectors, but they are often less efficient than those algorithms that exploit the specific case of a solitary known vector.

Newton's method for solving systems of linear equations iteratively computes an approximate inverse of the system matrix in a manner analogous to the Newton-Raphson iteration for computing the inverse of a scalar. The history of this quadratically convergent iterative method is a long one. First proposed by Schulz [Sc33] in 1933, this approach was later studied by Ben-Israel and Cohen [BC66] who found an appropriate initial approximation of the inverse and showed that the method can be used to compute pseudo-inverses (generalized inverses) too. Isaacson and Keller [IK66] also discussed the method but dismissed it as less attractive for sequential computation than the linearly convergent iterative methods (e.g., Jacobi's method, the Gauss-Seidel method, and simultaneous overrelaxation) that use only matrix-vector multiplication instead of the more complex matrix-matrix multiplication.

Newton's method was recently resurrected by Bojańczyk [B84] and Pan and Reif [PR85] in the context of parallel computation models. They realized that no time penalty is paid for using the quadratically convergent Newton's method since, in most models of parallel computation, the time complexity of matrix-matrix multiplication is of the same order as that of matrix-vector multiplication. Thus, for a well-conditioned $n \times n$ matrix A (i.e., $\kappa(A)$, the condition number of A, is polynomial in n) the PRAM algorithm of Pan and Reif (which uses an initial approximate inverse slightly different from that of [BC66] and that of [IK66]) obtains an accurate solution in time $O(\log^2 n)$ with $M(n)$ processors, where $M(n)$ is the number of processors needed to multiply two $n \times n$ matrices in $O(\log n)$ time. Prior to [B84] and [PR85] the only known polylog-time algorithm was that of Csanky [C76] (and the modified form of [PS78]), which is based on the Cayley-Hamilton theorem and Leverrier's method; however, barring arbitrarily accurate arithmetic, this method is numerically unstable [W65]. Newton's method, on the other hand, is suited to the real-world environment of finite precision [SS74].

Nevertheless, exclusive reference to the condition number in the upper bounds of [B84] and [PR85] seems excessively pessimistic. Indeed, even a system with a *singular* matrix A is solvable exactly if the known vector is in the range space of A. This elementary observation suggests that the attainability of an approximate solution rests not on a property of the matrix alone, but, rather, on joint properties of the matrix and the known vector.

Informally, an ill-conditioned matrix distorts the unit sphere into an ellipsoid with axes of markedly different lengths. Suppose, to aid intuition, that the ellipsoid axes can be broadly categorized as either "long" or "short". If the known vector projects mainly on the subspace spanned by the long axes, we would expect that an accurate solution is attainable with moderate effort. In this case, we say that the known vector is *attuned* to the matrix.

Formally, the major contribution of this paper is the introduction of the notion of *attunement number* $a(A, b, \varepsilon)$ of a vector b to a matrix A for relative error ε, which is the number of Newton iterations needed to reduce the relative error to ε. Well-attuned vectors b are those for which $a(A, b, \varepsilon)$ is logarithmic in n. For any fixed number l of iterations, the set of *l-attuned* vectors (i.e., those vectors for which satisfactory convergence is attained in l iterations) is a cone centered at the origin and containing in a suitably defined interior the ellipsoid axes corresponding to the largest singular values of A. Our main results are: (i) Newton's method converges in $O(\log n)$ iterations for a much broader class of linear systems than those for which A is well-conditioned; (ii) For a well-attuned b, the effects of roundoff errors only lengthen the duration of the process by constant factors.

2. Our Algorithm

In order to exploit the attunement of a given known vector, clearly we must not insist upon attaining an accurate approximation of A^{-1}, a least not when A is ill-conditioned. Instead, we terminate the iteration when Ax_l is an accurate approximation of b, where x_l is the current approximant of x. Therefore, our algorithm is just a simple modification of the basic Newton's method as discussed in [BC66], [B84], and [PR85]. It takes as input the $m \times n$ matrix A, the $m \times 1$ vector b, and the error criterion ε; it produces as output the $n \times 1$ vector x such that the relative error in the *solution of the linear system* (as opposed to the error in the pseudo-inverse) is no larger than the error criterion. In the event that such an x cannot be found (up to the precision used), the algorithm halts and indicates failure. It is given concisely in pidgin Algol as follows:

SOLVE (A, x, b, ε)
begin
$B := A^T A$;
$\alpha := \dfrac{1}{\max_i \sum_j |(B)_{ij}|}$;
$Y_0 := \alpha A^T$; (1)
$l := 0$;
repeat
$Y_{l+1} := 2Y_l - Y_l A Y_l$; (2)
$l := l + 1$;
$x_l := Y_l b$; (3)
$e_l := \dfrac{\| b - Ax_l \|_2}{\| b \|_2}$; (4)
until ($e_l \le \varepsilon$ **or** $e_{l-1} - e_l \le 0$)
$x := x_l$;
if ($e_l \le \varepsilon$) **then**
SOLVE := **success**;
else
SOLVE := **failure**;
end.

This algorithm is essentially the same as previous algorithms for computing the pseudo-inverse,[1] but it exploits the fact that the pseudo-inverse of A need not be computed to full accuracy in order to solve $Ax = b$. This idea is expounded upon in the next two sections, which analyze the operation and performance of this algorithm.

We also note that, in most models of parallel computation, the complexity of SOLVE is dominated by the two matrix multiplications that occur in each Newton iteration.[2] For such models, the overall time complexity is (on the order of) just the product of the number of iterations and the time required for a matrix multiplication. Although we will not dwell on issues of parallel complexity in this paper, this consideration motivates our study of the convergence of Newton's method.

3. Analysis of Newton Iteration Applied to A

In this section, we will derive a closed-form expression for the matrix Y_l and investigate the convergence properties of the sequence x_l. In order to facilitate this discussion, we first review the definition of the *singular value decomposition* (SVD) of the matrix A. The SVD of a matrix $A \in \mathbb{R}^{m \times n}$ is the representation of the matrix as the product of three matrices ($A = U \Sigma V^T$), where $U = [u_1, \cdots, u_m] \in \mathbb{R}^{m \times m}$ and $V = [v_1, \cdots, v_n] \in \mathbb{R}^{n \times n}$ (and both are unitary) and $\Sigma = \mathrm{diag}[\sigma_1, \cdots, \sigma_p] \in \mathbb{R}^{m \times n}$ ($p = \min\{m, n\}$ and $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_p \ge 0$). The σ_i are called the

[1] We have chosen the form of α given in [BC66], $\alpha = \| A^T A \|_\infty^{-1}$. An alternative form is $\alpha = [\mathrm{tr}(A^T A)]^{-1}$ [IK66] or $\alpha = [\| A \|_\infty \| A \|_1]^{-1}$ [PR85]. The reduction in the required number of iterations achieved by using the scaling factor of [BC66] instead of that of [IK66] or [PR85] more than compensates for the expense of the additional matrix multiplication.

[2] A variant of this algorithm reduces the number of matrix multiplications per iteration from two to one, which results in a constant factor improvement in the parallel complexity, but this variant is not the most appropriate for finite-precision computation.

singular values of A, and the vectors u_i and v_i are, respectively, the *i-th left singular vector* and the *i-th right singular vector* of A [GV83]. We can express the *pseudo-inverse* of A in terms of the SVD of A. Observing that $\sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$, where $r = \text{rank}(A)$, we define $\Sigma^+ \in \mathbb{R}^{n \times m}$ as $\text{diag}[\sigma_1^{-1}, \cdots, \sigma_r^{-1}, 0, \cdots, 0]$. The matrix $A^+ = V\Sigma^+ U^T$ is referred to as the *pseudo-inverse* of A [GV83].

While Newton's method does not explicitly compute the SVD of A, this tool is useful in the method's analysis; in the following lemma, we apply it, to derive a closed-form expression for Y_l.

Lemma 1: $Y_l = V\Delta_l U^T$, where Δ_l is a diagonal matrix given by

$$\delta_{ij}^{(l)} = (\Delta_l)_{ij} = \begin{cases} 0, & \text{if } i \neq j, \\ \sigma_i^{-1}[1 - (1 - \alpha\sigma_i^2)^{2^l}], & \text{if } i = j \text{ and } i \leq r, \\ 0, & \text{if } i = j \text{ and } i > r. \end{cases} \tag{5}$$

Proof (by induction on l):

Basis ($l = 0$): $Y_0 = \alpha A^T$ and $\Delta_0 = V^T Y_0 U = \alpha V^T V \Sigma^T U^T U = \alpha\Sigma^T$. Now, for $i \leq r$, $\alpha\sigma_i = \sigma_i^{-1}[1 - (1 - \alpha\sigma_i^2)] = \delta_{ii}^{(0)}$, and for $i > r$, $\alpha\sigma_i = 0 = \delta_{ii}^{(0)}$.

Inductive Step: Assume that the lemma holds for $l = k$. We now prove that it also holds for $l = k + 1$. By Equation (2) and the inductive hypothesis,

$$\begin{aligned} Y_{k+1} &= 2Y_k - Y_k A Y_k \\ &= 2V\Delta_k U^T - (V\Delta_k U^T)(U\Sigma V^T)(V\Delta_k U^T) \\ &= V[2\Delta_k - \Delta_k \Sigma \Delta_k] U^T. \\ &= V\Delta_{k+1} U^T, \end{aligned}$$

where the diagonal matrix $\Delta_{k+1} = 2\Delta_k - \Delta_k\Sigma\Delta_k$. Notice that the recurrence for Δ_l is precisely the Newton iteration for computing the inverse of Σ, if it exists. Since Σ is diagonal, the equations are decoupled, and it immediately follows from the initial conditions that $\delta_{ij}^{(l+1)} = 0$ for $i \neq j$, and, for $i = j$,

$$\delta_{ii}^{(l+1)} = 2\delta_{ii}^{(l)} - \delta_{ii}^{(l)}\sigma_i\delta_{ii}^{(l)} = \begin{cases} \sigma_i^{-1}[1 - (1 - \alpha\sigma_i^2)^{2^{l+1}}], & \text{if } \sigma_i \neq 0 \ (i \leq r), \\ 0, & \text{if } \sigma_i = 0 \ (i > r). \ \square \end{cases}$$

As Schulz noted [Sc33], Newton iteration for the inverse of a matrix converges only if all of the eigenvalues of $Y_0 A$ lie within the unit circle with origin $1 + 0\cdot\sqrt{-1}$ in the complex plane. In the case of SOLVE, all of the

eigenvalues of $\alpha A^T A = V(\alpha \Sigma^T \Sigma)V^T$ must be in the interval (0,2) to satisfy this condition (i.e., for all i, $\alpha\sigma_i^2 \in (0,2)$, the necessity of which can be observed also from (5)). Strict inequality in the lower bound is not required for computing the pseudo-inverse, but the upper bound is guaranteed by our choice of the scaling factor α. Therefore, as $l \to \infty$, Δ_l converges to Σ^+, and Y_l converges to A^+ [BC66,SS74]. Furthermore, since $x_l = Y_l b$, x_l converges to $A^+ b$, the minimum-Euclidean-norm, least-squared-error solution of $Ax = b$ [FM67].

4. Analysis of the Set of "Attuned" Vectors

While the above analysis guarantees the convergence of x_l, it gives no indication as to the number of iterations necessary to achieve a specified relative error in the solution of $Ax_l = b$. In this section, we derive an expression relating the error measure e_l to the number of iterations. From this expression, we first determine the number of iterations necessary to achieve the specified relative error in the worst case. Then, with a more careful analysis, we show that, for a known vector b which is "attuned" to the matrix A, an accurate solution can be obtained much more rapidly than in the pessimistic worst case. Finally, we derive an algebraic and geometric characterization of the vectors that are "attuned" to A, i.e., those b's such that $e_l \le \varepsilon$ for fixed l and ε. The notion of "attuned" vectors represents the principal contribution of this paper; prior to this, the convergence of Newton's method for solving a linear system has been related only to the condition number of the matrix.

In order to find an expression for e_l as a function of the number of iterations, let us define $\beta = U^T b$, i.e., $\beta = [\beta_1, \beta_2, \cdots, \beta_m]^T$ is the representation of b in the basis of the left singular vectors of A. From Equation (3) and Lemma 1,

$$\begin{aligned} Ax_l &= AY_l b \\ &= (U\Sigma V^T)(V\Delta_l U^T)U\beta \\ &= U\Sigma\Delta_l\beta. \end{aligned}$$

Manipulating Equation (4), we obtain:

$$e_l = \frac{\| b - Ax_l \|_2}{\| b \|_2} = \frac{\| U\beta - U\Sigma\Delta_l\beta \|_2}{\| U\beta \|_2} = \frac{\| (I - \Sigma\Delta_l)\beta \|_2}{\| \beta \|_2}. \qquad (6)$$

Let us define $\Omega_l = I - \Sigma\Delta_l = \mathrm{diag}[\omega_1^{(l)}, \omega_2^{(l)}, \cdots, \omega_m^{(l)}] \in \mathbb{R}^{m \times m}$, where

$$\omega_i^{(l)} = \begin{cases} (1 - \alpha\sigma_i^2)^{2^l}, & \text{if } i \le r, \\ 1, & \text{if } i > r, \end{cases} \tag{7}$$

to obtain the desired expression:

$$e_l = \frac{\| \Omega_l \beta \|_2}{\| \beta \|_2} = \left[\frac{\sum_{i=1}^{m} (\omega_i^{(l)} \beta_i)^2}{\sum_{i=1}^{m} \beta_i^2} \right]^{1/2} \tag{8}$$

With this expression for the relative error, we will determine the worst-case number of iterations necessary to achieve a specified relative error. To begin with, we restrict our attention to known vectors b in the range space $R(A)$ of A. Since the σ_i's form a nonincreasing sequence, the $\omega_i^{(l)}$'s form a nondecreasing sequence. Therefore, the relative error is maximized if $\beta_r = \| b \|_2$ and $\beta_i = 0$ for $i \neq r$. In this case,

$$e_l = \omega_r^{(l)} = (1 - \alpha\sigma_r^2)^{2^l}, \tag{9}$$

which exhibits the quadratic convergence typical of Newton's method.

With Equation (9) for the worst-case relative error, we can proceed to upper bound the required number of iterations by a function of $\kappa(A)$ and ε.

Theorem 1: For all $b \in R(A)$, there exists an $l^* = O(\log n + \log(\kappa(A)) + \log\log(1/\varepsilon))$ such that $e_l \le \varepsilon$ for all $l \ge l^*$.

Proof: We divide the iteration process into two regimes: the first l_1 iterations for which the relative error $e_i \ge 1/2$ ($i \le l_1$) and the next l_2 iterations for which the relative error $1/2 > e_i \ge \varepsilon$ ($l_1 < i \le l_1 + l_2$). By upper bounding l_1 and l_2 independently, we can upper bound the sum $l^* = l_1 + l_2$, such that $e_l \le \varepsilon$ if $l \ge l^*$.

For the first iteration regime, we consider the recurrence relation $e_{l+1} = e_l^2$, which describes the quadratic convergence of the relative error in Equation (9). Let us define $a_l = 1 - e_l$; $e_l \ge 1/2$ implies $a_l \le 1/2$, whence

$$\begin{aligned} e_{l+1} &= e_l^2 \\ 1 - a_{l+1} &= (1 - a_l)^2 \\ &= 1 - 2a_l + a_l^2 \\ 1 - a_{l+1} &\le 1 - (3/2)a_l \\ a_{l+1} &\ge (3/2)a_l. \end{aligned}$$

Thus, in the first regime, $e_l \le 1 - (3/2)^l a_0$. This bound is used until $1 - (3/2)^l a_0 < 1/2$, i.e.,

$$l > l_1 = \left\lfloor \frac{-\log a_0 - 1}{\log(3/2)} \right\rfloor = \left\lfloor \frac{\log(1/a_0) - 1}{\log(3/2)} \right\rfloor, \tag{10}$$

at which point $e_l < ½$, and we enter the second iteration regime. From Equation (9), we can see that $a_0 = \alpha\sigma_r^2$, and, by our choice of α (see, e.g. [St73]), $a_0 \geq \frac{\sigma_r^2}{\sqrt{n}\sigma_1^2} = \frac{1}{\sqrt{n}(\kappa(A))^2}$. Therefore,

$$l_1 < \left\lfloor \frac{½\log n + 2\log(\kappa(A))}{\log(3/2)} \right\rfloor. \tag{11}$$

For the second iteration regime, the recurrence remains the same, but now we may assume that $e_0 < ½$. Thus, $e_l < (½)^{2^l}$ and $e_l \leq \varepsilon$ if

$$l > l_2 = \lfloor \log\log(1/\varepsilon) \rfloor. \tag{12}$$

Adding the number of iterations from these two regimes, we verify that $l^* = O(\log n + \log(\kappa(A)) + \log\log(1/\varepsilon))$ iterations suffice to achieve the specified relative error. □

The preceding discussion of the convergence of the relative error was predicated on the restriction that $b \in R(A)$. Of course, if we eliminate this restriction, then the relative error is maximized if $b \in R(A)^C$ (i.e., b is in the orthogonal complement of the range space of A), which yields $e_l = 1$ for all l. Whenever $b \notin R(A)$, though, it is impossible to achieve an arbitrarily small specified relative error. In such cases, it may be more reasonable to consider the relative error

$$e_{l'} = \frac{\| b_A - Ax_l \|_2}{\| b \|_2}, \tag{13}$$

where b_A is the projection of b on $R(A)$. If we consider this error measure instead, the conclusion of Theorem 1 holds without the restriction that $b \in R(A)$.

Theorem 1 is essentially a generalized version of the result of Pan and Reif [PR85]. They considered only nonsingular matrices $r = m = n$ with $\kappa(A)$ polynomial in n and assumed an ε of the form 2^{-n^c}, where c is a constant. Their result follows as a corollary of Theorem 1, which indicates that $l^* = O(\log n)$ in such cases.

Newton's method is, however, more powerful than this worst-case analysis would indicate. Informally, a known vector b is "attuned" to A if $\beta = U^T b$ has most of its weight in its low-indexed components, i.e., those corresponding to large singular values of A. More formally, $e_l \leq \varepsilon$ is equivalent, by Equation (8), to

$$\sum_{i=1}^{m}[\varepsilon^2 - (\omega_i^{(l)})^2]\beta_i^2 \geq 0, \tag{14}$$

and we may quantify this notion by defining the *attunement number* $a(A, b, \varepsilon)$ of b to A for a given ε as

$$a(A, b, \varepsilon) = \min\{l : \sum_{i=1}^{m}[\varepsilon^2 - (\omega_i^{(l)})^2]\beta_i^2 \geq 0\}. \tag{15}$$

In other words, $a(A, b, \varepsilon)$ is just the number of iterations necessary to reduce the relative error to ε. Since the $\omega_i^{(l)}$'s form a nondecreasing sequence, $\varepsilon^2 - (\omega_1^{(l)})^2 \geq \varepsilon^2 - (\omega_2^{(l)})^2 \geq \cdots \geq \varepsilon^2 - (\omega_r^{(l)})^2 > \varepsilon^2 - (\omega_{r+1}^{(l)})^2 = \cdots = \varepsilon^2 - (\omega_m^{(l)})^2 = \varepsilon^2 - 1$, and the sequence of coefficients in (14) and (15) has at most one sign change. In particular, the ith coefficient changes from negative to positive when

$$\omega_i^{(l)} = \varepsilon$$

$$(1 - \alpha\sigma_i^2)^{2^l} = \varepsilon$$

$$2^l = \frac{\log\varepsilon}{\log(1 - \alpha\sigma_i^2)}$$

$$l = \left\lceil \log\left[\frac{\log\varepsilon}{\log(1 - \alpha\sigma_i^2)} \right] \right\rceil .$$

Theorem 1 dictates that, in the worst case, $a(A, b, \varepsilon) = O(\log n + \log(\kappa(A)) + \log\log(1/\varepsilon))$, but, as we will show in Theorem 2, if b is "attuned" to A, then $a(A, b, \varepsilon)$ may be $O(\log n)$ even if $\kappa(A)$ is not polynomial in n.

Equation (15) provides a concise algebraic definition of attunement number; geometric insight, however, is best attained by means of an example.

Example:

Consider the matrix

$$A = \begin{bmatrix} 8 & 2 & 20 \\ 19 & -14 & 10 \\ -2 & -2 & 1 \end{bmatrix}$$

with SVD

$$U\Sigma V^T = \begin{bmatrix} 3/5 & -4/5 & 0 \\ 4/5 & 3/5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 30 & 0 & 0 \\ 0 & 15 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2/3 & -1/3 & 2/3 \\ 1/3 & -2/3 & -2/3 \\ -2/3 & -2/3 & 1/3 \end{bmatrix}.$$

Then

$$B = A^T A = \begin{bmatrix} 429 & -246 & 348 \\ -246 & 204 & -102 \\ 348 & -102 & 501 \end{bmatrix},$$

which yields $\alpha = (429 + 246 + 348)^{-1} \approx 0.0009775$.

Note that the matrix A has condition number $\kappa(A) = 10$, and, therefore, the unit sphere is mapped into a flat ellipsoid by the transformation Ax (see Figure 1). By applying Equation (14) for $m = n = r = 3$, $a(A, b, \varepsilon) \le l$ if

$$[\varepsilon^2 - (\omega_1^{(l)})^2]\beta_1^2 + [\varepsilon^2 - (\omega_2^{(l)})^2]\beta_2^2 + [\varepsilon^2 - (\omega_3^{(l)})^2]\beta_3^2 \ge 0. \tag{16}$$

Let us assume $\varepsilon = .0001$ and see how the set of l-attuned known vectors evolves as we increase l.

Case 0 $\left(l < \left\lceil \log\left[\frac{\log\varepsilon}{\log(1 - \alpha\sigma_1^2)} \right] \right\rceil = 3\right)$: All of the coefficients in (16) are negative, so the only l-attuned β is the zero vector.

If we allow Newton's method to proceed, we reach

Case 1 $\left(3 = \left\lceil \log\left[\frac{\log\varepsilon}{\log(1 - \alpha\sigma_1^2)} \right] \right\rceil \le l < \left\lceil \log\left[\frac{\log\varepsilon}{\log(1 - \alpha\sigma_2^2)} \right] \right\rceil = 6\right)$: The first coefficient in (16) has become positive, but the other two remain negative, so the set of l-attuned β's forms the *interior* of an elliptic cone about the β_1 axis (see Figure 1).

Continuing further, we next reach

Case 2 $\left(6 = \left\lceil \log\left[\frac{\log\varepsilon}{\log(1 - \alpha\sigma_2^2)} \right] \right\rceil \le l < \left\lceil \log\left[\frac{\log\varepsilon}{\log(1 - \alpha\sigma_3^2)} \right] \right\rceil = 11\right)$: Now the second coefficient in (16) has also become positive, (although the third remains negative), so the set of l-attuned β's now forms the *exterior* of an elliptic cone about the β_3 axis (see Figure 1).

Continuing even further, we finally reach

Case 3 $\left(l \ge \left\lceil \log\left[\frac{\log\varepsilon}{\log(1 - \alpha\sigma_3^2)} \right] \right\rceil = 11\right)$: Now all coefficients in (16) are positive, so all β's are l-attuned, i.e., $l \ge l^*$, as defined in Theorem 1.

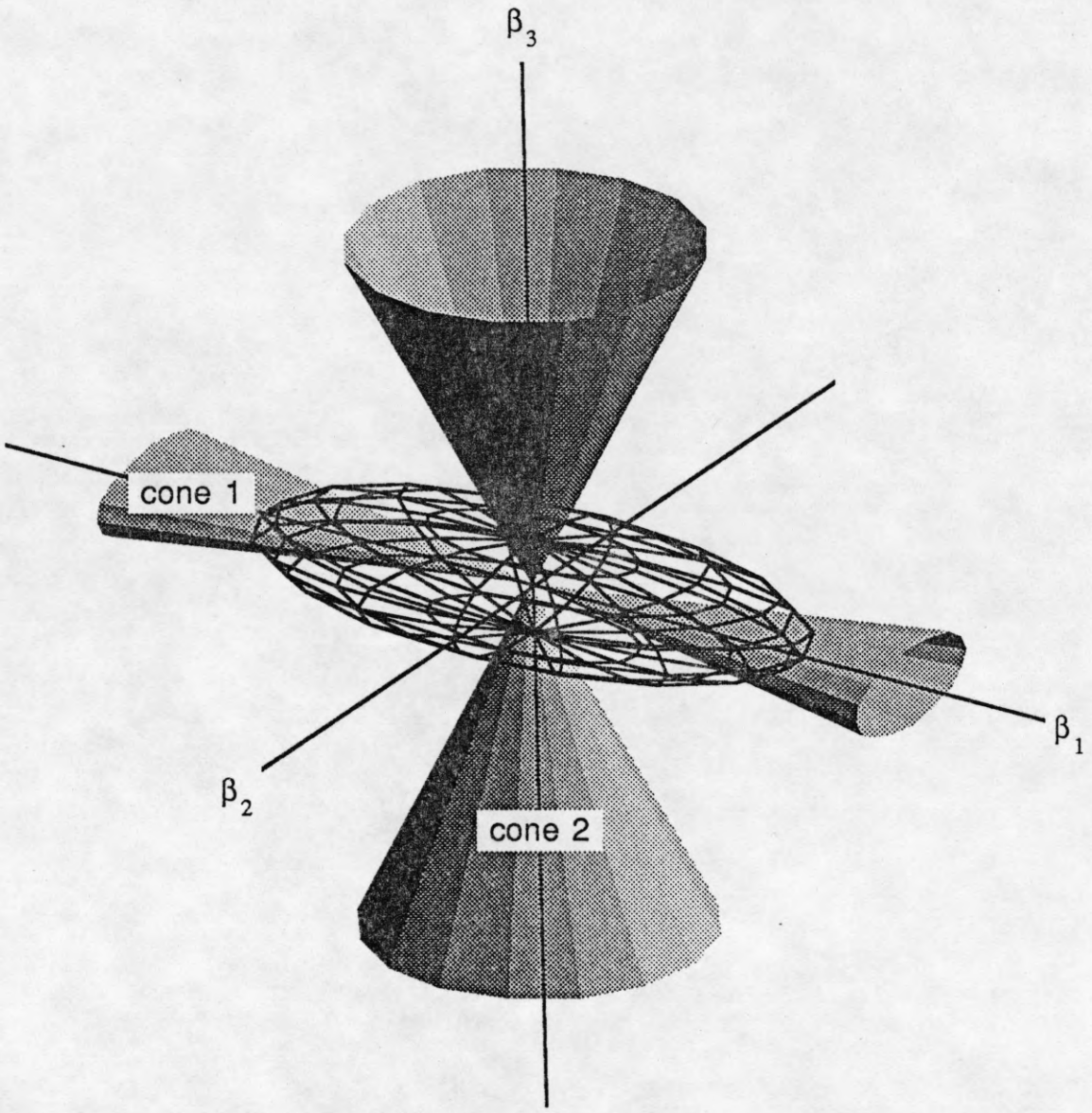

Figure 1. Image of the unit sphere under the linear transformation Ax and the set of l-attuned vectors for Case 1 (interior of cone 1) and Case 2 (exterior of cone 2)

Since $b = U\beta$ is a simple rotation, the set of l-attuned b's has the same shape and size as the set of β's in each of the above cases. For spaces of dimension greater than three, the algebraic and geometric description of this set of vectors easily generalizes, although it is impossible to visualize the elliptic hypercones that bound the set. Also, observe that the process becomes "stuck" in Case r if matrix A has rank $r < m$. In such situations, SOLVE eventually terminates because $e_{l-1} - e_l \leq 0$ (with finite-precision arithmetic).

Now that we have formally defined the attunement number, we show how this concept can be used to make statements regarding the asymptotic complexity of solving linear systems. The underlying intuition is that if the known vector does not project too much on the "unattuned" subspace of $R(A)$ and $R(A)^C$, then it should not be very difficult to solve the linear system. This intuition is substantiated by the following theorem.

Theorem 2: Let $C_A(n)$ be a class of $\mu(n) \times n$ matrices with $\sigma_1/\sigma_{\hat{r}(n)}$ polynomial in n, where μ and $\hat{r}$ are positive integer functions and $\hat{r}(n) \leq \min\{n, \mu(n)\}$. Let $C_b(n)$ be a class of $\mu(n) \times 1$ vectors, and let $q(n)$ be the smallest value such that

$$\sum_{i=\hat{r}(n)+1}^{\mu(n)} (u_i^T b)^2 = \sum_{i=\hat{r}(n)+1}^{\mu(n)} \beta_i^2 \leq (\varepsilon^2 - \varepsilon^{2q(n)})(\|b\|_2)^2 = (\varepsilon^2 - \varepsilon^{2q(n)}) \sum_{i=1}^{\mu(n)} \beta_i^2 \tag{17}$$

for all $A \in C_A(n)$ and all $b \in C_b(n)$. Then, if $A \in C_A(n)$, $b \in C_b(n)$, and $q(n)$ is at most polynomial in n, $a(A, b, \varepsilon) = O(\log n + \log\log(1/\varepsilon))$.

Proof: We divide the square of the relative error into two terms: one contributed by the projection of b on the "attuned" subspace of $R(A)$ and the other contributed by the projection of b (spanned by the first $\hat{r}(n)$ coordinates) on the "unattuned" subspace of $R(A)$ and $R(A)^C$ (collectively spanned by the remaining coordinates). From Equation (8),

$$e_l^2 \leq \frac{\sum_{i=1}^{\hat{r}(n)} (\omega_i^{(l)} \beta_i)^2 + \sum_{i=\hat{r}(n)+1}^{\mu(n)} (\omega_i^{(l)} \beta_i)^2}{\sum_{i=1}^{\mu(n)} \beta_i^2}.$$

If we define

$$\hat{e}_l = \left[\frac{\sum_{i=1}^{\hat{r}(n)} (\omega_i^{(l)} \beta_i)^2}{\sum_{i=1}^{\mu(n)} \beta_i^2} \right]^{1/2} \tag{18}$$

and recall from Equation (7) that $|\omega_i^{(l)}| \leq 1$, then we can write

$$e_l^2 \leq \hat{e}_l^2 + \frac{\sum_{i=\hat{r}(n)+1}^{\mu(n)} \beta_i^2}{\sum_{i=1}^{\mu(n)} \beta_i^2}.$$

Then, by Equation (17) of the hypothesis,

$$e_l^2 \le \hat{e}_l^2 + \varepsilon^2 - \varepsilon^{2q(n)} \tag{19}$$

so $e_l \le \varepsilon$ if $\hat{e}_l \le \varepsilon^{q(n)}$. This is analogous to the situation in Theorem 1, with e_l replaced by $\hat{e}_l$, $\kappa(A)$ replaced by $\sigma_1/\sigma_{\hat{r}(n)}$, and ε replaced by $\varepsilon^{q(n)}$. Consequently,

$$\begin{aligned} a(A, b, \varepsilon) &= O(\log n + \log(\sigma_1/\sigma_{\hat{r}(n)}) + \operatorname{loglog}(1/\varepsilon^{q(n)})) \\ &= O(\log n + \log(\sigma_1/\sigma_{\hat{r}(n)}) + \log(q(n)) + \operatorname{loglog}(1/\varepsilon)). \end{aligned}$$

By hypothesis, $\sigma_1/\sigma_{\hat{r}(n)}$ and $q(n)$ are at most polynomial in n. Therefore, $a(A, b, \varepsilon) = O(\log n + \operatorname{loglog}(1/\varepsilon))$. □

This theorem shows that Newton's method can accurately solve a much broader class of linear systems with $O(\log n)$ iterations than was previously realized; in particular, A does not need to be well-conditioned, and b does not even need to lie entirely in the "attuned" subspace of $R(A)$.

5. Analysis of the Effect of Roundoff Error

Until now, we have assumed that all computations were carried out with infinite precision. In this section, we will examine the effect of the roundoff errors induced by computing with finite precision. Henceforth, we assume that fixed-point numbers with g bits to the left of the radix point, h bits to the right of it, and one sign bit are used in all computations. Thus, we can represent numbers in the range $[-2^g+2^{-h}, 2^g-2^{-h}]$ with a precision of 2^{-h}. We also assume that the original entries of A are expressed with $\bar{g}$ bits to the left of the radix point, i.e., they are bounded in absolute value by $\gamma = 2^{\bar{g}}$.

As Söderström and Stewart observed [SS74], the stability of Newton's method for computing the pseudo-inverse of a matrix cannot be verified simply by asserting a self-correction property, as is often done in the analysis of the scalar Newton's method. Instead, we must undertake a thorough study of how the roundoff error accumulates. Our analysis differs from that of [SS74] in two aspects: first, we use a fixed-point model of computation, and second (and more important), we obtain precision requirements that depend on $a(A, b, \varepsilon)$ instead of $\kappa(A)$.

We will compute the same quantities that were computed in the preceding sections, accounting now for roundoff error, as indicated by tilde. The new equation for the Newton iteration on A is:

$$\tilde{Y}_{l+1} = 2\tilde{Y}_l - \tilde{Y}_l A \tilde{Y}_l + E_l, \tag{20}$$

where E_l is the error matrix, the generic entry e_{st} of which is bounded in absolute value by $\eta = 2^{-(h+1)}$. In the diagonal domain,

$$\tilde{\Delta}_{l+1} = 2\tilde{\Delta}_l - \tilde{\Delta}_l \Sigma \tilde{\Delta}_l + V^T E_l U. \tag{21}$$

For an individual entry $\tilde{\delta}_{ij}^{(l+1)}$ of the matrix $\tilde{\Delta}_{l+1}$, we have

$$\tilde{\delta}_{ij}^{(l+1)} = 2\tilde{\delta}_{ij}^{(l)} - \sum_{k=1}^{r} \tilde{\delta}_{ik}^{(l)} \sigma_k \tilde{\delta}_{kj}^{(l)} + \sum_{s=1}^{n}\sum_{t=1}^{m} v_{si} e_{st} u_{tj}. \tag{22}$$

Let us consider the last term of this equation. To obtain upper and lower bounds on $\tilde{\delta}_{ij}^{(l+1)}$, we must determine the maximum and minimum values for this summation, which we do in the following lemma.

Lemma 2: $-\eta\sqrt{mn} \le \sum_{s=1}^{n}\sum_{t=1}^{m} v_{si} e_{st} u_{tj} \le \eta\sqrt{mn}$.

Proof:

$$\sum_{s=1}^{n}\sum_{t=1}^{m} v_{si} e_{st} u_{tj} = \sum_{s=1}^{n} (v_{si} \sum_{t=1}^{m} e_{st} u_{tj})$$

Taking the absolute value:

$$\begin{aligned} |\sum_{s=1}^{n} (v_{si} \sum_{t=1}^{m} e_{st} u_{tj})| &\le \sum_{s=1}^{n} (|v_{si}| \sum_{t=1}^{m} e_{st} u_{tj}|) \\ &\le \sum_{s=1}^{n} (|v_{si}| \sum_{t=1}^{m} |e_{st}|\,|u_{tj}|) \\ &\le \sum_{s=1}^{n} (|v_{si}| \sum_{t=1}^{m} \eta\,|u_{tj}|) \\ &\le \eta(\sum_{s=1}^{n} |v_{si}|)(\sum_{t=1}^{m} |u_{tj}|) \end{aligned}$$

Since $\sum_{s=1}^{n} |v_{si}|^2 = 1$, $\sum_{s=1}^{n} |v_{si}| \le \sqrt{n}$, and, similarly, $\sum_{t=1}^{m} |u_{tj}| \le \sqrt{m}$. Therefore,

$$-\eta\sqrt{mn} \le \sum_{s=1}^{n}\sum_{t=1}^{m} v_{si} e_{st} u_{tj} \le \eta\sqrt{mn}. \ \square$$

It is also useful to establish the following lemma, the proof of which is omitted here for brevity, due to its similarity to that of Lemma 2.

Lemma 3: For $1 \le i \le p$, $\sigma_i \le \gamma\sqrt{mn}$.

Equation (22) and Lemma 2 yield the following recurrence relation:

$$2\tilde{\delta}_{ij}^{(l)} - \sum_{k=1}^{r} \tilde{\delta}_{ik}^{(l)} \sigma_k \tilde{\delta}_{kj}^{(l)} - \eta\sqrt{mn} \le \tilde{\delta}_{ij}^{(l+1)} \le 2\tilde{\delta}_{ij}^{(l)} - \sum_{k=1}^{r} \tilde{\delta}_{ik}^{(l)} \sigma_k \tilde{\delta}_{kj}^{(l)} + \eta\sqrt{mn}. \tag{23}$$

Based on this recurrence relation, the following lemma specifies a bound on the deviation of the lth finite-precision iterate from its exact value, i.e., its value if infinite precision were used.

Lemma 4: If $h \ge (\log 3)\cdot l + \overline{g} + \log r + \log m + \log n$, then

$$|\tilde{\delta}_{ij}^{(l)} - \delta_{ij}^{(l)}| \le \begin{cases} \sigma_i^{-1}(3^l - 1)\gamma\eta mn, & \text{if } \sigma_i > 1, \\ (3^l - 1)\eta\sqrt{mn}, & \text{if } \sigma_i \le 1. \end{cases}$$

Proof: (see Appendix A)

The following theorem, in turn, makes use of this result to obtain an upper bound $\tilde{e}_l$, the relative error after l finite-precision iterations.

Theorem 3: If $h \ge (\log 3)\cdot l + \overline{g} + \log r + \log m + \log n$, then $\tilde{e}_l \le e_l + (3^l - 1)\gamma\eta(mn)^{3/2}$.

Proof: If we define $\tilde{\Omega}_l = I - \Sigma\tilde{\Delta}_l$, then, analogous to Equation (8), we have

$$\tilde{e}_l = \frac{\| \tilde{\Omega}_l \beta \|_2}{\| \beta \|_2}. \tag{24}$$

By the properties of the Euclidean norm $\|\cdot\|_2$ and Equation (24),

$$\begin{aligned} \tilde{e}_l &= \frac{\| [\Omega_l + (\tilde{\Omega}_l - \Omega_l)]\beta \|_2}{\| \beta \|_2} \\ &\le \frac{\| \Omega_l \beta \|_2 + \| \Sigma(\Delta_l - \tilde{\Delta}_l)\beta \|_2}{\| \beta \|_2} \\ &\le e_l + \frac{\| \Sigma(\Delta_l - \tilde{\Delta}_l)\beta \|_2}{\| \beta \|_2} \\ &\le e_l + \| \Sigma(\Delta_l - \tilde{\Delta}_l) \|_2. \end{aligned}$$

Lemma 4 indicates that each entry in row i of $\Delta_l - \tilde{\Delta}_l$ has absolute value no greater than $\sigma_i^{-1}(3^l - 1)\gamma\eta mn$ for $\sigma_i > 1$ and no greater than $(3^l - 1)\eta\sqrt{mn}$ for $\sigma_i \le 1$, so each entry of $\Sigma(\Delta_l - \tilde{\Delta}_l)$ has absolute value no greater than

$(3^l-1)\gamma\eta mn$. Thus, $\|\Sigma(\Delta_l - \tilde{\Delta}_l)\|_\infty \le (3^l-1)\gamma\eta m^2 n$, and $\|\Sigma(\Delta_l - \tilde{\Delta}_l)\|_1 \le (3^l-1)\gamma\eta mn^2$,

so

$$\|\Sigma(\Delta_l - \tilde{\Delta}_l)\|_2 \le [\|\Sigma(\Delta_l - \tilde{\Delta}_l)\|_\infty \|\Sigma(\Delta_l - \tilde{\Delta}_l)\|_1]^{1/2} \le (3^l-1)\gamma\eta(mn)^{3/2}.$$

and the theorem follows. □

Theorem 3 shows that $\tilde{e}_l$ does not depart significantly from e_l if sufficient precision is used. In the following theorem, we show further that this additional error can be eliminated easily provided that b is "attuned" to A.

Theorem 4: If $A \in C_A(n)$ and $b \in C_b(n)$ (as defined in Theorem 2), then there exists an integer $h = O(\overline{g} + \log\mu(n) + \log n + q(n)\log(1/\varepsilon))$ such that $\tilde{a}(A, b, \varepsilon) = O(\log n + \log\log(1/\varepsilon))$.

Proof: From Theorem 3 (replacing m with $\mu(n)$) and Equation (19),

$$\begin{aligned}\tilde{e}_l &\le e_l + (3^l-1)\gamma\eta(\mu(n)n)^{3/2} \\ &\le [\hat{e}_l^2 + \varepsilon^2 - \varepsilon^{2q(n)}]^{1/2} + 3^l\gamma\eta(\mu(n)n)^{3/2}.\end{aligned}$$

Therefore, $\tilde{e}_l \le \varepsilon$ if

$$\begin{aligned}\hat{e}_l^2 + \varepsilon^2 - \varepsilon^{2q(n)} &\le [\varepsilon - 3^l\gamma\eta(\mu(n)n)^{3/2}]^2 \\ \hat{e}_l^2 &\le \varepsilon^{2q(n)} - [2\varepsilon - 3^l\gamma\eta(\mu(n)n)^{3/2}][3^l\gamma\eta(\mu(n)n)^{3/2}] \\ &\le \varepsilon^{2q(n)} - 2\varepsilon[3^l\gamma\eta(\mu(n)n)^{3/2}].\end{aligned}$$

Let us assume that $l = O(\log n + \log\log(1/\varepsilon))$; then, recalling that $\eta = 2^{-(h+1)}$, some integer $h = O(\overline{g} + \log\mu(n) + \log n + q(n)\log(1/\varepsilon))$ can be chosen so that $2\varepsilon[3^l\gamma\eta(\mu(n)n)^{3/2}] \le \varepsilon^{4q(n)}$. Therefore, we only require that

$$\begin{aligned}\hat{e}_l^2 &\le \varepsilon^{2q(n)} - \varepsilon^{4q(n)} \\ \hat{e}_l^2 &\le \varepsilon^{4q(n)} \\ \hat{e}_l &\le \varepsilon^{2q(n)}.\end{aligned}$$

As in Theorem 2, $l = O(\log n + \log\log(1/\varepsilon))$ is sufficient to ensure that $\hat{e}_l \le \varepsilon^{2q(n)}$. This justifies our assumption and verifies the theorem. □

The theorem above shows that the notion of attunement number remains valid in the presence of roundoff errors regardless of the condition number of the matrix, provided there are sufficiently many bits to the right of the radix point. We must also determine the number, g, of bits needed to the left of the radix point, an equally important, but somewhat simpler question.

Theorem 5: If $h \geq (\log 3)\cdot l + \bar{g} + 3/2\log m + 3/2\log n + 1$, then $g = l$ is sufficient to represent $\tilde{Y}_l$.

Proof: For an individual element $\tilde{y}_{ij}^{(l)}$ of the matrix $\tilde{Y}_l = V\tilde{\Delta}_l U^T$, we have

$$|\tilde{y}_{ij}^{(l)}| = |\sum_{s=1}^{n}\sum_{t=1}^{m} v_{is}\tilde{\delta}_{st}^{(l)} u_{jt}|.$$

Using an approach similar to that used in Lemma 2, along with the bounds obtained in Lemma 4 (which can be applied since $r \leq m$ and $r \leq n$), the Cauchy-Schwarz inequality, and the fact that U and V are unitary, we obtain:

$$\begin{aligned}
|\tilde{y}_{ij}^{(l)}| &= |\sum_{s=1}^{n}\sum_{t=1,t\neq s}^{m} v_{is}\tilde{\delta}_{st}^{(l)} u_{jt} + \sum_{s=1}^{\min(m,n)} v_{is}\tilde{\delta}_{ss}^{(l)} u_{js}| \\
&\leq (3^l-1)\gamma\eta(mn)^{3/2} + \{\max_{k\leq l,\ 0<\sigma\leq\alpha^{1/2}} [\sigma^{-1}[1-(1-\alpha\sigma^2)^{2^k}] + (3^l-1)\gamma\eta mn\} \sum_{s=1}^{\min(m,n)} |v_{is}||v_{js}| \\
&\leq (3^l-1)\gamma\eta(mn)^{3/2} + [\alpha^{1/2}2^l + (3^l-1)\gamma\eta mn]\left[\sum_{s=1}^{n}|v_{is}|^2 \sum_{s=1}^{m}|u_{js}|^2\right]^{1/2} \\
&\leq \alpha^{1/2}2^l + 2(3^l-1)\gamma\eta(mn)^{3/2}
\end{aligned}$$

By our choice of h, η is small enough that taking the integer part of this expression leaves only $\left\lfloor |\tilde{y}_{ij}^{(l)}| \right\rfloor \leq \alpha^{1/2}2^l$. We can assume without loss of generality that $A \neq 0$ and that A is scaled so that $\alpha \leq 1$. Therefore, $g = l$ bits suffice to represent $\tilde{y}_{ij}^{(l)}$. □

Acknowledgements

The authors gratefully acknowledge the helpful suggestions and insights of Ahmed Sameh and Gianfranco Bilardi.

Appendix A

Proof of Lemma 4 (by induction on l): First, we decompose the lemma into ten different inequalities, each of which will be proven separately:

$$\text{For } \sigma_i > 1\text{: } \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^l} - (3^l-1)\gamma\eta mn] \le \tilde{\delta}_{ii}^{(l)} \le \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^l} + (3^l-1)\gamma\eta mn]$$

$$\text{For } 0 < \sigma_i \le 1\text{: } \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^l}] - (3^l-1)\eta\sqrt{mn} \le \tilde{\delta}_{ii}^{(l)} \le \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^l}] + (3^l-1)\eta\sqrt{mn}$$

$$\text{For } \sigma_i = 0\text{: } -(3^l-1)\eta\sqrt{mn} \le \tilde{\delta}_{ii}^{(l)} \le (3^l-1)\eta\sqrt{mn}$$

$$\text{For } \sigma_i > 1 \text{ and } i \ne j\text{: } -\sigma_i^{-1}(3^l-1)\gamma\eta mn \le \tilde{\delta}_{ij}^{(l)} \le \sigma_i^{-1}(3^l-1)\gamma\eta mn$$

$$\text{For } \sigma_i \le 1 \text{ and } i \ne j\text{: } -(3^l-1)\eta\sqrt{mn} \le \tilde{\delta}_{ij}^{(l)} \le (3^l-1)\eta\sqrt{mn}$$

As a basis, we note that the initial condition $\tilde{\Delta}_0 = \Delta_0 = \alpha\Sigma^T$ satisfies these inequalities. Inductively, we now assume that the theorem holds for $\delta_{ij}^{(l)}$. We will show that it also must hold for $\delta_{ij}^{(l+1)}$.

To prove the upper bound for $\tilde{\delta}_{ii}^{(l+1)}$, $\sigma_i > 1$:

$$\tilde{\delta}_{ii}^{(l+1)} \le \tilde{\delta}_{ii}^{(l)}(2 - \tilde{\delta}_{ii}^{(l)}\sigma_i) - \sum_{k\ne i}\tilde{\delta}_{ik}^{(l)}\sigma_k\tilde{\delta}_{ki}^{(l)} + \eta\sqrt{mn};$$

by the inductive hypothesis and since $\gamma\sqrt{mn}$:

$$\tilde{\delta}_{ii}^{(l+1)} \le \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^l} + (3^l-1)\gamma\eta mn]\{2 - [1-(1-\alpha\sigma_i^2)^{2^l} + (3^l-1)\gamma\eta mn]\} + \sum_{k\ne i,\ k\le r}\sigma_i^{-1}[(3^l-1)\gamma\eta mn]^2 + \eta\sqrt{mn}$$

$$= \sigma_i^{-1}\{1-(1-\alpha\sigma_i^2)^{2^{l+1}} + 2(1-\alpha\sigma_i^2)^{2^l}(3^l-1)\gamma\eta mn - [(3^l-1)\gamma\eta mn]^2 + (r-1)[(3^l-1)\gamma\eta mn]^2\} + \eta\sqrt{mn};$$

since $\alpha\sigma_i^2 \ge 0$:

$$\tilde{\delta}_{ii}^{(l+1)} \le \sigma_i^{-1}\{1-(1-\alpha\sigma_i^2)^{2^{l+1}} + 2(3^l-1)\gamma\eta mn + r[(3^l-1)\gamma\eta mn]^2 + \sigma_i\eta\sqrt{mn}\};$$

by Lemma 3 and our hypothesis on h:

$$\tilde{\delta}_{ii}^{(l+1)} \le \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}} + 3(3^l-1)\gamma\eta mn + \gamma\eta mn] < \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}} + (3^{l+1}-1)\gamma\eta mn].$$

To prove the lower bound for $\tilde{\delta}_{ii}^{(l+1)}$, $\sigma_i > 1$:

$$\tilde{\delta}_{ii}^{(l+1)} \ge \tilde{\delta}_{ii}^{(l)}(2 - \tilde{\delta}_{ii}^{(l)}\sigma_i) - \sum_{k\ne i}\tilde{\delta}_{ik}^{(l)}\sigma_k\tilde{\delta}_{ki}^{(l)} - \eta\sqrt{mn};$$

by the inductive hypothesis and since $\gamma\sqrt{mn} \ge 1$:

$$\tilde{\delta}_{ii}^{(l+1)} \ge \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^l} - (3^l-1)\gamma\eta mn]\{2 - [1-(1-\alpha\sigma_i^2)^{2^l} - (3^l-1)\gamma\eta mn]\} - \sum_{k\ne i,\ k\le r}\sigma_i^{-1}[(3^l-1)\gamma\eta mn]^2 - \eta\sqrt{mn}$$

$$= \sigma_i^{-1}\{1-(1-\alpha\sigma_i^2)^{2^{l+1}} - 2(1-\alpha\sigma_i^2)^{2^l}(3^l-1)\gamma\eta mn - [(3^l-1)\gamma\eta mn]^2 - (r-1)[(3^l-1)\gamma\eta mn]^2\} - \eta\sqrt{mn};$$

since $\alpha\sigma_i^2 \geq 0$:

$$\tilde{\delta}_{ii}^{(l+1)} \geq \sigma_i^{-1}\{1-(1-\alpha\sigma_i^2)^{2^{l+1}} - 2(3^l-1)\gamma\eta mn - r[(3^l-1)\gamma\eta mn]^2 - \sigma_i\eta\sqrt{mn}\};$$

by Lemma 3 and our hypothesis on h:

$$\tilde{\delta}_{ii}^{(l+1)} \geq \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}} - 3(3^l-1)\gamma\eta mn - \gamma\eta mn] > \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}} - (3^{l+1}-1)\gamma\eta mn].$$

To prove the upper bound for $\tilde{\delta}_{ii}^{(l+1)}$, $0 < \sigma_i \leq 1$:

$$\tilde{\delta}_{ii}^{(l+1)} \leq \tilde{\delta}_{ii}^{(l)}(2 - \tilde{\delta}_{ii}^{(l)}\sigma_i) - \sum_{k\neq i}\tilde{\delta}_{ik}^{(l)}\sigma_k\tilde{\delta}_{ki}^{(l)} + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \leq \{\sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^l}] + (3^l-1)\eta\sqrt{mn}\}\{2 - [1-(1-\alpha\sigma_i^2)^{2^l}] - \sigma_i(3^l-1)\eta\sqrt{mn}\}$$
$$+ \sum_{k\neq i,\ k\leq r}[(3^l-1)\eta\sqrt{mn}][(3^l-1)\gamma\eta mn] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \leq \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}}] + 2(1-\alpha\sigma_i^2)^{2^l}(3^l-1)\eta\sqrt{mn} - \sigma_i[(3^l-1)\eta\sqrt{mn}]^2 + [(3^l-1)\eta\sqrt{mn}][(r-1)(3^l-1)\gamma\eta mn] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \leq \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}}] + 2(3^l-1)\eta\sqrt{mn} + r[(3^l-1)\gamma\eta mn][(3^l-1)\eta\sqrt{mn}] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \leq \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}}] + 3(3^l-1)\eta\sqrt{mn} + \eta\sqrt{mn} < \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}}] + (3^{l+1}-1)\eta\sqrt{mn}$$

To prove the lower bound for $\tilde{\delta}_{ii}^{(l+1)}$, $0 < \sigma_i \leq 1$:

$$\tilde{\delta}_{ii}^{(l+1)} \geq \tilde{\delta}_{ii}^{(l)}(2 - \tilde{\delta}_{ii}^{(l)}\sigma_i) - \sum_{k\neq i}\tilde{\delta}_{ik}^{(l)}\sigma_k\tilde{\delta}_{ki}^{(l)} - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \geq \{\sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^l}] - (3^l-1)\eta\sqrt{mn}\}\{2 - [1-(1-\alpha\sigma_i^2)^{2^l}] + \sigma_i(3^l-1)\eta\sqrt{mn}\}$$
$$- \sum_{k\neq i,\ k\leq r}[(3^l-1)\eta\sqrt{mn}][(3^l-1)\gamma\eta mn] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \geq \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}}] - 2(1-\alpha\sigma_i^2)^{2^l}(3^l-1)\eta\sqrt{mn} - \sigma_i[(3^l-1)\eta\sqrt{mn}]^2 - [(3^l-1)\eta\sqrt{mn}][(r-1)(3^l-1)\gamma\eta mn] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \geq \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}}] - 2(3^l-1)\eta\sqrt{mn} - r[(3^l-1)\gamma\eta mn][(3^l-1)\eta\sqrt{mn}] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \geq \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}}] - 3(3^l-1)\eta\sqrt{mn} - \eta\sqrt{mn} > \sigma_i^{-1}[1-(1-\alpha\sigma_i^2)^{2^{l+1}}] - (3^{l+1}-1)\eta\sqrt{mn}$$

To prove the upper bound for $\tilde{\delta}_{ii}^{(l+1)}$, $\sigma_i = 0$:

$$\tilde{\delta}_{ii}^{(l+1)} \leq 2\tilde{\delta}_{ii}^{(l)} - \sum_{k\neq i}\tilde{\delta}_{ik}^{(l)}\sigma_k\tilde{\delta}_{ki}^{(l)} + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \leq 2[(3^l-1)\eta\sqrt{mn}] + \sum_{k\leq r}[(3^l-1)\eta\sqrt{mn}][(3^l-1)\gamma\eta mn] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \leq 2[(3^l-1)\eta\sqrt{mn}] + [(3^l-1)\eta\sqrt{mn}][r(3^l-1)\gamma\eta mn] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \leq 3[(3^l-1)\eta\sqrt{mn}] + \eta\sqrt{mn} < (3^{l+1}-1)\eta\sqrt{mn}$$

To prove the lower bound for $\tilde{\delta}_{ii}^{(l+1)}$, $\sigma_i = 0$:

$$\tilde{\delta}_{ii}^{(l+1)} \geq 2\tilde{\delta}_{ii}^{(l)} - \sum_{k \neq i} \tilde{\delta}_{ik}^{(l)} \sigma_k \tilde{\delta}_{ki}^{(l)} - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \geq 2[-(3^l-1)\eta\sqrt{mn}] - \sum_{k \leq r} [(3^l-1)\eta\sqrt{mn}][(3^l-1)\gamma\eta mn] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \geq 2[-(3^l-1)\eta\sqrt{mn}] - [(3^l-1)\eta\sqrt{mn}][r(3^l-1)\gamma\eta mn] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ii}^{(l+1)} \geq 3[-(3^l-1)\eta\sqrt{mn}] - \eta\sqrt{mn} > -(3^{l+1}-1)\eta\sqrt{mn}$$

To prove the upper bound for $\tilde{\delta}_{ij}^{(l+1)}$, $\sigma_i > 1$ and $i \neq j$:

$$\tilde{\delta}_{ij}^{(l+1)} \leq \tilde{\delta}_{ij}^{(l)}(2 - \sigma_i\tilde{\delta}_{ii}^{(l)} - \sigma_j\tilde{\delta}_{jj}^{(l)}) - \sum_{k \neq i,j} \tilde{\delta}_{ik}^{(l)} \sigma_k \tilde{\delta}_{kj}^{(l)} + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \leq \sigma_i^{-1}(3^l-1)\gamma\eta mn\{2 - [1-(1-\alpha\sigma_i^2)^{2^l} - (3^l-1)\gamma\eta mn] - [1-(1-\alpha\sigma_j^2)^{2^l} - (3^l-1)\gamma\eta mn]\}$$
$$+ \sum_{k \neq i,j,\ k \leq r} \sigma_i^{-1}[(3^l-1)\gamma\eta mn]^2 + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \leq \sigma_i^{-1}(3^l-1)\gamma\eta mn[(1-\alpha\sigma_i^2)^{2^l} + (3^l-1)\gamma\eta mn + (1-\alpha\sigma_j^2)^{2^l} + (3^l-1)\gamma\eta mn + (r-2)(3^l-1)\gamma\eta mn] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \leq \sigma_i^{-1}(3^l-1)\gamma\eta mn[2 + r(3^l-1)\gamma\eta mn] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \leq 3[\sigma_i^{-1}(3^l-1)\gamma\eta mn] + \eta\sqrt{mn} < \sigma_i^{-1}(3^{l+1}-1)\gamma\eta mn$$

To prove the lower bound for $\tilde{\delta}_{ij}^{(l+1)}$, $\sigma_i > 1$ and $i \neq j$:

$$\tilde{\delta}_{ij}^{(l+1)} \geq \tilde{\delta}_{ij}^{(l)}(2 - \sigma_i\tilde{\delta}_{ii}^{(l)} - \sigma_j\tilde{\delta}_{jj}^{(l)}) - \sum_{k \neq i,j} \tilde{\delta}_{ik}^{(l)} \sigma_k \tilde{\delta}_{kj}^{(l)} - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \geq -\sigma_i^{-1}(3^l-1)\gamma\eta mn\{2 - [1-(1-\alpha\sigma_i^2)^{2^l} - (3^l-1)\gamma\eta mn] - [1-(1-\alpha\sigma_j^2)^{2^l} - (3^l-1)\gamma\eta mn]\}$$
$$- \sum_{k \neq i,j,\ k \leq r} \sigma_i^{-1}[(3^l-1)\gamma\eta mn]^2 - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \geq -\sigma_i^{-1}(3^l-1)\gamma\eta mn[(1-\alpha\sigma_i^2)^{2^l} + (3^l-1)\gamma\eta mn + (1-\alpha\sigma_j^2)^{2^l} + (3^l-1)\gamma\eta mn + (r-2)(3^l-1)\gamma\eta mn] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \geq -\sigma_i^{-1}(3^l-1)\gamma\eta mn[2 + r(3^l-1)\gamma\eta mn] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \geq 3[-\sigma_i^{-1}(3^l-1)\gamma\eta mn] - \eta\sqrt{mn} > -\sigma_i^{-1}(3^{l+1}-1)\gamma\eta mn$$

To prove the upper bound for $\tilde{\delta}_{ij}^{(l+1)}$, $\sigma_i \leq 1$ and $i \neq j$:

$$\tilde{\delta}_{ij}^{(l+1)} \leq \tilde{\delta}_{ij}^{(l)}(2 - \sigma_i\tilde{\delta}_{ii}^{(l)} - \sigma_j\tilde{\delta}_{jj}^{(l)}) - \sum_{k \neq i,j} \tilde{\delta}_{ik}^{(l)} \sigma_k \tilde{\delta}_{kj}^{(l)} + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \leq (3^l-1)\eta\sqrt{mn}\{2 - [1-(1-\alpha\sigma_i^2)^{2^l} - \sigma_i(3^l-1)\eta\sqrt{mn}] - [1-(1-\alpha\sigma_j^2)^{2^l} - (3^l-1)\gamma\eta mn]\}$$
$$+ \sum_{k \neq i,j,\ k \leq r} [(3^l-1)\eta\sqrt{mn}][(3^l-1)\gamma\eta mn] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \leq (3^l-1)\eta\sqrt{mn}[(1-\alpha\sigma_i^2)^{2^l} + (3^l-1)\eta\sqrt{mn} + (1-\alpha\sigma_j^2)^{2^l} + (3^l-1)\gamma\eta mn + (r-2)(3^l-1)\gamma\eta mn] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \le (3^l-1)\eta\sqrt{mn}\,[2 + r\,(3^l-1)\gamma\eta mn\,] + \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \le 3[(3^l-1)\eta\sqrt{mn}\,] + \eta\sqrt{mn} < (3^{l+1}-1)\eta\sqrt{mn}$$

To prove the lower bound for $\tilde{\delta}_{ij}^{(l+1)}$, $\sigma_i \le 1$ and $i \ne j$:

$$\tilde{\delta}_{ij}^{(l+1)} \ge \tilde{\delta}_{ij}^{(l)}(2 - \sigma_i\tilde{\delta}_{ii}^{(l)} - \sigma_j\tilde{\delta}_{jj}^{(l)}) - \sum_{k\ne i,j} \tilde{\delta}_{ik}^{(l)}\sigma_k\tilde{\delta}_{kj}^{(l)} - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \ge -(3^l-1)\eta\sqrt{mn}\,\{2 - [1-(1-\alpha\sigma_i^2)^{2^l} - \sigma_i(3^l-1)\eta\sqrt{mn}\,] - [1-(1-\alpha\sigma_j^2)^{2^l} - (3^l-1)\gamma\eta mn\,]\}$$

$$+ \sum_{k\ne i,j,\ k\le r} [(3^l-1)\eta\sqrt{mn}\,][(3^l-1)\gamma\eta mn\,] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \ge -(3^l-1)\eta\sqrt{mn}\,[(1-\alpha\sigma_i^2)^{2^l} + (3^l-1)\eta\sqrt{mn} + (1-\alpha\sigma_j^2)^{2^l} + (3^l-1)\gamma\eta mn + (r-2)(3^l-1)\gamma\eta mn\,] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \ge -(3^l-1)\eta\sqrt{mn}\,[2 + r\,(3^l-1)\gamma\eta mn\,] - \eta\sqrt{mn}$$

$$\tilde{\delta}_{ij}^{(l+1)} \ge 3[-(3^l-1)\eta\sqrt{mn}\,] - \eta\sqrt{mn} > -(3^{l+1}-1)\eta\sqrt{mn}$$

This concludes the inductive step and verifies the lemma. □

REFERENCES

[BC66] A. Ben-Israel and D. Cohen, "On Iterative Computation of Generalized Inverses and Associated Projections," *SIAM Journal on Numerical Analysis*, vol. 3, no. 3, Jun. 1966, pp. 410-419.

[B84] A. Bojańczyk, "Complexity of Solving Linear Systems in Different Models of Computation," *SIAM Journal on Numerical Analysis*, vol. 21, no. 3, Jun. 1984, pp. 591-603.

[C76] L. Csanky, "Fast Parallel Matrix Inversion Algorithms," *SIAM Journal on Computing*, vol. 5, no. 4, Apr. 1976, pp. 618-623.

[FM67] G. Forsythe and C. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, 1967.

[GV83] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1983.

[IK66] E. Isaacson and H. Keller, *Analysis of Numerical Methods*, John Wiley and Sons, 1966.

[PR85] V. Pan and J. Reif, "Efficient Parallel Solution of Linear Systems," *Proceedings of the 17th Annual ACM Symposium on the Theory of Computing*, Providence, RI, May 1985, pp. 143-152.

[PS78] F. Preparata and D. Sarawate, "An Improved Parallel Processor Bound in Fast Matrix Inversion," *Information Processing Letters*, vol. 7, no. 3, Apr. 1978, pp. 148-150.

[Sc33] G. Schulz, "Iterative Berechnung der reziproken Matrix," *Zeitschrift fur angewandte Math. und Mech.*, vol. 13, no. 1, Feb. 1933, pp. 57-59.

[SS74] T. Söderström and G. Stewart, "On the Numerical Properties of an Integrative Method for Computing the Moore-Penrose Generalized Inverse," *SIAM Journal on Numerical Analysis*, vol. 11, no. 1, Mar. 1974, pp. 61-74.

[St73] G. Stewart, *Introduction to Matrix Computations*, Academic Press, 1973.

[W65] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, 1965.