The Stability of Inversion Formulas for Toeplitz Matrices

Martin H. Gutknecht*
Interdisciplinary Project Center for Supercomputing ETH Zurich
ETH-Zentrum
CH-8092 Zurich, Switzerland

and

Marlis Hochbruck†
Mathematisches Institut
Universität Tübingen
Auf der Morgenstelle 10
D-72076 Tübingen, Germany

Dedicated to M. Fiedler and V. Pták.

Submitted by Angelika Bunse-Gerstner

ABSTRACT

It is shown that under suitable assumptions the well-known formulas for the inverse of Toeplitz matrices that are due to Gohberg and Semencul and Heinig are weakly stable, i.e., they are numerically forward stable if the matrices that are by assumption nonsingular are actually well conditioned. The same is true for another, less-known pair of inversion formulas that only involve the left biorthogonal Szegő polynomials.

1. INTRODUCTION

The asymptotically fastest methods for solving Toeplitz systems, which have been proposed by Musicus [23], de Hoog [9], and Ammar and Gragg [3, 2, 4] for Hermitian systems, by Bitmead and Anderson [6] and Morf [22] for...
strongly regular systems, and by Gutknecht [13] and Gutknecht and Hochbruck [15, 14] for general non-Hermitian Toeplitz systems, normally make use at the end of inversion formulas for Toeplitz matrices. Two well-known such inversion formulas are due to Gohberg and Semencul [11]; further similar formulas were found by Heinig [16], and Heinig and Rost [17], who outlined a general technique that is in particular applicable to other matrices with small displacement rank, but also to Hankel matrices. See also Friedlander, Morf, Kailath, and Ljung [10] and, ten years later, Heinig and Rost [18, 19] and Kailath and Chun [20] for similar generalizations. This is just a very limited selection of the work on inversion formulas for Toeplitz and related matrices. In particular, there are also a number of papers on generalizations to block matrices, which are not addressed here.

In an equivalent recursive form, the first Gohberg-Semencul formula was already given by Trench [26], and it has been observed by Kailath et al. [21] and others that this formula is just a matrix reformulation of an analogue of the Christoffel-Darboux relation, which for Hermitian Toeplitz matrices dates back to Szegö [25], and for non-Hermitian Toeplitz matrices to Baxter [5].

It is a widespread belief that such inversion formulas, and in particular the Gohberg-Semencul formula, are numerically unstable. For example, this belief has been nourished by a remark of Bunch [7], who pointed out correctly that for well-conditioned Toeplitz matrices, the Gohberg-Semencul formula may be unstable: the roundoff errors occurring during the evaluation of the formula may become arbitrarily large; in fact, the formula can even break down due to division by zero. However, it is well known that the formula is only applicable to a nonsingular Toeplitz matrix whose last leading principal submatrix is also nonsingular. Hence, numerical stability can only be expected if both the full matrix and this particular submatrix are well conditioned. Under that assumption there is also a similar formula for the inverse of this submatrix. As a pair these inversion formulas can be considered as weakly stable in the sense that the effects of roundoff errors occurring during the evaluation remain bounded as long as the condition and the order of the Toeplitz matrix and its last leading principal submatrix are bounded. Another pair of inversion formulas presented here has a similar property. Like the other examples we discuss, these two turned out to be special cases of a general formula due to Heinig and Rost [19]. In contrast to the pair of Gohberg-Semencul formulas and this other pair, Heinig's inversion formula, which comes in two equivalent versions and was independently also found by Russakowski [24], only requires one matrix to be well conditioned. For this reason, Heinig's formula is the most powerful, and it is easy to derive the other ones from it.

Heinig's formula is the appropriate one for the sawtooth algorithms proposed in [13], while the Gohberg-Semencul pair fits the lookahead
Levinson and Schur algorithms of [15], and the other pair is adjusted to the row algorithms of [13].

Given is a real or complex nonsingular \( n \times n \) Toeplitz matrix

\[
T := T_n := \begin{bmatrix}
\mu_0 & \mu_{-1} & \cdots & \mu_{-n+1} \\
\mu_1 & \mu_0 & \cdots & \mu_{-n+2} \\
\vdots & \vdots & \ddots & \vdots \\
\mu_{n-1} & \mu_{n-2} & \cdots & \mu_0
\end{bmatrix}.
\]

We consider also \( T_{n-1} \), its leading principal submatrix of order \( n - 1 \).

Let \( e_k \) be the \( k \)th unit vector in \( \mathbb{C}^n \), and define the right-hand sides

\[
f := -\begin{bmatrix}
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_n
\end{bmatrix}^T, \quad g := -\begin{bmatrix}
\mu_{-n} \\
\mu_{-n+1} \\
\vdots \\
\mu_{-1}
\end{bmatrix}^T.
\]

Denote by \( u, v, x, \) and \( y \) the solutions of the four \( n \times n \) Toeplitz systems

\[
Tu = f, \quad (2.1)
\]
\[
Tv = g, \quad (2.2)
\]
\[
Tx = e_1, \quad (2.3)
\]
\[
Ty = e_n. \quad (2.4)
\]

Note that \( f \) and \( g \) contain the coefficients \( \mu_n \) and \( \mu_{-n} \), respectively, which do not occur in \( T \). Hence, these two coefficients can be considered as free parameters if the task is just to invert \( T \). In addition to (2.1)-(2.4) we consider the following two systems of order \( n - 1 \):

\[
T_{n-1} \bar{u} = \bar{f}, \quad \bar{f} := -\begin{bmatrix}
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_{n-1}
\end{bmatrix}^T, \quad (2.5)
\]
\[
T_{n-1} \bar{v} = \bar{g}, \quad \bar{g} := -\begin{bmatrix}
\mu_{-n+1} \\
\mu_{-n+2} \\
\vdots \\
\mu_{-1}
\end{bmatrix}^T. \quad (2.6)
\]

(The bar does not denote complex conjugation.) The systems (2.1)-(2.2) and (2.5)-(2.6) are often called Yule-Walker equations.
The coefficients of the solution vectors are chosen as follows:

\[ u = \begin{bmatrix} \psi_1 & \psi_2 & \cdots & \psi_n \end{bmatrix}^T, \]
\[ v = \begin{bmatrix} \phi_0 & \phi_1 & \cdots & \phi_{n-1} \end{bmatrix}^T, \]
\[ x = \begin{bmatrix} \xi_0 & \xi_1 & \cdots & \xi_{n-1} \end{bmatrix}^T, \]
\[ y = \begin{bmatrix} \eta_0 & \eta_1 & \cdots & \eta_{n-1} \end{bmatrix}^T, \]
\[ \bar{u} = \begin{bmatrix} \bar{\psi}_1 & \bar{\psi}_2 & \cdots & \bar{\psi}_{n-1} \end{bmatrix}^T, \]
\[ \bar{v} = \begin{bmatrix} \bar{\phi}_0 & \bar{\phi}_1 & \cdots & \bar{\phi}_{n-2} \end{bmatrix}^T. \]

Note the index shift in \( u \) and \( \bar{u} \). We set additionally

\[ \psi_0 = 1, \quad \phi_n = 1, \quad \bar{\psi}_0 = 1, \quad \bar{\phi}_{n-1} = 1. \quad (2.7) \]

There is a well-known, easily verified relationship between (2.5) and (2.4), and between (2.6) and (2.3):

**Lemma 2.1.**

(i) If \( T_{n-1} \) is nonsingular and the solution \( \bar{u} \) of (2.5) (extended by (2.7)) satisfies

\[ \sigma^- := \begin{bmatrix} \mu_0 & \cdots & \mu_{-n+1} \end{bmatrix} \begin{bmatrix} 1 \\ \bar{u} \end{bmatrix} = \sum_{k=0}^{n-1} \mu_{-k} \bar{\psi}_k \neq 0, \]

then \( T \) is also nonsingular, and

\[ x = \frac{1}{\sigma^-} \begin{bmatrix} 1 \\ \bar{u} \end{bmatrix}, \quad \text{i.e.,} \quad \xi_k = \bar{\psi}_k / \sigma^-, \quad k = 0, \ldots, n-1. \]

(ii) If \( T_{n-1} \) is nonsingular and the solution \( \bar{v} \) of (2.6) (extended by (2.7)) satisfies

\[ \sigma^+ := \begin{bmatrix} \mu_{-n-1} & \cdots & \mu_0 \end{bmatrix} \begin{bmatrix} \bar{v} \\ 1 \end{bmatrix} = \sum_{k=0}^{n-1} \mu_k \bar{\phi}_{n-1-k} \neq 0, \]
then $T$ is also nonsingular, and

$$ y = \frac{1}{\sigma^+} \begin{bmatrix} \bar{v} \\ 1 \end{bmatrix}, \text{ i.e., } \eta_k = \bar{\phi}_k / \sigma^+, \ k = 0, \ldots, n - 1. $$

(iii) If $T$ is nonsingular and the solution $x$ of (2.3) satisfies $\xi_0 \neq 0$, then $T_{n-1}$ is also nonsingular, and

$$ \begin{bmatrix} 1 \\ u \end{bmatrix} = \frac{1}{\xi_0} x, \text{ i.e., } \bar{\psi}_k = \xi_k / \xi_0, \ k = 0, \ldots, n - 1. \quad (2.8) $$

(iv) If $T$ is nonsingular and the solution $y$ of (2.4) satisfies $\eta_{n-1} \neq 0$, then $T_{n-1}$ is also nonsingular, and

$$ \begin{bmatrix} \bar{v} \\ 1 \end{bmatrix} = \frac{1}{\eta_{n-1}} y, \text{ i.e., } \bar{\phi}_k = \eta_k / \eta_{n-1}, \ k = 0, \ldots, n - 1. $$

Here are some well-known inversion formulas, proved, e.g., in [17].

**Theorem 2.1** (Heinig [16; 17, Theorem 1.1]). Assume (2.2) and (2.3) have solutions. Then $T$ is nonsingular and

$$ T^{-1} = \begin{bmatrix} \xi_0 & & & 0 \\ \bar{\xi}_1 & \xi_0 & & \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{n-1} & \ldots & \bar{\xi}_1 & \xi_0 \end{bmatrix} \begin{bmatrix} 1 & \phi_{n-1} & \ldots & \phi_1 \\ 0 & 1 & \ddots & \ddots \\ \vdots & \ddots & \ddots & \phi_{n-1} \\ 0 & \ddots & \ddots & 1 \end{bmatrix} $$

$$ - \begin{bmatrix} \phi_0 & & & 0 \\ \phi_1 & \phi_0 & & \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n-1} & \ldots & \phi_1 & \phi_0 \end{bmatrix} \begin{bmatrix} 0 & \xi_{n-1} & \ldots & \xi_1 \\ \xi_{n-1} & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \ddots & \ddots & \xi_{n-1} \end{bmatrix}. \quad (2.9) $$

**Theorem 2.2** [17, Remark 1.1]. Assume (2.1) and (2.4) have solutions.
Then $T$ is nonsingular and

$$
T^{-1} = \begin{bmatrix}
\eta_{n-1} & \eta_{n-2} & \ldots & \eta_0 \\
& \eta_{n-1} & \ldots & \\
& & \eta_{n-2} & \\
& & & \eta_{n-1} \\
0 & & & \\
\end{bmatrix}
\begin{bmatrix}
1 & \eta_0 \\
\psi_1 & 1 \\
& \ldots & \ldots \\
\psi_{n-1} & \psi_1 & 1 \\
\end{bmatrix}
- \begin{bmatrix}
\psi_n & \psi_{n-1} & \ldots & \psi_1 \\
& \psi_n & \ldots & \\
& & \psi_{n-1} & \\
& & & \psi_n \\
0 & & & \\
\end{bmatrix}
\begin{bmatrix}
0 & \eta_0 \\
& \ldots & \ldots \\
& & \eta_{n-2} & \\
& & & \eta_0 \\
\end{bmatrix}.
(2.10)
$$

**Theorem 2.3** (Gohberg and Semencul [11; 17, Theorems 1.2, 1.3]). Assume (2.3) and (2.4) have solutions, and $\xi_0 \neq 0$ or $\eta_{n-1} \neq 0$. Then $\xi_0 = \eta_{n-1}$, both $T$ and $T_{n-1}$ are nonsingular, and they can be represented as

$$
T_1 = \frac{1}{\xi_0}
\begin{bmatrix}
\xi_0 & \eta_0 & \ldots & \eta_{n-1} \\
\xi_1 & \xi_0 & \ldots & \\
& \ldots & \ldots & \\
\xi_{n-1} & \ldots & \xi_1 & \xi_0 \\
\end{bmatrix}
\begin{bmatrix}
\eta_{n-1} & \eta_{n-2} & \ldots & \eta_0 \\
& & \ldots & \\
& & & \eta_{n-1} \\
\end{bmatrix}
- \begin{bmatrix}
0 & \eta_0 \\
& \ldots & \ldots \\
& & \eta_{n-2} & \\
& & & \eta_0 \\
\end{bmatrix}
\begin{bmatrix}
0 & \xi_{n-1} & \ldots & \xi_1 \\
& & \ldots & \\
& & & 0 \\
\end{bmatrix}.
(2.11)
$$

$$
T_n = \frac{1}{\xi_0}
\begin{bmatrix}
\xi_0 & \eta_0 & \ldots & \eta_{n-1} \\
& \xi_1 & \xi_0 & \\
& & \ldots & \\
& & & \xi_0 \\
\end{bmatrix}
\begin{bmatrix}
\eta_{n-1} & \ldots & \eta_1 \\
& & \ldots & \\
\end{bmatrix}
- \begin{bmatrix}
\eta_0 \\
& \ldots & \ldots \\
& & \eta_{n-2} & \\
& & & \eta_0 \\
\end{bmatrix}
\begin{bmatrix}
\xi_{n-1} & \ldots & \xi_0 \\
& & \ldots & \\
\end{bmatrix}.
(2.12)
$$
Note that in view of Lemma 2.1 we could replace in Theorem 2.3 \( x \) and \( y \) by \( \tilde{u} \) and \( \tilde{v} \), respectively. Moreover, Theorem 2.2 is readily deduced from Theorem 2.1, and vice versa. Multiplication of (2.2) and (2.3) by the reversal matrix

\[
J := \begin{bmatrix}
  & & & 1 \\
 & & & \\
 & 1 & & \\
1 & & & \\
\end{bmatrix}
\]

yields, in view of

\[
JTJ = T^T, \tag{2.13}
\]

systems of type (2.1) and (2.4) for \( T^T \), and thus an inversion formula of type (2.10) for \( T^T \). Applying \( J \) on both sides to this formula and inserting factors \( J^2 \) between the factors of the matrix products shows that it is equivalent to (2.9). In the following we can therefore restrict our attention to Theorems 2.2 and 2.3. This remark makes it also clear why in (2.9) and (2.10) the upper and lower triangular matrices are exchanged.

Next, we derive another, less-known inversion formula based on solutions of (2.1) and (2.3) that satisfy \( \psi_n \neq 0 \). It turns out that the existence of such solutions depends not only on the nonsingularity of \( T \) but additionally on the nonsingularity of

\[
\tilde{T} := \begin{bmatrix}
\mu_1 & \mu_0 & \cdots & \mu_{-n+1} \\
\mu_2 & \mu_1 & \cdots & \mu_{-n+3} \\
\vdots & \vdots & \ddots & \vdots \\
\mu_n & \mu_{n-1} & \cdots & \mu_1 \\
\end{bmatrix}
\]

By analogy to the Gohberg-Semencul pair, it is no surprise that we will also find an inversion formula for \( \tilde{T} \) based on (2.1) and (2.3). Like all the other inversion formulas we have considered here, these two formulas can be seen to be special cases of Theorem 4.2 of Heinig and Rost [19]; see also [1, 18]. However, here we give a simple derivation from Heinig's formula. First, recall that if

\[
S := \begin{bmatrix}
0 & & & \\
1 & & & \\
& \ddots & \ddots & \\
& & 1 & 0 \\
\end{bmatrix}
\]
denotes the $n \times n$ shift matrix, then
\begin{equation}
TS^T - S^T T = fe_1^T - e_n f^T J. \tag{2.14}
\end{equation}

**Lemma 2.2.** Assume (2.1) and (2.3) have solutions. Then
\[ T(u_0 - Sx) = e_n \psi_n. \]
Moreover, if the solution $u$ of (2.1) satisfies $\psi_n \neq 0$, then $T$ and $\tilde{T}$ are nonsingular, and thus
\[ y = \left(u_0 - Sx\right) \frac{1}{\psi_n} \]
is the solution of (2.4), i.e.,
\[ \eta_k = \left(\xi_0 \psi_{k-1} - \xi_{k+1}\right) \frac{1}{\psi_n}, \quad k = 0, \ldots, n - 1, \quad \text{where} \quad \xi_n := 0. \tag{2.15} \]

**Proof.** By assumption, $Tu = f$ and $Tx = e_1$. From (2.13) we see that
\[ f^T Jx = u^T T^T Jx = u^T JTx = u^T e_n = \psi_n, \]
and by (2.14) we get
\[ T(u_0 - Sx) = f \xi_0 - S^T e_1 - f \xi_0 + e_n u^T J e_1 = e_n \psi_n. \]
If $\psi_n \neq 0$, it follows that $(u_0 - Sx) \psi_n^{-1}$ is a solution of (2.4), $Ty = e_n$. Hence, by Theorem 2.2, $T$ is nonsingular. Now notice that
\[ \tilde{T} = S^T T - e_n f^T J, \]
and therefore, since $JT^{-1} = T^{-T} J$,
\[ \tilde{T} T^{-1} = S^T - e_n u^T J. \tag{2.16} \]
The matrix on the right-hand side is a companion matrix; the element in the lower left corner is $-\psi_n$. Clearly, this matrix is nonsingular if and only if $\psi_n \neq 0$. Hence, when $T$ is nonsingular, $\psi_n \neq 0$ is equivalent to $\tilde{T}$ being nonsingular.

**Theorem 2.4.** Assume (2.1) and (2.3) have solutions, and $\psi_n \neq 0$. Then both $T$ and $\tilde{T}$ are nonsingular, and they can be represented as

\[
T^{-1} = \frac{1}{\psi_n} \begin{pmatrix}
\psi_n & \psi_{n-1} & \cdots & \psi_1 & 0 \\
\psi_n & \cdots & \psi_{n-1} & 0 & 0 \\
0 & \psi_n & \cdots & \psi_{n-1} & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \psi_n & 0
\end{pmatrix}
\begin{pmatrix}
\xi_0 \\
\xi_1 \\
\vdots \\
\xi_{n-1} \\
0
\end{pmatrix}
\begin{pmatrix}
1 \\
\psi_1 \\
\vdots \\
\psi_{n-1} \\
0
\end{pmatrix}
\begin{pmatrix}
0 \\
\xi_0 \\
\vdots \\
\xi_{n-2} \\
0
\end{pmatrix}
\]

and

\[
\tilde{T}^{-1} = \frac{1}{\psi_n} \begin{pmatrix}
\psi_n & \psi_{n-1} & \cdots & \psi_1 & 0 \\
\psi_n & \cdots & \psi_{n-1} & 0 & 0 \\
0 & \psi_n & \cdots & \psi_{n-1} & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \psi_n & 0
\end{pmatrix}
\begin{pmatrix}
\xi_{n-1} \\
\xi_{n-2} \\
\vdots \\
\xi_0 \\
0
\end{pmatrix}
\begin{pmatrix}
1 \\
\psi_1 \\
\vdots \\
\psi_{n-1} \\
0
\end{pmatrix}
\begin{pmatrix}
0 \\
\xi_0 \\
\vdots \\
\xi_{n-2} \\
0
\end{pmatrix}
\]

**Proof.** If (2.1) and (2.3) have solutions with $\psi_n \neq 0$, then, by Lemma 2.2, $T$ and $\tilde{T}$ are nonsingular. The inversion formula (2.17) follows immediately on replacing $\eta_k$ in (2.10) by the right-hand side of (2.15). Note that two matrix products cancel out.
For the proof of (2.18), recall that by (2.16), the inverse of $\tilde{T}$ is given by

$$\tilde{T}^{-1} = T^{-1}(S^T - e_nu^TJ)^{-1}.$$ 

Applying the inversion formula for companion matrices and inserting (2.17) yields

$$\tilde{T}^{-1} = T^{-1}[S - e_1\psi_n^{-1}(u^TJS + e_n^T)]$$

$$= \frac{1}{\psi_n} \left[ \begin{array}{ccccc} \psi_n & \psi_{n-1} & \cdots & \psi_1 & 0 \\ \psi_n & \cdots & \cdots & \cdots & \xi_0 \\ \vdots & \vdots & \ddots & \cdots & \xi_0 \\ 0 & \cdots & \cdots & \psi_n & 0 \\ 0 & \cdots & \cdots & 0 & \cdots & 0 \end{array} \right]$$

$$= \frac{1}{\psi_n} \left[ \begin{array}{cccccc} 0 & \xi_{n-1} & \xi_{n-2} & \cdots & \xi_1 \\ \xi_{n-1} & \cdots & \cdots & \cdots & \xi_0 \\ \vdots & \vdots & \ddots & \cdots & \xi_0 \\ 0 & \cdots & \cdots & \xi_{n-2} & \xi_{n-1} \\ 0 & \cdots & \cdots & 0 & \cdots & \cdots & 0 \end{array} \right]$$

$$\times \left[ \begin{array}{cccc} 0 & \cdots & \psi_1 & \cdots \\ \psi_1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \cdots \\ 0 & \cdots & \psi_{n-2} & \cdots \psi_1 & 0 \end{array} \right] = -x\psi_n^{-1}\left[ \begin{array}{cccc} \psi_{n-1} & \cdots & \psi_1 & 1 \end{array} \right].$$

(2.19)

Inserting the last term into the second matrix product completes the proof.

It is important that the assumptions of the above theorems hold whenever the respective matrices are nonsingular:

**Lemma 2.3.**

(i) If $T$ is nonsingular, the assumptions of Theorems 2.1 and 2.2 hold.

(ii) If $T$ and $T_{n-1}$ are nonsingular, the assumptions of Theorem 2.3 hold.
(iii) If $T$ and $\check{T}$ are nonsingular, the assumptions of Theorem 2.4 hold.

**Proof.** Obviously, the nonsingularity of $T$ implies the solvability of (2.1)-(2.4). Therefore, it remains to show that the assumptions in (ii) yield $\xi_0 \neq 0$ and the assumption in (iii) yield $\psi_n \neq 0$.

By Cramer's rule, we obtain from (2.3) and (2.4)

$$\xi_0 = \eta_{n-1} = \frac{\det T_{n-1}}{\det T}.$$ 

Hence, if $T$ is nonsingular, then the nonsingularity of $T_{n-1}$ is equivalent to $\xi_0 \neq 0$. This proves (ii). From (2.16), we have seen that if $T$ is nonsingular, $\psi_n \neq 0$ is equivalent to $\check{T}$ being nonsingular. ■

The assumptions for the inversion formulas of Theorems 2.1, 2.3, and 2.4 are exactly those in the definition of, respectively, regular, column-regular, and row-regular Padé forms [13] and basic pairs [14], which are fundamental for various fast and superfast Toeplitz solvers.

3. **WEAK STABILITY**

We want to show that the evaluation of the inversion formulas of the previous section is weakly stable. In general, an algorithm is called weakly stable [8] if for all well-conditioned problems the computed solution $\tilde{x}$ is close to the true solution $x$ in the sense that the relative error $\|x - \tilde{x}\|/\|x\|$ is small. (Some readers may prefer to call this forward stable.) Stability in the sense of Wilkinson (backward stability) or Bunch [8] means that the computed solution is the true solution of a slightly perturbed problem; it implies weak stability but not vice versa. In our situation, the evaluation is a well-conditioned problem if and only if the relevant matrices, namely those whose inverses can be computed from the data, are well conditioned.

We will also include perturbations of the data in our estimates. Assume we have computed solutions $\tilde{u}$, $\tilde{v}$, $\tilde{x}$, and $\tilde{y}$ of (2.1)-(2.4) which are perturbed by a normwise relative error bounded by $\tilde{e}$:

$$\|\tilde{u}\| \leq \|u\|(1 + \tilde{e}),$$

$$\|\tilde{v}\| \leq \|v\|(1 + \tilde{e}),$$

$$\|\tilde{x}\| \leq \|x\|(1 + \tilde{e}),$$

$$\|\tilde{y}\| \leq \|y\|(1 + \tilde{e}).$$
For our analysis, we use the following bounds on roundoff errors occurring
during matrix computations. Here, and in the sequel, $\| \cdot \|$ always denotes the
Euclidean or spectral norm, $\| \cdot \|_F$ the Frobenius norm.

**Lemma 3.1** [12, Section 2.4.8]. Let $A, B \in \mathbb{C}^{n \times n}$ and $\alpha \in \mathbb{C}$. Then in floating-point arithmetic with machine precision $\varepsilon$,

\[
\begin{align*}
\text{fl}(\alpha A) &= \alpha A + E, & \|E\|_F \leq \varepsilon |\alpha| \|A\|_F \leq \varepsilon |\alpha| \sqrt{n} \|A\|_F, \\
\text{fl}(A + B) &= A + B + E, & \|E\|_F \leq \varepsilon \|A + B\|_F \leq \varepsilon \sqrt{n} \|A + B\|_F, \\
\text{fl}(AB) &= AB + E, & \|E\|_F \leq n \varepsilon \|A\|_F \|B\|_F.
\end{align*}
\]

As usual, errors of order $O(\varepsilon^2)$, $O(\delta^2)$, and $O(\varepsilon \delta)$ will be ignored.

**Theorem 3.1.** If $T$ is well conditioned, and $|\mu_n|, |\mu_{-n}| \leq \gamma \|T\|$, then each of the formulas (2.9) and (2.10) is forward stable.

**Proof.** For abbreviation we will write the inversions formula (2.9) as

\[
T^{-1} = L_x R_v^1 - L_v R_x^0.
\]

Its evaluation in floating-point arithmetic, starting from the perturbed solutions $\tilde{x}$ and $\tilde{v}$, can be expressed as

\[
\tilde{T}^{-1} = \text{fl}\{(L_x + \delta L_x)(R_v^1 + \delta R_v^1) - (L_v + \delta L_v)(R_x^0 + \delta R_x^0)\}
\]

\[
= T^{-1} + \delta L_x R_v^1 + L_x \delta R_v^1 - \delta L_v R_x^0 - L_v \delta R_x^0 + E + F. \tag{3.20}
\]

Here, $E$ is the matrix containing the error which results from computing the
matrix products, and $F$ contains the error from subtracting the matrices. For the error matrices we have

\[
\begin{align*}
\|\delta R_x^0\|_F &\leq \|\delta L_x\|_F \leq \varepsilon \|L_x\|_F \leq \varepsilon \sqrt{n} \|x\|, \\
\|\delta R_v^1\|_F &\leq \varepsilon \|R_v^1\|_F \leq \varepsilon \sqrt{n} \sqrt{1 + \|v\|^2}, \\
\|\delta L_v\|_F &\leq \varepsilon \|L_v\|_F \leq \varepsilon \sqrt{n} \|v\|.
\end{align*}
\]
Then, Lemma 3.1 yields the following bounds on $E$ and $F$:

\[
\|E\|_F \leq n \varepsilon \left( \|L_x\|_F \|R_x^0\|_F + \|L_y\|_F \|R_y^0\|_F \right)
\]

\[
\leq n^2 \varepsilon \|x\| \left( \sqrt{1 + \|v\|^2} + \|v\| \right)
\]

\[
\leq n^2 \varepsilon \|x\| \left( 1 + \frac{3}{2} \|v\| \right),
\]

\[
\|F\| \leq \sqrt{n} \varepsilon \|T^{-1}\|.
\]

By (2.2) and (2.3), and by the assumption on $\mu_n$, $\|v\| \leq \sqrt{1 + \gamma^2 \|T\| \|T^{-1}\|}$ and $\|x\| \leq \|T^{-1}\|$. Consequently, by (3.20),

\[
\|T^{-1} - T^{-1}\| \leq n(2 \varepsilon + n \varepsilon) \|x\|(1 + \frac{3}{2} \|v\|) + \sqrt{n} \varepsilon \|T^{-1}\|
\]

\[
\leq n(2 \varepsilon + n \varepsilon) \|T^{-1}\| \left( 1 + \frac{3}{2} \sqrt{1 + \gamma^2 \|T^{-1}\| \|T\|} \right) + \sqrt{n} \varepsilon \|T^{-1}\|.
\]

Thus the relative error is bounded by

\[
\frac{\|T^{-1} - T^{-1}\|}{\|T^{-1}\|} \leq n(2 \varepsilon + n \varepsilon) \left[ 1 + \frac{3}{2} \sqrt{1 + \gamma^2 \kappa(T)} \right] + \sqrt{n} \varepsilon,
\]

where $\kappa(T)$ is the condition number of $T$.

The proof for the forward stability of (2.10) is completely analogous. The resulting bound is the same. \hfill \blacksquare

**Theorem 3.2.** If $T$ and $T_{n-1}$ are both well conditioned, then the Gohberg-Semencul formulas (2.11) and (2.12) are forward stable.

**Proof.** Assume $\|T\| \leq \tau$, $\|T^{-1}\| \leq \tau'$, and $\|T_{n-1}^{-1}\| \leq \tau_{n-1}'$. For abbreviation we will write (2.11) as

\[
T^{-1} = \frac{1}{\xi_0} \left( L_x R_y - L_y^0 R_x^0 \right).
\]

Now, the floating-point result of this formula applied to the perturbed
solutions $\tilde{x}$ and $\tilde{y}$ can be written as

$$
\tilde{T}^{-1} = \mathcal{Q}\left(\frac{1}{\xi_0} \left( L_x + \delta L_x \right) \left( R_y + \delta R_y \right) - \frac{1}{\xi_0} \left( L_y + \delta L_y \right) \left( R_x^0 + \delta R_x^0 \right)\right)
$$

$$
\equiv T^{-1} + \frac{1}{\xi_0} \left( \delta L_x R_y + L_x \delta R_y - \delta L_y R_x^0 - L_y \delta R_x^0 + E + F \right) + G.
$$

$E$ and $F$ have the same meaning as in the previous proof; $G$ represents the error of the multiplication by $1/\xi_0$. For the error matrices we have

$$
\| \delta R_x^0 \|_F \leq \| \delta L_x \|_F \leq \tilde{\varepsilon} \| L_x \|_F \leq \tilde{\varepsilon} \sqrt{n} \| x \|,
$$

$$
\| \delta L_y^0 \|_F \leq \| \delta R_y \|_F \leq \tilde{\varepsilon} \| R_y \|_F \leq \tilde{\varepsilon} \sqrt{n} \| y \|.
$$

Lemma 3.1 leads to the following bounds for $E$, $F$, and $G$:

$$
\| E \|_F \leq n \varepsilon \left( \| L_x \|_F \| R_y \|_F + \| L_y^0 \|_F \| R_x^0 \|_F \right) \leq 2n^2 \varepsilon \| x \| \| y \|,
$$

$$
\| F \| \leq \sqrt{n} \varepsilon \| T^{-1} \|,
$$

$$
\| G \| \leq \varepsilon \frac{1}{|\xi_0|} \left( \| L_x \|_F \| R_y \|_F + \| L_y^0 \|_F \| R_x^0 \|_F \right) \leq \frac{1}{|\xi_0|} 2n^2 \varepsilon \| x \| \| y \|.
$$

For an upper bound on $1/|\xi_0|$, note that (2.5) yields

$$
\| u \| \leq \| T_{n-1}^{-1} \| \| f \| \leq \tau_{n-1} \| T \| \leq \tau_{n-1} \tau.
$$

The equality (2.8) shows that

$$
\| x \| = |\xi_0| \sqrt{1 + \| u \| ^2},
$$

and hence, with (2.3) we obtain

$$
\frac{1}{|\xi_0|} \leq \frac{1}{|\xi_0|} \| T \| \| x \| \leq \sqrt{1 + \left( \tau_{n-1} \tau \right)^2} \leq \tau \left( 1 + \frac{1}{2} \tau_{n-1} \tau \right).
$$
Adding the bounds finally yields

\[
\|\tilde{T}^{-1} - T^{-1}\| \leq \frac{1}{|\xi_0|} \left[ \|x\|\|y\| (4n \tilde{e} + 2n^2 e + sn e) + \sqrt{n} e \|T^{-1}\| \right]
\]

\[
\leq \frac{1}{|\xi_0|} \|T^{-1}\| \left[ (2n\|T^{-1}\| [2 \tilde{e} + (n + 1) e] + \sqrt{n} e \right].
\]

Thus the relative error for this inversion formula is bounded by

\[
\frac{\|\tilde{T}^{-1} - T^{-1}\|}{\|T^{-1}\|} \leq \tau \left( 1 + \frac{1}{2} \tau'_n \right) \left[ 2n \tau' \left( 2 \tilde{e} + (n + 1) e \right) + \sqrt{n} e \right].
\]

This proves the forward stability of (2.11). The proof for (2.12) is done analogously. 

**Theorem 3.3.** If $T$ and $\tilde{T}$ are both well conditioned, and $|\mu_n| \leq \gamma \|T\|$, then the formulas (2.17) and (2.18) are forward stable.

**Proof.** Assume $\|T\| \leq \tau$, $\|T^{-1}\| \leq \tau'$, and $\|\tilde{T}^{-1}\| \leq \tilde{\tau}'$. For abbreviation we will write (2.17) as

\[
T^{-1} = \frac{1}{\psi_n} \left( R_u L_x - R_x^0 L_u^1 \right).
\]

Evaluation in floating-point arithmetic, with the perturbed solutions $\hat{x}$ and $\hat{u}$, yields

\[
\tilde{T}^{-1} = \text{fl} \left\{ \frac{1}{\psi_n} \left( R_u + \delta R_u \right)(L_x + \delta L_x) - \frac{1}{\psi_n} \left( R_x^0 + \delta R_x^0 \right)(L_u^1 + \delta L_u^1) \right\}
\]

\[
= T^{-1} + \frac{1}{\psi_n} \left( \delta R_u L_x + R_u \delta L_x - \delta R_x^0 L_u^1 - R_x^0 L_u^1 + E + F \right) + G.
\]

$E$, $F$, and $G$ are defined as in the proof of Theorem 3.2. Here, we have the
following bounds on the norms of the error matrices:

\[ \| \delta R^0_x \|_F \leq \| \delta L_x \|_F \leq \tilde{\delta} \| L_x \|_F \leq \tilde{\delta} \sqrt{n} \| x \|, \]
\[ \| \delta L^1_u \|_F \leq \tilde{\delta} \| L^1_u \|_F \leq \tilde{\delta} \sqrt{n} \sqrt{1 + \| u \|^2}, \]
\[ \| \delta R_u \|_F \leq \tilde{\delta} \| R_u \|_F \leq \tilde{\delta} \sqrt{n} \| u \| \]

and

\[ \| E \|_F \leq n \varepsilon \left( \| R_u \|_F \| L_x \|_F + \| R^0_x \|_F \| L^1_u \|_F \right) \]
\[ \leq n^2 \varepsilon \| x \| \left( \| u \| + \sqrt{1 + \| u \|^2} \right). \]
\[ \| F \| \leq \sqrt{n} \varepsilon \| T^{-1} \|, \]
\[ \| G \| \leq \varepsilon \frac{1}{| \psi_n |} \left( \| R_u \|_F \| L_x \|_F + \| R^0_x \|_F \| L^1_u \|_F \right) \]
\[ \leq \frac{1}{| \psi_n |} n \varepsilon \| x \| \left( \| u \| + \sqrt{1 + \| u \|^2} \right). \]

To derive an upper bound for \( 1/| \psi_n | \), note that (2.19) yields

\[ e_n^T \tilde{T} T^{-1} e_n = e_n^T \left[ \mathbf{S} - e_n \psi_n^{-1} (u^T J S + e_n^T) \right] e_n = - \frac{1}{\psi_n}. \]

Therefore,

\[ \frac{1}{| \psi_n |} \leq \| T \| \| \tilde{T}^{-1} \| \leq \tau \tilde{\tau}'. \]

With the assumption for \( \mu_n \), this yields

\[ \| T^{-1} - T \|
\[ \leq \frac{1}{| \psi_n |} \left( \| x \|(1 + \frac{3}{2} \| u \|)(2n \tilde{\delta} + n^2 \varepsilon + n \varepsilon) + \sqrt{n} \varepsilon \| T^{-1} \| \right) \]
\[ \leq \tau \tilde{\tau}' \| T^{-1} \| \left( n \left[ 1 + \frac{3}{2} \sqrt{1 + \gamma^2 \| T \| \| T^{-1} \|} \right] \left[ 2 \tilde{\delta} + (n + 1) \varepsilon \right] + \sqrt{n} \varepsilon \right). \]
Hence, the relative error is bounded by

\[
\frac{\|\hat{T}^{-1} - T^{-1}\|}{\|T^{-1}\|} \leq \tau' \left\{ n \left[ 1 + \frac{3}{2} \sqrt{1 + \gamma^2 \kappa(T)} \right] \left[ 2\tilde{\varepsilon} + (n + 1)\varepsilon \right] + \sqrt{n} \varepsilon \right\}.
\]

The proof of stability of (2.18) follows the same lines and yields the same bound.

The authors are indebted to a referee for bringing References [1, 18, 19, 24] to their attention and for pointing out that the inversion formulas (2.17) and (2.18) are special cases of a general result in [19].

REFERENCES


---


---


Received 15 November 1993; final manuscript accepted 26 September 1994