Numerical Matrix Analysis
Notes #13 — Conditioning and Stability:
Stability of Back Substitution

Peter Blomgren
⟨blomgren@sdsu.edu⟩

Department of Mathematics and Statistics
Dynamical Systems Group
Computational Sciences Research Center
San Diego State University
San Diego, CA 92182-7720

http://terminus.sdsu.edu/

Spring 2023
(Revised: March 1, 2023)
1. Looking Back
   - Stability of Householder Triangularization

2. Backward Stability of Back Substitution
   - Introduction: Algorithm, Conventions, Axioms, and Theorem
   - Proof
   - Comments
Looking Back

Backward Stability of Back Substitution

Last Time: Stability of Householder Triangularization

— We discussed the stability properties of QR-factorization by Householder Triangularization (HT-QR).
— Numerical “evidence” that HT-QR is backward stable.
— Statement (proof by reference to Higham’s *Accuracy and Stability of Numerical Algorithms*) that HT-QR is backward stable

— Showed that solving \( A\vec{x} = \vec{b} \) using HT-QR and backward substitution is backward stable, assuming that

1. \( QR = A \) by HT-QR is backward stable
2. \( \tilde{w} = Q^*\vec{b} \) is backward stable
3. \( R\tilde{x} = \tilde{w} \) by back substitution is backward stable

— Today: Explicit proof of (3), and implicit proof of (2).
Backward Stability of Back Substitution

Back substitution is one of the **easiest non-trivial algorithms** we study in numerical linear algebra, and is therefore a good venue for a full backward stability proof.

The proof for backward stability of Householder triangularization follows the same pattern, but the details become more cumbersome.

Back-substitution applies to $R\vec{x} = \vec{b}$, where

$$
\begin{bmatrix}
    r_{11} & r_{12} & \cdots & r_{1m} \\
    r_{22} & r_{2m} & & \\
    \vdots & \vdots & & \\
    r_{mm} & & & \\
\end{bmatrix}
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_m \\
\end{bmatrix}
= 
\begin{bmatrix}
    b_1 \\
    b_2 \\
    \vdots \\
    b_m \\
\end{bmatrix}
$$

Upper (and lower) triangular matrices are generated by, *e.g.* the QR-factorization [NOTES#6–7], Gaussian elimination [NOTES#16–17], and the Cholesky factorization [NOTES#17].

Peter Blomgren (blomgren@sdsu.edu)

13. Stability of Back Substitution
Looking Back
Backward Stability of Back Substitution

Algorithm: Back-Substitution

Algorithm (Back-Substitution)

1: \( x_m \leftarrow b_m / r_{mm} \)
2: for \( \ell \in \{(m - 1), \ldots, 1\} \) do
3: \( x_\ell \leftarrow \left( b_\ell - \sum_{k=\ell+1}^{m} x_k r_{\ell k} \right) / r_{\ell \ell} \)
4: end for

Note that the algorithm breaks if \( r_{\ell \ell} = 0 \) for some \( \ell \).

For this discussion we make the assumption that \( b_\ell - \sum (x_k r_{\ell k}) \) is computed as \((m - \ell)\) subtractions performed in \(k\)-increasing order.

Simplification: In the theorem/proof, we use the convention that if the denominator in a statement like \( \frac{\delta r_{i \ell}}{|r_{i \ell}|} \leq m \varepsilon_{\text{mach}} \) is zero, we implicitly assert that the numerator is also zero, as \( \varepsilon_{\text{mach}} \to 0 \).
Reference: Key Floating Point Axioms

Floating Point Representation Axiom

∀x ∈ ℝ, there exists ε with |ε| ≤ ε_{mach}, such that fl(x) = x(1 + ε).

The Fundamental Axiom of Floating Point Arithmetic

For all x, y ∈ ℱₙ (where ℱₙ is the set of n-bit floating point numbers), there exists ε with |ε| ≤ ε_{mach}, such that

\[ x ⊕ y = (x + y)(1 + \epsilon), \quad x ⊖ y = (x - y)(1 + \epsilon), \]
\[ x ⊗ y = (x * y)(1 + \epsilon), \quad x ⊙ y = (x/y)(1 + \epsilon) \]
Back-Substitution: Backward Stability Theorem

Theorem (Solving an Upper Triangular System $R\tilde{x} = \tilde{b}$ Using Back-Substitution is Backward Stable)

Let the back-substitution algorithm be applied to $R\tilde{x} = \tilde{b}$, where $R \in \mathbb{C}^{m \times m}$ is upper triangular, $\tilde{b}, \tilde{x} \in \mathbb{C}^m$, in a floating-point environment satisfying the floating point axioms. The algorithm is backward stable in the sense that the computed solution $\tilde{x} \in \mathbb{C}^m$ satisfies

$$(R + \delta R)\tilde{x} = \tilde{b}$$

for some upper triangular $\delta R \in \mathbb{C}^{m \times m}$ with

$$\frac{\|\delta R\|}{\|R\|} = \mathcal{O}(\varepsilon_{\text{mach}}).$$

Specifically, for each $i, \ell$

$$\frac{|\delta r_{i\ell}|}{|r_{i\ell}|} \leq m\varepsilon_{\text{mach}} + \mathcal{O}(\varepsilon_{\text{mach}}^2).$$
Proof: \( m = 1 \)

When \( m = 1 \), back substitution terminates in one step

\[
\tilde{x}_1 = b_1 \odot r_{11}
\]

The error introduced in this step is captured by

\[
\tilde{x}_1 = \frac{b_1}{r_{11}} (1 + \epsilon_1 \odot), \quad |\epsilon_1 \odot| \leq \varepsilon_{\text{mach}}.
\]

Since we want to express the error in terms of perturbations of \( R \), we write

\[
\tilde{x}_1 = \frac{b_1}{r_{11}(1 + \epsilon_1')}, \quad |\epsilon_1'| \leq \varepsilon_{\text{mach}} + O(\varepsilon^2_{\text{mach}}).
\]

Hence,

\[
(r_{11} + \delta r_{11})\tilde{x}_1 = b_1, \quad \frac{|\delta r_{11}|}{|r_{11}|} \leq \varepsilon_{\text{mach}} + O(\varepsilon^2_{\text{mach}}) = O(\varepsilon_{\text{mach}}).
\]
A Note on \((1 + \epsilon)\) and \(1/(1 + \epsilon')\)

In backward stability proofs we frequently need to move terms of the type \((1 + \epsilon)\) from/to the numerator to/from the denominator. We do this because we want to express all the floating point errors as perturbations to a specific part of the expression, e.g. the matrix \(R\) in the instance of backward substitution.

When \(\epsilon\) is small, we can set

\[
\epsilon' = \frac{-\epsilon}{1 + \epsilon} \sim -\epsilon(1 - \epsilon + O(\epsilon^2)) = -\epsilon + O(\epsilon^2)
\]

and thus (throwing away \(O(\epsilon^2)\)-terms)

\[
1 + \epsilon' = \frac{1 + \epsilon}{1 + \epsilon} - \frac{\epsilon}{1 + \epsilon} = \frac{1 + \epsilon - \epsilon}{1 + \epsilon} = \frac{1}{1 + \epsilon} \quad \Rightarrow \quad \frac{1}{1 + \epsilon'} = 1 + \epsilon.
\]

**Bottom line:** we can move \((1 + \epsilon)\) terms (where \(|\epsilon| \leq \epsilon_{\text{mach}} \ll 1\)) between the numerator and denominator, and only introduce errors of the order \(O(\epsilon_{\text{mach}}^2)\), i.e. \(|\epsilon'| \leq \epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}^2)\).
Step one (which computes $\tilde{x}_2$) is exactly like the $m = 1$ case:

$$\tilde{x}_2 = \frac{b_2}{r_{22}(1 + \epsilon_1^\Theta)}, \quad |\epsilon_1| \leq \epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}^2).$$

The second step is defined by

$$\tilde{x}_1 = (b_1 \ominus (\tilde{x}_2 \odot r_{12})) \odot r_{11}.$$
Proof: \( m = 2 \)

As before, we can shift the \((1 + \epsilon_3^\oplus)\) and \((1 + \epsilon_4^\otimes)\) terms to the denominator

\[
\tilde{x}_1 = \frac{b_1 - \tilde{x}_2 r_{12}(1 + \epsilon_2^\otimes)}{r_{11}(1 + \epsilon_3^\oplus)(1 + \epsilon_4^\otimes)} = \frac{b_1 - \tilde{x}_2 r_{12}(1 + \epsilon_2^\otimes)}{r_{11}(1 + 2\epsilon_5^\oplus, \otimes)}
\]

where \(|\epsilon_3^\text{'}_4|, |\epsilon_5| \leq \epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}^2)\).

Now

\[
(R + \delta R)\tilde{x} = \tilde{b}
\]

since \(r_{11}\) is perturbed by the factor \((1 + 2\epsilon_5^\oplus, \otimes)\), \(r_{12}\) by the factor \((1 + \epsilon_2^\otimes)\), and \(r_{22}\) by the factor \((1 + \epsilon_1^\otimes)\). The entries satisfy

\[
\begin{bmatrix}
|\delta r_{11}|/|r_{11}| & |\delta r_{12}|/|r_{12}| \\
|\delta r_{22}|/|r_{22}|
\end{bmatrix} = \begin{bmatrix}
2|\epsilon_5^\oplus, \otimes| & |\epsilon_2^\otimes| \\
|\epsilon_5^\otimes| & |\epsilon_1^\otimes|
\end{bmatrix} \leq \begin{bmatrix}
2 & 1 \\
1 & 1
\end{bmatrix} \epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}^2)
\]

Thus \(\|\delta R\|/\|R\| = O(\epsilon_{\text{mach}})\).
Proof: $m = 3$

The first two steps are as before, and we get

\[
\begin{align*}
\tilde{x}_3 &= b_3 \odot r_{33} = \frac{b_3}{r_{33}(1 + \epsilon_1)} \\
\tilde{x}_2 &= (b_2 \ominus (\tilde{x}_3 \otimes r_{23})) \odot r_{22} = \frac{b_2 - \tilde{x}_3 r_{23} (1 + \epsilon_2)}{r_{22}(1 + 2\epsilon_3, \ominus)}
\end{align*}
\]

where superscripts on $\epsilon$s indicate the source operation; now

\[
\begin{bmatrix}
2|\epsilon_3| & |\epsilon_2| \\
|\epsilon_1|
\end{bmatrix}
\leq
\begin{bmatrix}
2 & 1 \\
1 & 1
\end{bmatrix}
\varepsilon_{\text{mach}} + O(\varepsilon^2_{\text{mach}})
\]

We take a deep breath, and write down the third step

\[
\tilde{x}_1 = [(b_1 \ominus (\tilde{x}_2 \otimes r_{12})) \ominus (\tilde{x}_3 \otimes r_{13})] \odot r_{11}
\]
Proof: $m = 3$

We expand the two $\otimes$ operations, and write

$$\tilde{x}_1 = [(b_1 \oplus \tilde{x}_2 r_{12}(1 + \epsilon_4^\otimes)) \oplus \tilde{x}_3 r_{13}(1 + \epsilon_5^\otimes)] \otimes r_{11}$$

We introduce error bounds for the $\oplus$ operations

$$\tilde{x}_1 = [(b_1 - \tilde{x}_2 r_{12}(1 + \epsilon_4^\otimes))(1 + \epsilon_6^\oplus) - \tilde{x}_3 r_{13}(1 + \epsilon_5^\otimes)] (1 + \epsilon_7^\oplus) \oplus r_{11}$$

Finally, we convert $\otimes$ to a mathematical division with a perturbation $\epsilon_8$; and move both the $(1 + \epsilon_{7,8})$ expressions to the denominator

$$\tilde{x}_1 = \frac{(b_1 - \tilde{x}_2 r_{12}(1 + \epsilon_4^\otimes))(1 + \epsilon_6^\oplus) - \tilde{x}_3 r_{13}(1 + \epsilon_5^\otimes)}{r_{11}(1 + \epsilon_7^\oplus)(1 + \epsilon_8^\otimes)}$$

As it stands, we have introduced a perturbation in $b_1$. This was not our intention, so we ship $(1 + \epsilon_7^\oplus)$ to the denominator as well...
Proof: \( m = 3 \)

We now have an expression with perturbations in only \( r_{1\ell} \):

\[
\tilde{x}_1 = \frac{b_1 - \tilde{x}_2 r_{12} (1 + \epsilon_4) - \tilde{x}_3 r_{13} (1 + \epsilon_5) (1 + \epsilon_6')}{r_{11} (1 + \epsilon_6') (1 + \epsilon_7')(1 + \epsilon_8')}
\]

where \( |\epsilon_{4,5}| \leq \varepsilon_{\text{mach}} \), and \( |\epsilon_{6,7,8}'| \leq \varepsilon_{\text{mach}} + O(\varepsilon_{\text{mach}}^2) \).

If we collect the limits on the relative sizes of the perturbations \( |\delta r_{i\ell}|/|r_{i\ell}| \) we get the following 6 relations

\[
\begin{bmatrix}
|\delta r_{11}|/|r_{11}| & |\delta r_{12}|/|r_{12}| & |\delta r_{13}|/|r_{13}|
|\delta r_{22}|/|r_{22}| & |\delta r_{23}|/|r_{23}|
|\delta r_{33}|/|r_{33}|
\end{bmatrix} \leq \begin{bmatrix}
3 & 1 & 2
2 & 1 & 1
\end{bmatrix} \varepsilon_{\text{mach}} + O(\varepsilon_{\text{mach}}^2)
\]

We are now ready to identify the pattern for general values of \( m \)...
The division by $r_{ij}$ induces perturbations $\delta r_{ij}$ only, since we always immediately shift that $(1 + \varepsilon_*)$-term to the denominator $1/(1 + \varepsilon')$, hence the perturbation pattern is of the form

$$\otimes \leadsto I_{n \times n} \varepsilon_{\text{mach}} + O(\varepsilon_{\text{mach}}^2)$$

The multiplications $\tilde{x}_i r_{\ell i}$ induces perturbations $\delta r_{\ell i}$ of relative size $\leq \varepsilon_{\text{mach}}$, the perturbation pattern is of the form

$$\otimes \leadsto \begin{bmatrix}
0 & 1 & 1 & \ldots & 1 \\
0 & 1 & \ldots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 \\
0 & 0
\end{bmatrix} \varepsilon_{\text{mach}}$$
Proof: General $m$

The most complicated contribution comes from the subtractions (and this is where the order of evaluation has an effect on the answer) — in computing $\tilde{x}_k$

\[
\begin{array}{ll}
    r_{k,k} & \text{is perturbed by} \quad (1 + \epsilon'_*)^{m-k} \\
    r_{k,k+1} & \text{is perturbed by} \quad 0 \\
    r_{k,k+2} & \text{is perturbed by} \quad (1 + \epsilon'_*) \\
    r_{k,k+3} & \text{is perturbed by} \quad (1 + \epsilon'_*)^2 \\
    \vdots & \\
    r_{k,m} & \text{is perturbed by} \quad (1 + \epsilon'_*)^{m-k-1}
\end{array}
\]

See next slide for the pattern.
Proof: General $m$

\[
\Theta \sim \begin{pmatrix}
(m-1) & 0 & 1 & 2 & 3 & \ldots & (m-2) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
4 & 0 & 1 & 2 & 3 \\
3 & 0 & 1 & 2 \\
2 & 0 & 1 \\
1 & 0 \\
0
\end{pmatrix} \epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}^2)
\]

Putting all this together gives...

Peter Blomgren (blomgren@sdsu.edu)
Proof: General $m —$ Collecting It All

$$\frac{|\delta R|}{|R|} \leq \begin{bmatrix} m & 1 & 2 & 3 & 4 & \ldots & (m - 1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 5 & 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 1 \\ 1 \end{bmatrix} \varepsilon_{\text{mach}} + O(\varepsilon^2_{\text{mach}})$$

Which completes the proof. $\square$
This is the standard approach for a backward stability analysis.

Errors introduced by the floating point operations $\oplus$, $\ominus$, $\otimes$, and $\oslash$ (in accordance with the axiom) are reinterpreted as errors in the initial data / or “problem.”

Where appropriate, errors $\sim \mathcal{O}(\varepsilon_{\text{mach}})$ are freely moved between numerators and denominators.

Perturbations of order $\mathcal{O}(\varepsilon_{\text{mach}})$ are accumulated additively, e.g.

$$(1 + \epsilon_1)(1 + \epsilon_2) = (1 + 2\epsilon_3) + \mathcal{O}(\varepsilon_{\text{mach}}^2)$$

where $|\epsilon_{1,2,3}| \leq \varepsilon_{\text{mach}}$. 

Peter Blomgren  ⟨blomgren@sdsu.edu⟩  13. Stability of Back Substitution  — (19/20)
Next, we turn our attention back to least squares problems.

— We take a detailed look at the **conditioning** of least squares problems; it is a subtle topic and has nontrivial implications for the **stability** (and ultimately, the **accuracy**) of least squares algorithms.

— Further, this will serve as our main example on detailed conditioning analysis (as Back-substitution served as the main example on detailed backward stability analysis).