7.6 Accuracy of a Backward Stable Algorithm

Theorem: Suppose a backward stable algorithm is used to solve $f : x \to y$ with condition number κ on a computer satisfying

$$fl(x \odot y) = (x \odot y)(1 + \delta) \qquad |\delta| \leq \epsilon_{\text{machine}} \tag{7.17}$$

then

$$\frac{\left\| \tilde{f}(x) - f(x) \right\|}{\|f(x)\|} = O \quad (\kappa(x) \cdot \epsilon_{\text{machine}}) \tag{7.18}$$

Proof: By definition $\tilde{f}(x) = f(\tilde{x})$ for $\tilde{x} : \frac{\|\tilde{x}-x\|}{\|x\|} = O(\epsilon_{\text{machine}})$.

(Continued on next page.)

$$\kappa(x) = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \frac{\frac{\|\delta f\|}{\|f\|}}{\frac{\|\delta x\|}{\|x\|}} \tag{7.19}$$

$$\begin{aligned} \frac{\|\delta f\|}{\|f\|} &\leq (\kappa(x) + O(1)) \frac{\|\delta x\|}{\|x\|} \\ &= O(\kappa(x)\epsilon_{\text{machine}}) \end{aligned} \tag{7.20}$$

Let P be an exact Householder or Givens, $\tilde{P} = fl(P)$, then

$$fl(\tilde{P}A) = PA + E \tag{7.21}$$

$$\|E\|_2 = O(\epsilon) \|A\|_2 \tag{7.22}$$

Proof: Only for Givens

$$G = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \tag{7.23}$$

$$a = \begin{bmatrix} x \\ y \end{bmatrix} \tag{7.24}$$

$$\begin{aligned} fl(\tilde{G}a) &= fl\left(\begin{bmatrix} \tilde{c} & \tilde{s} \\ -\tilde{s} & \tilde{c} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right) \\ &= fl\begin{pmatrix} \tilde{c}x + \tilde{s}y \\ -\tilde{s}x + \tilde{c}y \end{pmatrix} \\ &= \begin{pmatrix} cx(1+\delta_1)(1+\delta_2)(1+\delta_3) + sy(1+\delta_4)(1+\delta_5)(1+\delta_6) \\ \cdots \end{pmatrix} \end{aligned} \tag{7.25}$$

$$\left\| fl(\tilde{G}a) - Ga \right\|_2 = O(\epsilon) \|Ga\|_2 \tag{7.26}$$

$$\begin{aligned} \|E\|_2 &= \left\| fl(\tilde{G}A) - GA \right\|_2 \\ &= O(\epsilon) \|GA\|_2 \\ &= O(\epsilon) \|A\|_2 \end{aligned} \tag{7.27}$$

Similarly for Householder.

7.7 Backward Substitution (L-lower Triangular)

If $Lx = b$, and $\tilde{x}$ is the floating point solution, then

$$(L + \delta L)\tilde{x} = b \tag{7.28}$$

$$\|\delta L\| \leq \|L\| O(\epsilon_{\text{machine}}) \tag{7.29}$$

Proof:

$$
\begin{aligned}
l_{11}x_1 &= b_1 && (7.30)\\
l_{21}x_1 + l_{22}x_2 &= b_2 && (7.31)\\
&\cdots\\
l_{n1}x_1 + \cdots + l_{nn}x_n &= b_n && (7.32)
\end{aligned}
$$

then

$$
\begin{aligned}
\tilde{x}_1 &= \frac{b_1}{l_{11}}(1+\delta_1)\\
&= \frac{b_1}{\frac{l_{11}}{(1+\delta_1)}}\\
&= \frac{b_1}{\tilde{l}_{11}} && (7.33)\\
\tilde{x}_2 &= \frac{(b_2 - l_{21}\tilde{x}_1(1+\delta_2))(1+\delta_3)}{l_{22}}(1+\delta_4)\\
&= \frac{b_2 - l_{21}(1+\delta_2)\tilde{x}_1}{\frac{l_{22}}{(1+\delta_3)(1+\delta_4)}}\\
&= \frac{b_2 - \tilde{l}_{21}\tilde{x}_1}{\tilde{l}_{22}} && (7.34)\\
&\cdots\\
\tilde{x}_n &= \frac{((b_n - l_{n1}\tilde{x}_1(1+\delta_{n1}))(1+\delta'_{n1}) - l_{n2}\tilde{x}_2(1+\delta_{n2}))(1+\delta_{n2})}{l_{nn}}\\
&= \frac{b_n - l_{n1}(1+\delta_{ij}\ \ products)\tilde{x}_1 - l_{n2}(1+\delta_{ij}\ \ products)\tilde{x}_2}{\frac{l_{nn}}{1+\delta_{ij}\ \ products}} && (7.35)
\end{aligned}
$$