Precisión de Algoritmos Backward-Stable

De las discusiones anteriores vimos que la precisión de un algoritmo back-st. depende del condicionamiento del problema. Podemos concluir este resultado con el siguiente teorema.

Teo: Sea $S: I \to O$ un problema (input-output map) y sea $\tilde{S}$ un algoritmo back-st que aproxima S. Luego la precisión del output satisface.

$$\|\tilde{S}(a) - S(a)\| \leq O(\mu(a)\,\varepsilon_{machine}) \cdot \|S(a)\|$$

Demo. Como $\tilde{S}$ es back-st. si tiene $\tilde{S}(a) = S(\tilde{a})$ para cierto $\tilde{a}$ / $\dfrac{\|\tilde{a}-a\|}{\|a\|} = O(\varepsilon_{machine})$. Luego tenemos

$$\|\tilde{S}(a) - S(a)\| = \|S(\tilde{a}) - S(a)\| \leq \|DS(a)\| \cdot \|a - \tilde{a}\| + o(\|a-\tilde{a}\|)$$

$$= \left(\frac{\|DS(a)\|}{\|S(a)\|} \cdot \|a\|\right) \frac{\|a-\tilde{a}\|}{\|a\|} \cdot \|S(a)\| + o(\|\tilde{a}-a\|)$$

$$= \mu(a) \cdot O(\varepsilon_{machine}) \|S(a)\| + o(\|\tilde{a}-a\|).$$

Pensando que podemos asumir que $\varepsilon_{machine} \to 0$ tenemos que el término $o(\|\tilde{a}-a\|) = o(\|a\| \cdot O(\varepsilon_{mach})) \to 0$. □

Estabilidad de Triangularización Householder

Comencemos realizando experimentos en MATLAB para ver la precisión de este algoritmo. (Ver notebook Estabilidad-Householder)

Lo que está de fondo es el siguiente resultado (que enunciamos sin prueba)

Teo: Consideremos una máquina con la Aritmética de P.F. (satisfaciendo axiomas • $\forall x \in \mathbb{R}$, $\exists \varepsilon$, $|\varepsilon| \leq \varepsilon_{mach}$ tq $fl(x) = x(1+\varepsilon)$
• $\forall x, y \in \mathbb{F}$, $\exists \varepsilon$ tq $|\varepsilon| \leq \varepsilon_{mach}$ tq $x \circledast y = x * y (1+\varepsilon)$)

El algoritmo Householder para computar la descomp. QR de una matriz $A \in \mathbb{C}^{m \times n}$, es back-st., i.e. el output $\hat{Q}, \hat{R}$ del algoritmo satisface

$$\hat{Q} \cdot \hat{R} = A + \delta A \quad \text{con} \quad \frac{\|\delta A\|}{\|A\|} = O(\varepsilon_{mach})$$

para alguna matriz $\delta A \in \mathbb{C}^{m \times n}$.

Comentarios: En otras palabras, la $\hat{Q}$ computada ($\hat{Q}$) y la $\hat{R}$ computada ($\hat{R}$) son ~~solución~~ la factorización QR de un problema (levemente) perturbado.

La matriz $\hat{R}$ es exactamente la matriz triangular superior que resulta de las aplicaciones de las reflexiones Householder en punto flotante.

Sin embargo la $\tilde{Q}$ es una verdadera matriz ortogonal que resulta de tomar el producto $\overbrace{Id - 2\frac{\tilde{v}_k \tilde{v}_k^t}{\tilde{v}_k^t \tilde{v}_k}}^{H}$ (las reflexiones H) pero donde $\tilde{v}_k$ son los vectores en punto flotante ($\tilde{Q} = \tilde{Q}_1 \cdots \tilde{Q}_n$, $\tilde{Q}_k = H$)

Esta restricción en el resultado en realidad no es tal dado que si recordamos el algoritmo Householder la matriz Q no la escribimos explícita (solo actualizamos la matriz A hasta llegar a una triangular superior).

- Como vimos, individualmente $\tilde{Q}$ y $\tilde{R}$ no son precisos pero en realidad lo que ocurre es que se supone que el problema de encontrar Q (y R) está mal condicionado. (Sería interesante saber hasta qué punto este resultado está probado teóricamente.)
- En general, la descomposición QR no es en sí el mismo pero un medio para estudiar o resolver otro problema. Recordar que por ejemplo nos da una forma sencilla de resolver un sistema lineal $Ax = b$.

La pregunta es si los errores individuales en Q y R pueden afectar la precisión de resolver el sistema. La respuesta es que felizmente no, alcanza con la precisión conjunta de Q.R.

Analicemos el algoritmo de resolver $Ax=b$: Alg-SL

Recordar que el algoritmo consiste en

i) $QR=A$ # realizar factorización vía Reflexiones H.

ii) $y=Q^*b$ # se construye Q^*b

iii) $x=R^{-1}y$ # resolvemos el sistema triangular $Rx=y$ por sustitución "hacia atrás".

Se puede probar que i), ii) y iii) son back-st. (i) ya lo mencionamos) y esto tiene como resultado que el Alg-SL también lo es.

El paso ii) realiza el cálculo $\tilde{Q}^*b$ mediante el alg. ya visto en clase (calculo implícito). Dice que back-st. implica que el output $\tilde{y}$ satisface

$$(\tilde{Q}+\delta Q)\tilde{y}=b \quad \text{con} \quad \|\delta Q\|=O(\varepsilon_{mach}) \qquad (I)$$

Análogamente el paso iii) da un output $\tilde{x}$ que satisface

$$(\tilde{R}+\delta R)\tilde{x}=\tilde{y} \quad \text{con} \quad \frac{\|\delta R\|}{\|\tilde{R}\|}=O(\varepsilon_{mach}) \qquad (II)$$

Juego de (I) y (II) tenemos que

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

Utilizando que el Alg. Householder es b-s. tenemos ($A+\delta A = \tilde{Q}\tilde{R}$)

$$b = (A + \underbrace{\delta A + \delta Q \cdot \tilde{R} + \tilde{Q} \cdot \delta R + \delta Q \cdot \delta R}) \tilde{x}$$

i.e. $b = (A + \Delta A)\tilde{x}$, por lo que resta probar que $\|\Delta A\|$ es "relativamente chico", i.e.

$$\frac{\|\Delta A\|}{\|A\|} \leq O(\varepsilon_{mach}).$$ Para eso observamos que

$\tilde{R} = \tilde{Q}^*(A+\delta A) \Rightarrow \frac{\|\tilde{R}\|}{\|A\|} = \frac{\|A+\delta A\|}{\|A\|} = O(1)$

$$\frac{\|\delta Q \tilde{R}\|}{\|A\|} \leq \|\delta Q\| \cdot \frac{\|\tilde{R}\|}{\|A\|} \leq O(\varepsilon_{mach}) \underbrace{\frac{\|\tilde{R}\|}{\|A\|}}_{=O(1)} = O(\varepsilon_{mach})$$

Análogo para $\frac{\|\tilde{Q}\delta R\|}{\|A\|} = O(\varepsilon_{mach})$

Y además $\frac{\|\delta Q \delta R\|}{\|A\|} \leq O(\varepsilon_{mach}^2)$, por lo tanto,

$$\frac{\|\Delta A\|}{\|A\|} \leq \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta Q \tilde{R}\|}{\|A\|} + \frac{\|\tilde{Q}\delta R\|}{\|A\|} + \frac{\|\delta Q \delta R\|}{\|A\|}$$

$$\leq O(\varepsilon_{mach}) \square$$