Multiplication by a unitary matrix is backward-stable

This is an idea which I think should have been in the text[1] itself, and not just in Exercise 16.1 (a). Its proof uses an idea not seen in other "show the algorithm is backward-stable" arguments. We start in an unexpected way, by bounding the forward error $\|\tilde{f}(A) - f(A)\|$. Then the combination of unitarity and linearity allows us to transfer the forward error to a backward error $\|\tilde{A} - A\|$ using an input $\tilde{A}$ for which $\tilde{f}(A) = f(\tilde{A})$.

Theorem 16.0. *Fix $Q \in \mathbb{C}^{m\times m}$ unitary. On a computer satisfying (13.5) and (13.7), the obvious matrix-matrix multiplication algorithm is backward-stable for the problem*

$$f(A) = QA, \qquad A \in \mathbb{C}^{m\times n}.$$

Proof. Each entry of the product QA is an inner product $g(y) = x^* y$. The obvious algorithm for inner products is backward stable, so that $\tilde{g}(y) = g(\tilde{y})$ where $\tilde{y} = y + \delta y$ with $\|\delta y\|_2 \le c(m)\epsilon_{\text{m}} \|y\|_2$ with some constant $c(m)$ independent of y and ϵ_{m}.

Consider the i, j entry of the product QA. To apply the above idea, let $x = q_i^*$ be the ith row of Q and denote the jth column of A by a_j as usual. Note that a row of a unitary matrix has unit 2-norm. By the Cauchy-Schwarz inequality,

$$\begin{aligned} |\tilde{f}(A)_{ij} - f(A)_{ij}| &= |\tilde{g}(a_j) - g(a_j)| = |q_i^*(a_j + \delta a_j) - q_i^* a_j| \\ &= |q_i^* \delta a_j| \le \|q_i^*\|_2 \|\delta a_j\|_2 = \|\delta a_j\|_2 \le c(m)\epsilon_{\text{m}} \|a_j\|_2. \end{aligned}$$

In this calculation "δa_j" actually varies with (depends on) both i and j, but the final bound is independent of i.

This entry-wise bound can be advanced to a Frobenius norm bound. That is,

$$\begin{aligned} \|\tilde{f}(A) - f(A)\|_F^2 &= \sum_{\substack{i=1,\dots,m \\ j=1,\dots,n}} |\tilde{f}(A)_{ij} - f(A)_{ij}|^2 \le \sum_{i,j} c(m)^2 \epsilon_{\text{m}}^2 \|a_j\|_2^2 \\ &= m\, c(m)^2 \epsilon_{\text{m}}^2 \sum_j \|a_j\|_2^2 = m\, c(m)^2 \epsilon_{\text{m}}^2 \|A\|_F^2. \end{aligned}$$

Note that the sum over i simply gives a factor of m and that $\sum_{j=1}^n \|a_j\|_2^2 = \|A\|_F^2$. Thus

$$\|\tilde{f}(A) - f(A)\|_F \le \sqrt{m}\, c(m)\epsilon_{\text{m}} \|A\|_F.$$

Now we change tacks and describe the forward error as a backward error. Let

$$\delta A = Q^*(\tilde{f}(A) - f(A))$$

[1] Trefethen & Bau, *Numerical Linear Algebra*, SIAM Press, 1997.

so that $Q\delta A = \tilde{f}(A) - f(A)$. Observe that

$$\tilde{f}(A) = \tilde{f}(A) - f(A) + f(A) = Q\delta A + QA = Q(A + \delta A).$$

Let $\tilde{A} = A + \delta A$. We have

$$\tilde{f}(A) = f(\tilde{A}).$$

We now show that the backward error $\|\tilde{A} - A\|_F$ is relatively small by using the unitary invariance of the Frobenius norm:

$$\begin{aligned} \frac{\|\tilde{A} - A\|_F}{\|A\|_F} &= \frac{\|\delta A\|_F}{\|A\|_F} = \frac{\|Q\delta A\|_F}{\|A\|_F} = \frac{\|\tilde{f}(A) - f(A)\|_F}{\|A\|_F} \\ &\leq \frac{\sqrt{m}C(m)\epsilon_{\text{m}}\|A\|_F}{\|A\|_F} = \sqrt{m}\,c(m)\epsilon_{\text{m}}. \end{aligned}$$

□