Condition number and diagonal systems[1]

Background. In class we mentioned the estimate

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right), \tag{1}$$

where $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ and the "computed solution" $\mathbf{x}_c = \mathbf{x} + \delta\mathbf{x}$ obeys $(A + \delta A)\mathbf{x}_c = \mathbf{b} + \delta\mathbf{b}$. We view $\mathbf{x}$ as the "exact solution" obeying $A\mathbf{x} = \mathbf{b}$. We did not derive (1) in class, only quoted it. However, if $\delta A = O$ (zero matrix), then (1) becomes the estimate we did derive in class,

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(A)\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{2}$$

Here we consider the example of a simple diagonal system. For the example the righthand sides of the inequalities (1) and (2) overestimate the error associated with the computed solution, with the solution process viewed as generating (i) a perturbation δA (but no perturbation $\delta\mathbf{b}$) or (ii) a perturbation $\delta\mathbf{b}$ (but no perturbation δA). For both cases, the solution process is *backward stable*: $\mathbf{x}_c$ is the exact solution to a perturbed problem, with the perturbations $O(\varepsilon_{\mathrm{mach}})$ in a relative sense. In what follows, all (matrix and vector) norms $\|\cdot\|$ are the infinity norm $\|\cdot\|_\infty$.

Example. Suppose that we solve the real system $A\mathbf{x} = \mathbf{b}$, where explicitly

$$\begin{pmatrix} 1 & 0 \\ 0 & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \qquad 0 < a_{22} \ll 1. \tag{3}$$

Clearly, the exact solution $\mathbf{x}$ is

$$x_1 = b_1, \qquad x_2 = b_2/a_{22}. \tag{4}$$

Assume for simplicity that a_{22}, b_1, and b_2 are already floating point numbers, that is, represented exactly as double-precision machine numbers. The *numerical solution* $\mathbf{x}_c$ is then

$$x_{c1} = b_1, \qquad x_{c2} = \mathrm{fl}(b_2/a_{22}) = (b_2/a_{22})(1+\alpha), \tag{5}$$

[1] Supplemental notes for CS-Math 375, University of New Mexico, Fall 2022.

where the α factor takes into account the inexact machine division. In our model of floating point arithmetic $|\alpha| \leq \varepsilon_{\text{mach}}$.

Backward stability. The solution (5) is the exact solution to the system

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & a_{22}/(1+\alpha) \end{pmatrix}}_{A+\begin{pmatrix} 0 & 0 \\ 0 & a_{22}\left[1/(1+\alpha)-1\right] \end{pmatrix}} \begin{pmatrix} x_{c1} \\ x_{c2} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}. \tag{6}$$

The perturbation here is

$$\delta A = \begin{pmatrix} 0 & 0 \\ 0 & a_{22}\left[1/(1+\alpha)-1\right] \end{pmatrix}, \qquad \delta\mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{7}$$

Clearly then, $\|\delta\mathbf{b}\| = 0$ and

$$\|\delta A\| = |a_{22}|\left[1/(1+\alpha)-1\right] = |a_{22}|(-\alpha+\alpha^2-\alpha^3+\cdots) = O(\varepsilon_{\text{mach}}). \tag{8}$$

Since $\|A\| = 1$, we have $\|\delta A\|/\|A\| = O(\varepsilon_{\text{mach}})$. Thus the solution process on the computer is backward stable.

Alternatively, the solution (5) may be viewed as exact solution to

$$\begin{pmatrix} 1 & 0 \\ 0 & a_{22} \end{pmatrix} \begin{pmatrix} x_{c1} \\ x_{c2} \end{pmatrix} = \underbrace{\begin{pmatrix} b_1 \\ b_2(1+\alpha) \end{pmatrix}}_{\mathbf{b}+\begin{pmatrix} 0 \\ \alpha b_2 \end{pmatrix}}. \tag{9}$$

The perturbation now is

$$\delta A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \delta\mathbf{b} = \begin{pmatrix} 0 \\ \alpha b_2 \end{pmatrix}, \quad \|\delta\mathbf{b}\| = \alpha|b_2| \leq \|\mathbf{b}\|\alpha. \tag{10}$$

This shows $\|\delta\mathbf{b}\|/\|\mathbf{b}\| = O(\varepsilon_{\text{mach}})$. With this view, the numerical solution process is again backward stable.

Condition number. By inspection $\|A\| = 1$, and $\|A^{-1}\| = 1/|a_{22}|$. Therefore, $\kappa(A) = 1/|a_{22}| \gg 1$. Then, with the second viewpoint, $\delta A = O$ and $\delta\mathbf{b} \neq 0$, the estimate (2) is

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(A)\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = \frac{1}{|a_{22}|}O(\varepsilon_{\text{mach}}). \tag{11}$$

The righthand side here could be large if $|a_{22}|$ is sufficiently small. This suggests that the relative forward error will then **not** be small. However, as shown now, *this bound is too pessimistic*.

Using the solution (5) and the observation that $x_{c2} = x_2(1+\alpha)$,

$$\delta\mathbf{x} = \mathbf{x}_c - \mathbf{x} = \begin{pmatrix} 0 \\ x_2[(1+\alpha)-1] \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha x_2 \end{pmatrix}. \tag{12}$$

This shows $\|\delta\mathbf{x}\| = \alpha|x_2| \leq \alpha\|\mathbf{x}\|$. We therefore find

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} = O(\varepsilon_{\text{mach}}). \tag{13}$$

Despite the condition number being large, the numerical solution is accurate.