The Condition of a System of Linear Equations: Alternative Derivation

Roland Angst
Computer Vision and Geometry Lab, Eidgenössische Technische Hochschule Zürich
Zürich, Switzerland
rangst@inf.ethz.ch

1. Motivation

Sec. 1.9 of the script [1] already presents a motivation for the condition number

$$\kappa\left(\mathbf{A}\right) = \left\|\mathbf{A}\right\| \left\|\mathbf{A}^{-1}\right\| \tag{1}$$

of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$. As shown in [1], the condition number leads to an upper bound for the error $\left\|\Delta\mathbf{x}\right\|$ in the solution $\mathbf{x} + \Delta\mathbf{x}$ of a system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ with slightly changed initial data $\mathbf{A} + \Delta\mathbf{A}$. Note that the derivation in [1] leads to an upper bound for $\frac{\left\|\Delta\mathbf{x}\right\|}{\left\|\mathbf{x}+\Delta\mathbf{x}\right\|}$, rather than for the relative error

$$\frac{\left\|\Delta\mathbf{x}\right\|}{\left\|\mathbf{x}\right\|}. \tag{2}$$

Moreover, only a change $\Delta\mathbf{A}$ in the matrix is considered. The influence of a change in the right-hand side $\Delta\mathbf{b}$ is not treated.

Hence, these notes present an alternative derivation for the condition of a system of linear equations. The goal is to bound the relative error in Eq. (2) when both the matrix, and the right-hand-side are perturbed. The derivation will closely follow Sec. 1.2.2 in [2].

2. Alternative Derivation for Matrix Condition Number

We start with the perturbed problem

$$\left[\mathbf{A} + \Delta\mathbf{A}\right]\left(\mathbf{x} + \Delta\mathbf{x}\right) = \mathbf{b} + \Delta\mathbf{b} \tag{3}$$

of the original problem

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{4}$$

The derivation will require the following inequalities:

i) Matrix and vector norm must be compatible: $\left\|\mathbf{A}\mathbf{x}\right\| \leq \left\|\mathbf{A}\right\| \left\|\mathbf{x}\right\|$

ii) We assume that $\left\|\mathbf{A}^{-1}\right\| \left\|\Delta\mathbf{A}\right\| < 1$.

Expanding Eq. (3) gives

$$\mathbf{A}\mathbf{x} + \mathbf{A}\Delta\mathbf{x} + \Delta\mathbf{A}\mathbf{x} + \Delta\mathbf{A}\Delta\mathbf{x} = \mathbf{b} + \Delta\mathbf{b}, \tag{5}$$

and using the exact solution, *i.e.* $\mathbf{A}\mathbf{x} = \mathbf{b}$, we get

$$\Delta\mathbf{x} = \mathbf{A}^{-1}\left(\Delta\mathbf{b} - \Delta\mathbf{A}\mathbf{x} - \Delta\mathbf{A}\Delta\mathbf{x}\right). \tag{6}$$

Taking the norm on both sides and using i) yields

$$\begin{aligned}\left\|\Delta\mathbf{x}\right\| &\leq \left\|\mathbf{A}^{-1}\right\| \left\|\Delta\mathbf{b} + \Delta\mathbf{A}\mathbf{x} + \Delta\mathbf{A}\Delta\mathbf{x}\right\| & (7)\\ &\leq \left\|\mathbf{A}^{-1}\right\| \left(\left\|\Delta\mathbf{b}\right\| + \left\|\Delta\mathbf{A}\right\| \left\|\mathbf{x}\right\| + \left\|\Delta\mathbf{A}\right\| \left\|\Delta\mathbf{x}\right\|\right), & (8)\end{aligned}$$

where we have used the triangle inequality for norms $\left\|\mathbf{a}+\mathbf{b}\right\| \leq \left\|\mathbf{a}\right\| + \left\|\mathbf{b}\right\|$ in the last step. This previous inequality is equivalent to

$$\left(1 - \left\|\mathbf{A}^{-1}\right\| \left\|\Delta\mathbf{A}\right\|\right) \left\|\Delta\mathbf{x}\right\| \leq \left\|\mathbf{A}^{-1}\right\| \left(\left\|\Delta\mathbf{b}\right\| + \left\|\Delta\mathbf{A}\right\| \left\|\mathbf{x}\right\|\right). \tag{9}$$

Using ii) results in

$$\frac{\left\|\Delta\mathbf{x}\right\|}{\left\|\mathbf{x}\right\|} \leq \frac{\left\|\mathbf{A}^{-1}\right\|}{1 - \left\|\mathbf{A}^{-1}\right\| \left\|\Delta\mathbf{A}\right\|} \left(\frac{\left\|\Delta\mathbf{b}\right\|}{\left\|\mathbf{x}\right\|} + \left\|\Delta\mathbf{A}\right\|\right), \tag{10}$$

and using $\left\|\mathbf{b}\right\| = \left\|\mathbf{A}\mathbf{x}\right\| \leq \left\|\mathbf{A}\right\| \left\|\mathbf{x}\right\|$ finally gives

$$\frac{\left\|\Delta\mathbf{x}\right\|}{\left\|\mathbf{x}\right\|} \leq \frac{\left\|\mathbf{A}^{-1}\right\| \left\|\mathbf{A}\right\|}{1 - \left\|\mathbf{A}^{-1}\right\| \left\|\Delta\mathbf{A}\right\|} \left(\frac{\left\|\Delta\mathbf{b}\right\|}{\left\|\mathbf{b}\right\|} + \frac{\left\|\Delta\mathbf{A}\right\|}{\left\|\mathbf{A}\right\|}\right). \tag{11}$$

Introducing the condition number as defined in Eq. (1) leads to the following upper bound for the relative error

$$\frac{\left\|\Delta\mathbf{x}\right\|}{\left\|\mathbf{x}\right\|} \leq \frac{\kappa\left(\mathbf{A}\right)}{1 - \kappa\left(\mathbf{A}\right)\frac{\left\|\Delta\mathbf{A}\right\|}{\left\|\mathbf{A}\right\|}} \left(\frac{\left\|\Delta\mathbf{b}\right\|}{\left\|\mathbf{b}\right\|} + \frac{\left\|\Delta\mathbf{A}\right\|}{\left\|\mathbf{A}\right\|}\right). \tag{12}$$

If we make the stronger assumption in ii) that $\left\|\mathbf{A}^{-1}\right\| \left\|\Delta\mathbf{A}\right\| \ll 1$, *i.e.* $\left\|\mathbf{A}^{-1}\right\| \left\|\Delta\mathbf{A}\right\|$ is clearly smaller than one, we get the simplified formula[1]

$$\frac{\left\|\Delta\mathbf{x}\right\|}{\left\|\mathbf{x}\right\|} \quad \tilde{\leq} \quad \kappa\left(\mathbf{A}\right)\left(\frac{\left\|\Delta\mathbf{b}\right\|}{\left\|\mathbf{b}\right\|} + \frac{\left\|\Delta\mathbf{A}\right\|}{\left\|\mathbf{A}\right\|}\right). \tag{13}$$

[1] The symbol $\tilde{\leq}$ denotes that the inequality does not follow from Eq. (12) in a strict mathematical sense. However, together with $\left\|\mathbf{A}^{-1}\right\| \left\|\Delta\mathbf{A}\right\| \ll 1$, the inequality in Eq. (13) holds in practice almost always nevertheless.

3. Remarks and Interpretation

Based on the upper bound for the relative error in Eq. (12) or in Eq. (13), we observe the following things:

- If the entries of the unknown vector $\mathbf{x}$ are not of the same order of magnitude, then small entries can be completely wrong even if both the condition number $\kappa(\mathbf{A})$ and the perturbations $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$ are small. The reason for that can be seen in Eq. (12) or Eq. (13): the upper bound is based on the *norm* of $\Delta\mathbf{x}$, $\Delta\mathbf{A}$, and $\Delta\mathbf{b}$. Hence, large relative errors of small entries might be negligible in $\|\Delta\mathbf{x}\|$ compared to small relative errors of large entries, *i.e*. large relative errors of small entries do not contribute significantly to the norm $\|\Delta\mathbf{x}\|$. For example, consider an exact solution $\mathbf{x} = (1e6, 1e-6)^T$. A perturbed solution with $\Delta\mathbf{x} = (1e-3, 1e-6)$ can then still yield a much smaller relative error $\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}$ than the upper bound $\kappa(\mathbf{A})\left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right)$. However, an absolute error of $1e-6$ for the second entry $x_2 = 1e-6$ is really large and renders the small entry of the numerical solution $\tilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$ completely useless.
- $\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}$ and $\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$ are the relative error of the original input. These errors are usually at least on the order of the machine precision ϵ_{mach} because we need to represent the values of the entries of the input with machine numbers which leads to relative rounding errors on the order of ϵ_{mach}.
- Based on the previous remark, even if all the entries of the unknown vector are on the same order of magnitude, then we have to expect that up to the last $\log\kappa(\mathbf{A})$ digits in the numerical solution $\tilde{\mathbf{x}}$ may deviate from the exact solution $\mathbf{x}$.

4. Example Application

At the end of Sec 1.6.2 in [1], we concluded that ”we better not trust small residuals to always imply that we are close to a solution”. The immediate question is: In which cases can we trust small residuals? Equipped with the previous derivation and the upper bound in Eq. (13), we can now answer this question.

Specifically, consider the system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ with exact solution $\mathbf{x}$ and a solution $\mathbf{x}_k$ computed with some algorithm (*e.g*. an iterative algorithm providing successively refined solutions $\mathbf{x}_k, \mathbf{x}_{k+1}, \ldots$). The latter solution will lead to a small residual error, *i.e*. in the notation used in Sec 1.6.2 of [1]

$$\mathbf{A}\mathbf{x}_k = \mathbf{b} - \mathbf{r}. \tag{14}$$

In order to apply the bounds as derived in previous sections, we observe that $\mathbf{x}_k = \mathbf{x} + \Delta\mathbf{x}$ and $-\mathbf{r} = \Delta\mathbf{b}$ ($\Delta\mathbf{A}$ is equal to the zero-matrix in this example). Hence, we get the upper bound

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A})\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{15}$$

Therefore, a small residual $\|\mathbf{r}\|$ guarantees a ”good” solution if

- $\|\mathbf{r}\|$ is small compared to the right-hand-side $\|\mathbf{b}\|$, and
- the condition number $\kappa(\mathbf{A})$ is sufficiently small.

Note that in the example considered in Sec. 1.6.2, the matrix

$$\mathbf{A} = \begin{bmatrix} 0.4343 & 0.4340 \\ 0.4340 & 0.4337 \end{bmatrix} \tag{16}$$

has a condition number of $\kappa_2(\mathbf{A}) \approx 8 \cdot 10^6$. $\kappa_2(\mathbf{A})$ denotes the condition number with the spectral norm $\|\mathbf{A}\|_2 = \sqrt{\lambda_{max}(\mathbf{A}^T\mathbf{A})}$ where $\lambda_{max}(\mathbf{B})$ denotes the maximal Eigenvalue of the symmetric (or Hermitian) matrix $\mathbf{B}$.

References

[1] W. Gander. *Lineare Algebra: Endliche Arithmetik*. 2005.

[2] H. Schwarz. *Numerische Mathematik*. Lehrbuch Mathematik. Teubner Stuttgart, 4th edition, 1997.