When solving a linear system $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$, computer algorithms are only providing an approximation ($\boldsymbol{x}_{\text{app}}$) to the exact solution ($\boldsymbol{x}_{\text{ex}}$). This is due to factors such as finite precision, round off errors or even imperfect solution algorithms. In either case we have an *error* (error vector, in fact) defined as

$$\boldsymbol{e} = \boldsymbol{x}_{\text{app}} - \boldsymbol{x}_{\text{ex}}$$

Naturally, we would like to have an understanding of the *magnitude* of this error (e.g. some appropriate norm $||\boldsymbol{e}||$). The problem is that we do not know the exact, pristine solution, $\boldsymbol{x}_{\text{ex}}$!

One remedy is offered via the *residual vector* defined as:

$$\boldsymbol{r} = \boldsymbol{b} - \mathbf{A}\boldsymbol{x}_{\text{app}}$$

The vector $\boldsymbol{r}$ is something we can compute practically since it involves only known quantities ($\boldsymbol{b}, \mathbf{A}, \boldsymbol{x}_{\text{app}}$). Furthermore, we have:

$$\begin{aligned}
\boldsymbol{r} &= \boldsymbol{b} - \mathbf{A}\boldsymbol{x}_{\text{app}} \\
&= \mathbf{A}\boldsymbol{x}_{\text{ex}} - \mathbf{A}\boldsymbol{x}_{\text{app}} \\
&= -\mathbf{A}(\boldsymbol{x}_{\text{app}} - \boldsymbol{x}_{\text{ex}}) \\
&= -\mathbf{A}\boldsymbol{e} \Rightarrow \\
\boldsymbol{r} &= -\mathbf{A}\boldsymbol{e} \\
\boldsymbol{e} &= -\mathbf{A}^{-1}\boldsymbol{r}
\end{aligned}$$

The last equation links the error with the residual. Furthermore, we can write

$$||\boldsymbol{e}|| = ||\mathbf{A}^{-1}r|| \leq ||\mathbf{A}^{-1}|| \cdot ||r||$$

This equation provides a *bound* for the error, as a function of $||\mathbf{A}^{-1}||$ and the norm of the computable vector $\boldsymbol{r}$! Note that:

- We can obtain this estimate *without* knowing the exact solution, *but*
- We need $||\mathbf{A}^{-1}||$ and generally, computing $\mathbf{A}^{-1}$ is just as difficult (if not more) than finding $\boldsymbol{x}_{\text{ex}}$. *However* there are special cases where an estimate of $||\mathbf{A}^{-1}||$ can be obtained.

A different source of error Sometimes, the right-hand-side of $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ has errors that make it deviate from its intended value. For example, in the Vandermonde matrix method for polynomial interpolation, $\boldsymbol{b}$ contains the samples $(y_1 = f(x_1), y_2, \cdots, y_n)$ where $y_i = f(x_i)$. An error in a measuring device supposed to sample $f(x)$ could lead to erroneous readings y_i^* instead of y_i. In general, measuring inaccuracies can lead to the right-hand-side vector $\boldsymbol{b}$ being misrepresented as $\boldsymbol{b}^*$ $(\neq b)$.

In this case, instead of the intended solution $\boldsymbol{x} = \mathbf{A}^{-1}\boldsymbol{b}$ we in fact compute $\boldsymbol{x}^* = \mathbf{A}^{-1}\boldsymbol{b}^*$. How important is the error $\boldsymbol{e} = \boldsymbol{x}^* - \boldsymbol{x}$ that is caused by the misrepresentation of $\boldsymbol{b}$?

Let us introduce some notation:

$$\begin{aligned} \text{Let } \delta\boldsymbol{b} &:= \boldsymbol{b}^* - \boldsymbol{b} \\ \delta\boldsymbol{x} &:= \boldsymbol{x}^* - \boldsymbol{x} \\ \mathbf{A}\boldsymbol{x} &= \boldsymbol{b} \\ \mathbf{A}\boldsymbol{x}^* &= \boldsymbol{b}^* \end{aligned}$$

$$\begin{aligned} \mathbf{A}(\boldsymbol{x}^* - \boldsymbol{x}) &= \boldsymbol{b}^* - \boldsymbol{b} \\ \mathbf{A}\delta\boldsymbol{x} &= \boldsymbol{b}^* \\ \delta\boldsymbol{x} &= \mathbf{A}^{-1}\delta\boldsymbol{b} \end{aligned}$$

Taking norms:

$$||\delta\boldsymbol{x}|| = ||\mathbf{A}^{-1}\delta\boldsymbol{b}|| \leq ||\mathbf{A}^{-1}||||\delta\boldsymbol{b}|| \tag{12}$$

Thus the error in the computed solution $\delta\boldsymbol{x}$ is proportional to the error in $\boldsymbol{b}$.

An even more relevant question is: How does the *relative* error $\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}||} = \frac{||\boldsymbol{x}^* - \boldsymbol{x}||}{||\boldsymbol{x}||}$ compare to the relative error in $\frac{||\delta\boldsymbol{b}||}{||\boldsymbol{b}||}$? This may be more useful to know since $||\delta\boldsymbol{b}||$ may be impossible to compute (if we don't know the real $\boldsymbol{b}$!).

For this, we write:

$$\begin{aligned} \mathbf{A}\boldsymbol{x} &= \boldsymbol{b} \Rightarrow ||\boldsymbol{b}|| = ||\mathbf{A}\boldsymbol{x}|| \leq ||\mathbf{A}|| \cdot ||x|| \\ & \frac{1}{||x||} \leq ||\mathbf{A}|| \cdot \frac{1}{||\boldsymbol{b}||} \end{aligned} \tag{13}$$

Multiplying 12 and 13, we get:

$$\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}||} \leq \underbrace{||\mathbf{A}|| \cdot ||\mathbf{A}^{-1}||}_{\kappa(\mathbf{A})} \cdot \frac{||\delta\boldsymbol{b}||}{||\boldsymbol{b}||}$$

Thus, the relative error in $\boldsymbol{x}$ is bounded by a multiple of the relative error in $\boldsymbol{b}$! The multiplicative constant $\kappa(\mathbf{A}) = ||\mathbf{A}|| \cdot ||\mathbf{A}^{-1}||$ is called the *condition number* of $\mathbf{A}$ and is an important measure of the sensitivity of a linear system $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ to being solved on a computer in the presence of inaccurate values. e.g. If the relative error $\frac{||\delta\boldsymbol{b}||}{||\boldsymbol{b}||}$ is 0.0001%, but $\kappa(\mathbf{A}) = 10^5$ (which could certainly happen), then we could have up to a 10% error in the computed $\boldsymbol{x}$!

Why is this always relevant?

Simply, almost *any* $\boldsymbol{b}$ will have *some* small relative error due to the fact it is represented on a computer up to machine precision. The relative error will be at least as much as the machine epsilon due to round off.

$$\frac{||\delta \boldsymbol{b}||_\infty}{||\boldsymbol{b}||} \geq \varepsilon \approx 10^{-7} \text{ (for single-precision floating point numbers)}$$

But, how bad can the condition number get? *Very* bad, at times. For example:

Hilbert matrices $\mathbf{H}_n \in \mathbb{R}^{n \times n}$ are defined as $(H_n)_{ij} = \frac{1}{i+j-1}$

$$\mathbf{H}_5 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{bmatrix}$$

$$\kappa_\infty(\mathbf{H}_5) = ||\mathbf{H}_5||_\infty \cdot ||\mathbf{H}_5^{-1}||_\infty \approx 10^6!$$

Thus, any attempt at solving $\mathbf{H}_5 \boldsymbol{x} = \boldsymbol{b}$ would be subject to a relative error up to 10% just due to round off errors in $\boldsymbol{b}$!

Another case: near-singular matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 6+\varepsilon \end{bmatrix}$$

As $\varepsilon \to 0$, $\mathbf{A}$ becomes singular (non-invertible). In this case, $\kappa(\mathbf{A}) \to \infty$.

What is the best case for $\kappa(\mathbf{A})$?

Lemma: For any vector-induced matrix norm, we have $||\mathbf{I}|| = 1$.

Proof: From definition:

$$||\mathbf{I}|| = \max_{x \neq 0} \frac{||\mathbf{I}\boldsymbol{x}||}{||\boldsymbol{x}||} = \max_{x \neq 0} \frac{||\boldsymbol{x}||}{||\boldsymbol{x}||} = 1$$

Using property (iv) of matrix norms, we get:

$$\mathbf{I} = \mathbf{A} \cdot \mathbf{A}^{-1} \Rightarrow 1 = ||\mathbf{I}|| = ||\mathbf{A} \cdot \mathbf{A}^{-1}|| \leq ||\mathbf{A}|| \cdot ||\mathbf{A}^{-1}||$$

Thus, $\kappa(\mathbf{A}) \geq 1$ for any matrix. The "best" conditioned matrices that achieve this bound are of the form $\mathbf{A} = c \cdot \mathbf{I}$, for which $\kappa(\mathbf{A}) = 1$.