Stopping criteria

Assume $b \neq 0$ and choose a relative tolerance $0 < \varepsilon < 1$.

- Test on the iterates:

$$\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon \max(\|x^{(k)}\|, \|x^{(k-1)}\|).$$

- Test on the residual $r^{(k)} = b - Ax^{(k)}$:

$$\|r^{(k)}\| \leq \varepsilon \|b\| \quad \Longleftrightarrow \quad \frac{\|r^{(k)}\|}{\|b\|} \leq \varepsilon.$$

Both tests are relative to the indicated scales. The iterate test is written without division, so it also applies when both iterates vanish. Check the initial residual before iterating. At later steps, accept when both tests hold; an exactly zero residual can be accepted immediately. Reaching `maxiter` without acceptance is a failure to meet the tolerance.

Pseudocode for splitting methods

$$
\begin{aligned}
M\underline{x}^{(k)} &= \underline{b} + N\underline{x}^{(k-1)} = \underline{b} + (M - A)\underline{x}^{(k-1)} \\
&= \underline{b} - A\underline{x}^{(k-1)} + Mx^{(k-1)} \\
\Longrightarrow \underline{x}^{(k)} &= \underline{x}^{(k-1)} + M^{-1}\underline{r}^{(k-1)}
\end{aligned}
$$

M is usually referred as a *preconditioner*.

Splitting iterative method

Input: $A \in \mathbb{R}^{n\times n}$, $0 \neq \underline{b} \in \mathbb{R}^n$, $\underline{x}^{(0)} \in \mathbb{R}^n$, relative tolerance ε, maxiter $\in \mathbb{N}$
Choose nonsingular $M \in \mathbb{R}^{n\times n}$ and set $\underline{r}^{(0)} = \underline{b} - A\underline{x}^{(0)}$
If the initial residual test holds, return $x^{(0)}$.
for $k = 1, 2, \ldots,$ maxiter:
    Solve $M\underline{p}^{(k-1)} = \underline{r}^{(k-1)}$
    $\underline{x}^{(k)} = \underline{x}^{(k-1)} + \underline{p}^{(k-1)}$
    $\underline{r}^{(k)} = \underline{b} - A\underline{x}^{(k)}$
    If $r^{(k)} = 0$ or both tests hold, return $x^{(k)}$
end
Output: $x^{(k)}$ and report that the tolerance was not met.

Error analysis

Unfortunately, the fact that the residual is small does not guarantee that the error $\underline{x} - \underline{x}^{(k)}$ is small.

$$\underline{r}^{(k)} := \underline{b} - A\underline{x}^{(k)} = A\underline{x} - A\underline{x}^{(k)} \longrightarrow \underline{x} - \underline{x}^{(k)} = A^{-1}\underline{r}^{(k)}.$$

Assume A nonsingular and $b \neq 0$. Taking norms gives

$$\begin{aligned}\|\underline{x} - \underline{x}^{(k)}\| &= \|A^{-1}\underline{r}^{(k)}\| \leq |||A^{-1}|||\,\|\underline{r}^{(k)}\| \\ &\leq |||A^{-1}|||\,\frac{\|\underline{r}^{(k)}\|}{\|\underline{b}\|}\,\|A\underline{x}\| \leq |||A^{-1}|||\,|||A|||\,\|\underline{x}\|\,\frac{\|\underline{r}^{(k)}\|}{\|\underline{b}\|}.\end{aligned}$$

Then we obtain

$$\frac{\|\underline{x} - \underline{x}^{(k)}\|}{\|\underline{x}\|} \leq |||A^{-1}|||\,|||A|||\,\frac{\|\underline{r}^{(k)}\|}{\|\underline{b}\|}.$$

The **condition number** is $\kappa(A) = |||A^{-1}|||\,|||A|||$. If $\kappa(A) \gg 1$, the matrix is ill-conditioned: a small relative residual may accompany a much larger relative error. A relative residual at most $\varepsilon/\kappa(A)$ guarantees relative error at most ε.

Concept of conditioning

When dealing with ill-conditioned matrices, any numerical method (direct or iterative) might produce unsatisfactory results.

Roughly speaking, a problem is *well-conditioned* if "small" perturbations on the data determine "small" perturbations on the results.

To clarify the concept of *conditioning* of a problem, let us consider a generic problem: find u solution of

$$F(u, d) = 0,$$

where d are the data, and F is the law relating u to d.

Well-posedness and local conditioning

Write $u = S(d)$ for the solution of $F(u, d) = 0$. The problem is **well-posed** if a solution exists, is unique, and depends continuously on the admissible data. For $d \neq 0$ and $S(d) \neq 0$, the local relative condition number is

$$\kappa_{\mathrm{rel}}(d) = \limsup_{\delta d \to 0,\ \delta d \neq 0} \frac{\|S(d + \delta d) - S(d)\| / \|S(d)\|}{\|\delta d\| / \|d\|},$$

where the perturbations are admissible.

Example

Back to linear systems, let us make a conditioning analysis in a simple case. Assume that the possible errors are only on the right-hand side (and not on the matrix). Let $\delta\underline{b}$ be the errors on $\underline{b}$, small (in percentage), that is, for ε small,

$$\frac{\|\delta\underline{b}\|}{\|\underline{b}\|} \leq \varepsilon$$

The solution of the system with matrix A and right-hand side $\widetilde{\underline{b}} = \underline{b} + \delta\underline{b}$ will be $\widetilde{\underline{x}} = \underline{x} + \delta\underline{x}$:

$$A\underline{x} = \underline{b} \;\rightarrow\; A(\underline{x} + \delta\underline{x}) = \underline{b} + \delta\underline{b},$$

subtracting

$$A\delta\underline{x} = \delta\underline{b}, \qquad \text{giving} \qquad \delta\underline{x} = A^{-1}\delta\underline{b}.$$

Proceeding as we did before we have

$$\|\delta\underline{x}\| = \|A^{-1}\delta\underline{b}\| \leq |||A^{-1}|||\|\delta\underline{b}\| = |||A^{-1}|||\frac{\|\delta\underline{b}\|}{\|\underline{b}\|}\|\underline{b}\|$$

$$= |||A^{-1}|||\frac{\|\delta\underline{b}\|}{\|\underline{b}\|}\|A\underline{x}\| \leq |||A^{-1}|||\frac{\|\delta\underline{b}\|}{\|\underline{b}\|}|||A|||\|\underline{x}\|$$

We found

$$\|\delta \underline{x}\| \leq |||A^{-1}||| \frac{\|\delta \underline{b}\|}{\|\underline{b}\|} |||A||| \|\underline{x}\|$$

$$\Longrightarrow \frac{\|\delta \underline{x}\|}{\|\underline{x}\|} \leq |||A^{-1}||| \, |||A|||\varepsilon = \kappa(A)\varepsilon$$

A simple example to understand how a big condition number might affect the results.

$$\begin{pmatrix} 10^6 & 10^{-12} \\ 0 & 10^{-6} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10^6 \\ 10^{-6} \end{pmatrix}$$

Exact solution $x_2 = 1, x_1 = 1 - 10^{-18} \simeq 1; \kappa_\infty(A) \simeq 10^{12}$. We compare two separate perturbations: $\delta b = (10^{-6}, 0)^T$ and $\delta b = (0, 10^{-6})^T$. Both have relative size $\|\delta b\|_\infty / \|b\|_\infty = 10^{-12}$, but their effects on the solution are very different.

Same perturbation size, very different effects (1)

For the original system, backward substitution gives

$$10^{-6}x_2 = 10^{-6} \implies x_2 = 1, \qquad 10^{6}x_1 + 10^{-12} = 10^{6} \implies x_1 = 1 - 10^{-18}.$$

Thus $\|b\|_\infty = 10^6$ and $\|x\|_\infty = 1$.

Perturb only the first component: $\delta b = (10^{-6}, 0)^T$. Let $A\widehat{x} = b + \delta b$. Then

$$\begin{aligned} 10^{-6}\widehat{x}_2 = 10^{-6} \qquad &\implies \quad \widehat{x}_2 = 1, \\ 10^{6}\widehat{x}_1 + 10^{-12} = 10^{6} + 10^{-6} \qquad &\implies \quad \widehat{x}_1 = 1 - 10^{-18} + 10^{-12}. \end{aligned}$$

Consequently,

$$\delta x := \widehat{x} - x = \begin{pmatrix} 10^{-12} \\ 0 \end{pmatrix}, \qquad \frac{\|\delta x\|_\infty}{\|x\|_\infty} = 10^{-12} = \frac{\|\delta b\|_\infty}{\|b\|_\infty}.$$

There is no amplification of the relative perturbation in this direction.

Same perturbation size, very different effects (2)

Now perturb only the second component: $\delta b = (0, 10^{-6})^T$. This is a separate experiment with the original b.

$$\begin{pmatrix} 10^6 & 10^{-12} \\ 0 & 10^{-6} \end{pmatrix} \begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \end{pmatrix} = \begin{pmatrix} 10^6 \\ 2 \cdot 10^{-6} \end{pmatrix}.$$

Backward substitution gives

$$\begin{aligned} 10^{-6}\widehat{x}_2 = 2 \cdot 10^{-6} \qquad &\Longrightarrow \quad \widehat{x}_2 = 2, \\ 10^6\widehat{x}_1 + 2 \cdot 10^{-12} = 10^6 \qquad &\Longrightarrow \quad \widehat{x}_1 = 1 - 2 \cdot 10^{-18}. \end{aligned}$$

Therefore

$$\delta x = \begin{pmatrix} -10^{-18} \\ 1 \end{pmatrix}, \qquad \boxed{\frac{\|\delta x\|_\infty}{\|x\|_\infty} = 1}, \qquad \frac{\|\delta b\|_\infty}{\|b\|_\infty} = 10^{-12}.$$

The relative change in the solution is 100%, with an amplification factor of 10^{12}.

Why the large change is consistent with the error bound

Here the inverse and the condition number can be computed explicitly:

$$A^{-1} = \begin{pmatrix} 10^{-6} & -10^{-12} \\ 0 & 10^{6} \end{pmatrix}, \qquad |||A|||_\infty = 10^6 + 10^{-12}, \qquad |||A^{-1}|||_\infty = 10^6.$$

Hence $\kappa_\infty(A) = 10^{12} + 10^{-6}$, and the bound gives

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \le \kappa_\infty(A) \frac{\|\delta b\|_\infty}{\|b\|_\infty} = (10^{12} + 10^{-6})10^{-12} = 1 + 10^{-18}.$$

The second perturbation nearly attains this bound; the first does not. A large condition number describes possible amplification, not amplification in every direction.

Which relative change? The value 10^{-12} measures the data perturbation relative to $\|b\|_\infty$. In the second experiment, $|\delta b_2|/|b_2| = 1$: the small component of b changes by 100%. Thus a small normwise perturbation need not be small in every component.

All calculations above use exact arithmetic: this is sensitivity of the linear system itself, not an error introduced by the solver.