

Stopping criteria

Choose nonnegative absolute tolerances a_x, a_r and relative tolerances r_x, r_r .

- Test on the iterates:

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

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

$$\|r^{(k)}\| \leq a_r + r_r \|b\|.$$

These tests avoid division by zero. Absolute tolerances must be chosen for the scale and units of the problem. 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 tolerances.

Pseudocode for splitting methods

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

M is usually referred as a *preconditioner*.

Splitting iterative method

Input: $A \in \mathbb{R}^{n \times n}$, $\underline{b} \in \mathbb{R}^n$, $\underline{x}^{(0)} \in \mathbb{R}^n$, tolerances as above, $\text{maxiter} \in \mathbb{N}$

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 $\underline{x}^{(0)}$.

for $k = 1, 2, \dots, \text{maxiter}$:

 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 $\underline{r}^{(k)} = 0$ or both tests hold, return $\underline{x}^{(k)}$

end

Output: $\underline{x}^{(k)}$ and report that the tolerances were 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 $\underline{b} \neq 0$. 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

$$(P) \qquad F(u, d) = 0,$$

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

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

where the perturbations are admissible. It may be infinite: continuity alone does not imply a finite local Lipschitz bound. If S is differentiable on an open data domain, $\kappa_{\text{rel}}(d) = \|S'(d)\| \|d\| / \|S(d)\|$ for the associated operator norm. At zero data or zero solution, use absolute conditioning instead.

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 $\tilde{\underline{b}} = \underline{b} + \delta \underline{b}$ will be $\tilde{\underline{x}} = \underline{x} + \delta \underline{x}$:

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

subtracting

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

Proceeding as we did before we have

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

We found

$$\begin{aligned}\|\delta \underline{x}\| &\leq \|A^{-1}\| \frac{\|\delta \underline{b}\|}{\|\underline{b}\|} \|A\| \|\underline{x}\| \\ \Rightarrow \frac{\|\delta \underline{x}\|}{\|\underline{x}\|} &\leq \|A^{-1}\| \|A\| \varepsilon = \kappa(A) \varepsilon\end{aligned}$$

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 $x_2 = 1, x_1 \simeq 1$; $\kappa_\infty(A) \simeq 10^{12}$. Now perturb only the first component of the right-hand side by 10^{-6} , and then only the second component by 10^{-6} . In both cases $\|\delta \underline{b}\|_\infty / \|\underline{b}\|_\infty \leq 10^{-12}$. What happens to the solution?

Conditioning versus algorithmic stability

Conditioning concerns the problem. **Backward stability** concerns the algorithm: its computed answer solves a nearby problem. Consider

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad 0 < \varepsilon \ll 1.$$

A perturbation $\delta b = (0, \delta)^T$ gives $\delta x = (0, \delta/\varepsilon)^T$ and $\kappa_\infty(A) = 1/\varepsilon$. Even a backward-stable solver may have a large forward error. Conversely,

$q(x) = \sqrt{1+x} - 1$ is well-conditioned for small $x > 0$ (relative condition number tends to 1), but direct subtraction loses significant digits. The equivalent formula

$$q(x) = \frac{x}{\sqrt{1+x} + 1}$$

avoids cancellation. Reformulating an algorithm can improve stability without changing the problem's conditioning.