Principles of Numerical Mathematics

In an abstract setting, we describe a generic problem as

$$F(x, d) = 0 \tag{1}$$

where x is the unknown (generally a real number, a vector or a function) and $d \in D$ is the data. For each of the elements above we use an appropriate norm, which will enable us to measure quantities of interests from a numerical point of view, such as errors, stability, and dependency of the solution from the data. In particular we will use the symbols $\|\cdot\|_F$, $\|\cdot\|_x$ and $\|\cdot\|_d$ to indicate the various norms.

Stability of a problem

In general, not all problems can be approximated. If we write a problem as in (1), then its approximation is useful only if the continuous problem has a unique solution which depends continuously on the data. We call these problems *well posed* or *stable*:

Definition: A mathematical problem (1) is **well posed** or **stable** if it

- *admits a unique solution x:*

  $$\forall d \in D, \exists!\, x \text{ s.t. } F(x, d) = 0.$$

- *which depends with continuity on the data:*

  it means that small perturbations on the data d of D yield "small" changes in the solution x. Precisely, let $d \in D$ and denote by δd a perturbation admissible of the data, in the sense that $d + \delta d \in D$, and by δx the corresponding change in the solution x, so that $x + \delta x$ is the perturbed solution, in such a way that

  $$F(x + \delta x, d + \delta d) = 0, \tag{2}$$

  then we require that

  $$\begin{aligned} &\forall d \in D, \quad \exists \eta_0 = \eta_0(d), K_0 \text{ such that} \\ &\text{if } \|\delta d\|_d < \eta_0 \quad \Longrightarrow \quad \|\delta x\|_x < K_0 \|\delta d\|_d. \end{aligned} \tag{3}$$

Example: A simple instance of an ill-posed problem is finding the number of real roots of a polynomial. For example, the polynomial $p(x) = x^4 - x^2(2a-1) + a(a-1)$ exhibits a discontinuous variation of the number of real roots as a continuously varies in the real field. We have, indeed, 4 real roots if $a \geq 1$, 2 if $a \in [0, 1)$ while no real roots exist if $a < 0$.

Condition numbers

Definition: A measure of how accurately we can approximate the problem at hand is given by the **condition number**: it measures the problem sensitiveness w.r.t. the input data.

Relative condition number:

$$K := \sup_{d, \delta d} \left\{ \frac{\|\delta x\|_x / \|x\|_x}{\|\delta d\|_d / \|d\|_d},\ \delta d \neq 0,\ d + \delta d \in D \right\}. \tag{4}$$

Absolute condition number (to be used when either $x = 0$ or $d = 0$):

$$K_{abs} := \sup_{\delta d} \left\{ \frac{\|\delta x\|_x}{\|\delta d\|_d},\ \delta d \neq 0,\ d + \delta d \in D \right\}. \tag{5}$$

If problem (1) admits a unique solution x to each data d, then we can construct a **resolvent map** G between the sets of the data and of the
solutions, such that

$$G(d) = x, \quad \text{that is} \quad F(G(d), d) = 0. \tag{6}$$

According to this definition, (2) yields $x + \delta x = G(d + \delta d)$. Assuming that G is differentiable in d and denoting formally by $G'(d)$ its derivative with respect to d, then the Taylor expansion of G around d truncated at first order ensures that

$$G(d + \delta d) - G(d) = G'(d)\delta d + o(\delta d) \qquad \text{for } \delta d \to 0,$$

where $|| \cdot ||$ is a suitable vector norm and $o(\cdot)$ is the classical infinitesimal symbol denoting an infinitesimal term of higher order with respect to its argument.

Neglecting the infinitesimal of higher order with respect to δd, since we have that
$x + \delta x = G(d + \delta d)$ and $x = G(d)$, we have that

$$x + \delta x - x = G(d + \delta d) - G(d) \quad \Longrightarrow \quad \delta x = G(d) - G(d + \delta d) = G'(d)\delta d$$

Using this and from (4) and (5), we respectively deduce that

$$K \simeq ||G'(d)|| \frac{||d||_d}{||G(d)||_x} \qquad K_{abs} \simeq ||G'(d)||.$$

A *stable* problem is **well conditioned** when its condition number is "small", where the meaning of "small" depends on the problem at hand. If $K >> 1$, the problem is **ill-posed (sensitive, unstable)** and thus *not* approximable through numerical methods.

Stability of numerical methods

Once we have a *stable* problem (1), a numerical method for the approximate solution of (1) will consist, in general, of a sequence of approximate problems

$$F_n(x_n, d_n) = 0, \qquad n \geq 1 \tag{7}$$

depending on a certain parameter n (to be defined case by case), such that

$$\lim_{n\to\infty} ||F_n - F||_F = 0$$
$$\lim_{n\to\infty} ||x_n - x||_x = 0$$
$$\lim_{n\to\infty} ||d_n - d||_d = 0,$$

for some appropriate norms $|| \cdot ||_F, || \cdot ||_x$ and $|| \cdot ||_d$, so the understood expectation is that $x_n \to x$ as $n \to \infty$, i.e. that the numerical solution converges to the exact solution.

Equivalently to what happens in the continuous case, we can establish the stability of the approximate n-th problem.

Definition: A mathematical approximation of a stable problem is itself a **stable numerical method** if the following properties are satisfied:

- *Existence and uniqueness of solutions:*

  $$\forall n \geq 1, \forall d_n \in D_n, \; \exists!\, x_n \text{ such that } F_n(x_n, d_n) = 0.$$

- *Continuous dependence on data:*

  Let δd_n be a perturbation of the data, such that $d_n + \delta d_n \in D_n$, and let $x_n + \delta x_n$ be the corresponding perturbed solution, i.e., $F_n(x_n + \delta x_n, d_n + \delta d_n) = 0$, then

$$\begin{array}{ll}\forall d_n \in D_n, & \exists \eta_0 = \eta_0(d_n), K_0 = K_0(d_n) \text{ such that} \\ \text{if } \|\delta d_n\|_d < \eta_0 \in D_n & \implies \quad \|\delta x_n\|_x < K_0 \|\delta d_n\|_d.\end{array} \tag{8}$$

Consistency

Whenever the data d is admissible for F_n, then further properties of the approximations can be devised. In particular, we expect that $x_n \to x$ as $n \to \infty$, i.e. that the numerical solution converges to the exact solution. We could also say that the residual (the error produced by plugging the exact solution in the scheme) tends to zero as $n \to \infty$.

Definition: If the datum d of problem (1) is admissible for F_n, so $d \in D_n$ $\forall n$, we say that the numerical problem (7) is **consistent** if

$$\lim_{n\to\infty} F_n(x,d) = \lim_{n\to\infty} F_n(x,d) - F(x,d) = 0. \tag{9}$$

where x is the solution to problem (1) corresponding to the datum d.
Moreover, a numerical approximation is said to be **strongly consistent** if

$$F_n(x,d) = 0 \quad \text{for any value of } n.$$

Convergence

Definition: A numerical method (7) is **convergent** iff

$$\begin{array}{lll}\forall \varepsilon > 0, \quad \exists n_0 = n_0(\varepsilon), & \exists \delta = \delta(n_0, \varepsilon) & \text{such that} \\ \forall n > n_0, \quad \forall \delta d_n : & \|\delta d_n\|_d \le \delta \quad \Longrightarrow & \|x(d) - x_n(d + \delta d_n)\| \le \varepsilon,\end{array} \tag{10}$$

where d is an admissible datum for the problem (1), $x(d)$ is the corresponding solution and $x_n(d + \delta d_n)$ is the solution of the numerical problem (7) with datum $d + \delta d_n$.

To verify the implication (10) it suffices to check that under the same assumptions

$$||x(d + \delta d_n) - x_n(d + \delta d_n)|| \le \frac{\varepsilon}{2}. \tag{11}$$

Measures of the convergence of x_n to x are given by the **absolute error** or the **relative error**, respectively defined as

$$E(x_n) = |x - x_n|, \qquad E_{\text{rel}}(x_n) = \frac{|x - x_n|}{|x|} \quad (\text{if } x \neq 0).$$

The concepts of stability and convergence are strongly connected. First of all, if problem (1) is well posed, **a numerical problem (7) which is convergent is also stable**. So if the problem is well posed, a *necessary* condition in order for the numerical problem to be convergent is that it is stable.

Proof. Let us thus assume that the method is convergent, that is, (10) holds for an arbitrary $\varepsilon > 0$. We have

$$\begin{aligned} ||\delta x_n|| &= ||x_n(d + \delta d_n) - x_n(d)|| \\ &\le ||x_n(d) - x(d)|| + ||x(d) - x(d + \delta d_n)|| + ||x(d + \delta d_n) - x_n(d + \delta d_n)|| \\ &\le K(\delta(n_0, \varepsilon), d)||\delta d_n|| + \varepsilon, \end{aligned}$$

having used (3) and (11) twice. Choosing now δd_n such that $||\delta d_n|| \le \eta_0$, we deduce that $||\delta x_n||/||\delta d_n||$ can be bounded by $K_0 = K(\delta(n_0, \varepsilon), d) + 1$, provided that $\varepsilon \le ||\delta d_n||$, so that the method is stable.

□

Thus, we are interested in stable numerical methods since only these can be convergent (if a method is not stable it cannot be convergent, since stability $\Rightarrow$ convergence).

Lax-Richtmyer theorem

The stability of a numerical method becomes a *sufficient* condition for the numerical problem (7) to converge if this latter is also consistent with problem (1).

One of the fundamental theorems (a milestone) of numerical analysis is the so called **Lax-Richtmyer theorem** (or **equivalence theorem**):

Theorem. *For a consistent numerical method, stability is equivalent to convergence.*

Proof. Indeed, under these assumptions we have

$$||x(d+\delta d_n) - x_n(d+\delta d_n)|| \leq ||x(d+\delta d_n) - x(d)|| + ||x(d) - x_n(d)|| + ||x_n(d) - x_n(d+\delta d_n)||.$$

Thanks to (3), the first term at right-hand side can be bounded by $||\delta d_n||$ (up to a multiplicative constant independent of δd_n). A similar bound holds for the third term, due to the stability property (8). Finally, concerning the remaining term, if F_n is differentiable with respect to the variable x, an expansion in a Taylor series gives

$$F_n(x(d),d) - F_n(x_n(d),d) = \left.\frac{\partial F_n}{\partial x}\right|_{(\bar{x},d)} (x(d) - x_n(d)),$$

for a suitable $\bar{x}$ "between" $x(d)$ and $x_n(d)$. Assuming also that $\partial F_n/\partial x$ is invertible, we get

$$x(d) - x_n(d) = \left.\left(\frac{\partial F_n}{\partial x}\right)^{-1}\right|_{(\bar{x},d)} \left[F_n(x(d),d) - F_n(x_n(d),d)\right].$$

On the other hand, replacing $F_n(x_n(d),d)$ with $F(x(d),d)$ (since both terms are equal to zero) and passing to the norms, we find

$$||x(d) - x_n(d)|| \leq \left|\left|\left.\left(\frac{\partial F_n}{\partial x}\right)^{-1}\right|_{(\bar{x},d)}\right|\right| \, ||F_n(x(d),d) - F(x(d),d)||.$$

Thanks to (9) we can thus conclude that $||x(d) - x_n(d)|| \to 0$ for $n \to \infty$.

□