2. Nonlinear systems

Given a positive integer n and a continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$, find a zero of f, i.e., $x \in \operatorname{dom} f$ such that $f(x) = 0$.

This problem, ubiquitous in science and engineering, has been studied by some of the most brilliant minds and is still the subject of intense research nowadays. The main reason is that, in general, there is no algorithm that computes a zero of f within finitely many arithmetic operations. Thus, iterative methods are necessary.

Example. Example of a function f that has no zero: $f(x) := x^2 + 1$. Examples of a function f whose zeros cannot be computed within finitely many arithmetic operations: most polynomial functions of degree at least five, e.g., $f(x) := x^5 - 4x - 2$, and many other functions, e.g., $f(x) := \cos x - x$ and $f(x) := e^x - x - 2$.

The general approach in numerical analysis to tackle such problems is establishing conditions that ensure the existence of a solution and designing an iterative method, i.e., a method that, given an initial guess, generates a sequence that hopefully converges to a solution.

- The two are related: the existence theorems that we are going to see in this lecture admit a constructive proof that shows that an iterative method converges to a solution.
- The speed of convergence is a criterion to compare iterative methods: the faster the convergence the better.
- Another criterion is the computational cost per iteration.

Contents:

1. bisection method and intermediate value theorem;
2. fixed-point method and Banach fixed-point theorem;
3. Newton's method;
4. secant method;
5. stopping criteria.

1. Bisection method and intermediate value theorem

In this section, we focus on the case where $n = 1$. Let $a, b \in \mathbb{R}$ such that $a < b$. Let $f : \mathbb{R} \to \mathbb{R}$ be defined and continuous on $[a, b]$.

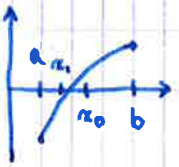
Theorem. If $f(a)\,f(b) < 0$, then there exists $c \in (a, b)$ such that $f(c) = 0$.

The proof simply establishes the convergence of the bisection method.
Without loss of generality, assume that $f(a) < 0$ and $f(b) > 0$.

```
a_0 ← a; b_0 ← b; x_0 ← (a+b)/2; i ← 0;
while f(x_i) ≠ 0 do
    if f(x_i) < 0 then
        a_{i+1} ← x_i; b_{i+1} ← b_i;
    else
        a_{i+1} ← a_i; b_{i+1} ← x_i;
    end
    i ← i+1; x_i ← (a_i + b_i)/2;
end
```

Bisection method

Either the method finds a zero of f after a finite number of iterations or it generates infinite sequences $(a_i)_{i\in\mathbb{N}}$, $(b_i)_{i\in\mathbb{N}}$, and $(x_i)_{i\in\mathbb{N}}$ such that, for all $i \in \mathbb{N}$, $a_i < x_i < b_i$ and $f(a_i) < 0$ and $f(b_i) > 0$. Since $(a_i)_{i\in\mathbb{N}}$ is monotonically nondecreasing and bounded from above, $(a_i)_{i\in\mathbb{N}}$ converges to $\sup_{i\in\mathbb{N}} a_i \in [a, b]$. Similarly, since $(b_i)_{i\in\mathbb{N}}$ is monotonically nonincreasing and bounded from below, $(b_i)_{i\in\mathbb{N}}$ converges to $\inf_{i\in\mathbb{N}} b_i \in [a, b]$. Moreover, since, for all $i \in \mathbb{N}$, $b_{i+1} - a_{i+1} = \frac{1}{2}(b_i - a_i)$ and thus $b_i - a_i = 2^{-i}(b - a)$, the sequences $(a_i)_{i\in\mathbb{N}}$ and $(b_i)_{i\in\mathbb{N}}$ have the same limit. Letting i tend to infinity in $f(a_i) < 0$ and $f(b_i) > 0$ shows that the limit is a zero f since f is continuous.



Unfortunately, no easy extension to the case where $n > 1$.

2. Fixed-point method and Banach fixed-point theorem

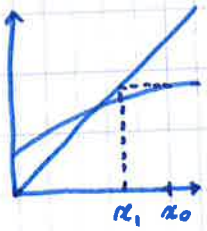
The following is a direct consequence of the preceding theorem.

Theorem (Brouwer fixed-point theorem).
Let $g : \mathbb{R} \to \mathbb{R}$ be defined and continuous on $[a,b]$. If $g([a,b]) \subseteq [a,b]$, then there exists $c \in [a,b]$ such that $g(c) = c$.

Proof. Apply the preceding theorem to the function f defined by $f(x) := g(x) - x$ for all $x \in [a,b]$. ■

Definition. A fixed point of a function $g : \mathbb{R}^n \to \mathbb{R}^n$ is a point $x \in \text{dom } g$ such that $g(x) = x$.

The proof of the preceding theorem shows that the problem of finding a zero of a function $f : \mathbb{R}^n \to \mathbb{R}^n$ can always be reformulated as the problem of finding a fixed point of a function $g : \mathbb{R}^n \to \mathbb{R}^n$. Furthermore, to find a fixed point of a function $g : \mathbb{R}^n \to \mathbb{R}^n$, it is natural to consider the iteration $x_{i+1} := g(x_i)$ for all $i \in \mathbb{N}$. Indeed, if g is continuous, then the generated sequence can converge only to fixed points of g.



For convergence, the slop must be smaller than 1.

We are going to establish a sufficient condition for convergence based on the following

Definition. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be defined on a nonempty subset S of $\mathbb{R}^n$. The function g is said to be Lipschitz continuous on S if

$$\text{Lip}_S(g) := \sup_{\substack{x,y \in S \\ x \neq y}} \frac{\|g(x) - g(y)\|}{\|x - y\|} < \infty.$$

The function g is called a contraction on S if $\text{Lip}_S(g) < 1$.
↳ distance between $g(x)$ and $g(y)$ < distance between x and y

Proposition. Let U be an open convex subset of $\mathbb{R}^n$ and let $g: U \to \mathbb{R}^n$ be differentiable. Then, g is Lipschitz continuous if and only if g' is bounded.

The proof shows that $\operatorname{Lip}_U(g) = \sup_U \|g'\|$

Proof. Assume that g' is bounded. Let x and y be two distinct points in U. By the mean value theorem,

$$\|g(x) - g(y)\| \leq \|x - y\| \sup_{t \in (0,1)} \|g'(x + t(y-x))\|.$$

Thus, $\operatorname{Lip}_U(g) \leq \sup_{x \in U} \|g'(x)\|$.

Conversely, assume that g is Lipschitz continuous. Let $x \in U$ and $h \in \mathbb{R}^n \setminus \{0\}$. For all $t \in (0, \infty)$ sufficiently small, $x + th \in U$ and

$$\frac{\|g'(x)h\|}{\|h\|} = \frac{\|g'(x)th\|}{\|th\|} \leq \underbrace{\frac{\|g(x+th) - g(x) - g'(x)th\|}{\|th\|}}_{\to 0 \text{ as } t \to 0} + \underbrace{\frac{\|g(x+th) - g(x)\|}{\|th\|}}_{\leq \operatorname{Lip}_U(g)}.$$

Thus, $\|g'(x)\| \leq \operatorname{Lip}_U(g)$. ∎

Theorem (Banach fixed-point theorem).
Let $g: \mathbb{R}^n \to \mathbb{R}^n$ be defined on a nonempty closed subset C of $\mathbb{R}^n$. If g is a contraction on C and $g(C) \subseteq C$, then g has a unique fixed point $x_* \in C$ and, for every $x_0 \in C$, $\|g^i(x_0) - x_*\| \leq \left(\operatorname{Lip}_C(g)\right)^i \|x_0 - x_*\|$ $\forall i \in \mathbb{N}$, which implies that $(g^i(x_0))_{i \in \mathbb{N}}$ converges to x_*.

Proof. ① Uniqueness of the fixed point. If x and y are fixed points of g, then $\|x - y\| = \|g(x) - g(y)\| \leq \operatorname{Lip}_C(g) \|x - y\|$, thus $(1 - \operatorname{Lip}_C(g)) \|x - y\| \leq 0$, hence $\|x - y\| \leq 0$, and therefore $x = y$.

② Inequality. Let $x_* \in C$ be a fixed point of g. Let $x_0 \in C$. For all $i \in \mathbb{N}$, $\|g^i(x_0) - x_*\| = \|g^i(x_0) - g^i(x_*)\| \leq \left(\operatorname{Lip}_C(g)\right)^i \|x_0 - x_*\|$.

③ Existence of the fixed point. Let $L := \operatorname{Lip}_C(g)$. First, for all $x, y \in C$,

$$\|x - y\| \leq \frac{\|x - g(x)\| + \|y - g(y)\|}{1 - L}$$

since

$$\|x - y\| \leq \|x - g(x)\| + \|g(x) - g(y)\| + \|g(y) - y\| \leq \|x - g(x)\| + L\|x - y\| + \|g(y) - y\|.$$

Based on this inequality, let us prove that, for every $x_0 \in C$, $(g^i(x_0))_{i \in \mathbb{N}}$ is a Cauchy sequence. For all $i, j \in \mathbb{N}$,

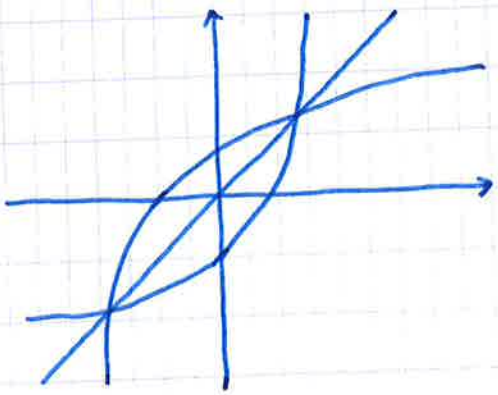
$$\|g^i(x_0) - g^j(x_0)\| \leq \frac{\|g^i(x_0) - g^{i+1}(x_0)\| + \|g^j(x_0) - g^{j+1}(x_0)\|}{1 - L} \leq \frac{L^i + L^j}{1 - L} \|x_1 - x_0\|.$$

Since $\mathbb{R}^n$ is complete, $(g^i(x_0))_{i\in\mathbb{N}}$ converges. Moreover, letting i tend to infinity in $g^{i+1}(x_0) = g(g^i(x_0))$ shows that the limit is a fixed point of g since g is continuous on C. ■

There are often several ways of turning the problem of finding a zero into that of finding a fixed point, and some ways may be better than others, as illustrated next.

Example. Let $f : \mathbb{R} \to \mathbb{R} : x \mapsto e^x - x - 2$. Then, $f(-2) > 0$, $f(-1) < 0$, $f(1) < 0$, and $f(2) > 0$. Thus, f has a zero on $(-2,-1)$ and a zero on $(1,2)$.

- Define $g : \mathbb{R} \to \mathbb{R} : x \mapsto e^x - 2$. Then, the zeros of f are exactly the fixed points of g. For all $x \in (-2,-1)$, $-2 < g(-2) < g(x) < g(-1) < -1$. For all $x \in \mathbb{R}$, $g'(x) = e^x$. Thus, for all $x \in (-2,-1)$, $0 < e^{-2} = g'(-2) < g'(x) < g'(-1) = e^{-1} < 1$. Hence, g is a contraction on $[-2,-1]$. However, for all $x \in [0,\infty)$, $1 \leq g'(x)$. Therefore, g is not a contraction on any subinterval of $[0,\infty)$. In conclusion, with this g, the fixed-point method can be used to find the fixed point in $(-2,-1)$ but not the fixed point in $(1,2)$.
- Define $g : (-2,\infty) \to \mathbb{R} : x \mapsto \ln(x+2)$. For all $x \in (1,2)$, $1 < \ln 3 = g(1) < g(x) < g(2) = 2\ln 2 < 2$. For all $x \in (-2,\infty)$, $g'(x) = \frac{1}{x+2}$. Thus, for all $x \in (1,2)$, $0 < \frac{1}{4} = g'(2) < g'(x) < g'(1) = \frac{1}{3} < 1$. Hence, g is a contraction on $[1,2]$. However, for all $x \in (-2,-1)$, $1 \leq g'(x)$. Therefore, g is not a contraction on $(-2,-1)$. In conclusion, with this g, the fixed-point method can be used to find the fixed point in $(1,2)$ but not the fixed point in $(-2,-1)$.



3. Newton's method

Faster convergence can be guaranteed if f is differentiable. The idea is to replace f with its first-order Taylor polynomial at every iteration. Thus, given $x_0 \in \operatorname{dom} f$, solve $f(x_0) + f'(x_0)(x - x_0) = 0$ instead of $f(x) = 0$. If $f'(x_0)$ is invertible, then the linear system has a unique solution $x_1 := x_0 - f'(x_0)^{-1} f(x_0)$. This yields the following method.

```
i ← 0
while f(x_i) ≠ 0 do
    try to find x ∈ ℝ^n such that f'(x_i) x = −f(x_i), e.g., by Gaussian el.
    if f'(x_i) is not invertible then
        stop
    else
        x_{i+1} ← x_i + x
        i ← i+1
    end
end
```

Newton's method

Newton's method is a fixed-point method for $g : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto x - f'(x)^{-1} f(x)$.

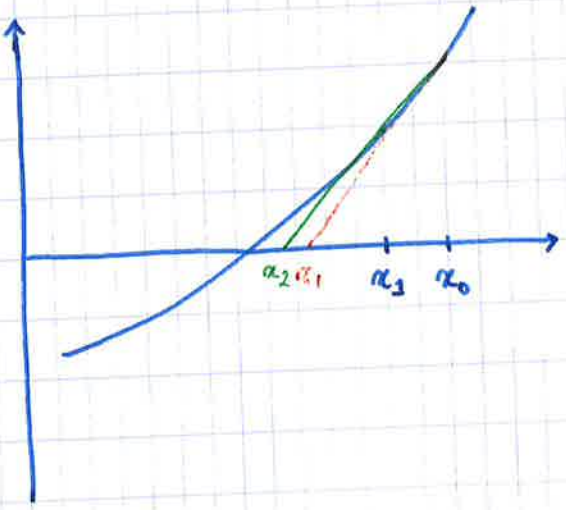
Theorem (Ortega & Rheinboldt 2000, §§ 10.2.2).
Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be defined and differentiable on a nonempty open subset U of $\mathbb{R}^n$. Let $x_* \in U$ be a zero of f. If f' is continuous at x_* and $f'(x_*)$ is invertible, then there exists $\delta \in (0, \infty)$ such that $B(x_*, \delta) \subseteq U$, $f'(x)$ is invertible for all $x \in B(x_*, \delta)$, and $\lim_{x \to x_*} \frac{\|g(x) - x_*\|}{\|x - x_*\|} = 0$, which implies that for all $x \in B(x_*, \delta)$ sufficiently close to x_*, the sequence $(g^i(x))_{i \in \mathbb{N}}$ converges to x_*. If, moreover, there exist $L \in [0, \infty)$ and $p \in (0, 1]$ such that, for all $x \in B(x_*, \delta)$, $\|f'(x) - f'(x_*)\| \leq L \|x - x_*\|^p$, then $\|g(x) - x_*\| \leq 4L \|f'(x_*)^{-1}\| \, \|x - x_*\|^{p+1}$ for all $x \in B(x_*, \delta)$.

4. Secant method

In this section, we focus on the case where $n = 1$. To avoid computing the derivative of f, we can replace f with its secant instead of its tangent. Thus, given $x_0, x_1 \in \operatorname{dom} f$, $x_0 \neq x_1$, we solve $f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}(x - x_0) = 0$ instead of $f(x) = 0$. If $f(x_0) \neq f(x_1)$, then the unique solution is $x_2 := \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)}$. This yields the following method.

$i \leftarrow 0$
while $0 \neq f(x_i) \neq f(x_{i+1}) \neq 0$ do
$x_{i+2} \leftarrow \frac{x_i f(x_{i+1}) - x_{i+1} f(x_i)}{f(x_{i+1}) - f(x_i)}$
$i \leftarrow i+1$
end

Secant method



5. Stopping criteria

Stop when $\|f(x_i)\| \leq \varepsilon$ or $\|x_{i+1} - x_i\| \leq \varepsilon$ for some $\varepsilon \in (0, \infty)$. The inequalities obtained in the convergence analyses sometimes enable to find $i \in \mathbb{N}$ such that the second stopping criterion is satisfied. For illustration, let us find $i \in \mathbb{N}$ such that $\|x_i - x_*\| \leq \varepsilon$ for the bisection method, Newton's method, and the fixed-point method.

- Bisection method: $\boxed{\left\lceil \log_2\left(\frac{b-a}{\varepsilon}\right) - 1 \right\rceil}$

  For all $i \in \mathbb{N}$, $|x_i - x_*| \leq \frac{1}{2}(b_i - a_i) = 2^{-i-1}(b-a)$ Thus $|x_i - x_*| \leq \varepsilon$ if $2^{-i-1}(b-a) \leq \varepsilon$, i.e., $i \geq \log_2\left(\frac{b-a}{\varepsilon}\right) - 1$.

- Fixed-point method: $\boxed{\left\lceil \ln\left(\frac{\varepsilon(1-\mathrm{Lip}_c(g))}{\|x_0 - x_1\|}\right) \Big/ \ln \mathrm{Lip}_c(g) \right\rceil}$

  For all $i \in \mathbb{N}$, $\|x_i - x_*\| \leq \left(\mathrm{Lip}_c(g)\right)^i \|x_0 - x_*\|$ Thus $\|x_i - x_*\| \leq \varepsilon$ if $\left(\mathrm{Lip}_c(g)\right)^i \|x_0 - x_*\| \leq \varepsilon$, i.e., $i \geq \frac{\ln\left(\frac{\varepsilon}{\|x_0 - x_*\|}\right)}{\ln \mathrm{Lip}_c(g)}$.

  Moreover, $\|x_0 - x_*\| \leq \|x_0 - x_1\| + \underbrace{\|x_1 - x_*\|}_{= \|g(x_0) - g(x_*)\| \leq \mathrm{Lip}_c(g)\|x_0 - x_*\|}$ Thus $\|x_0 - x_*\| \leq \frac{1}{1 - \mathrm{Lip}_c(g)} \|x_0 - x_1\|$.

  Hence, $\|x_i - x_*\| \leq \varepsilon$ if $\frac{\left(\mathrm{Lip}_c(g)\right)^i}{1 - \mathrm{Lip}_c(g)} \|x_0 - x_1\| \leq \varepsilon$, i.e.,

  $$i \geq \frac{\ln\left(\frac{\varepsilon(1-\mathrm{Lip}_c(g))}{\|x_0 - x_1\|}\right)}{\ln \mathrm{Lip}_c(g)}.$$

- Newton's method in the case where $p = 1$: $\boxed{\left\lceil \log_2\left(\frac{\ln(c(x_*)\varepsilon)}{\ln(c(x_*)\|x_0 - x_*\|)}\right) \right\rceil}$

  For all $x \in B(x_*, \delta)$, $\|g(x) - x_*\| \leq c(x_*)\|x - x_*\|^2$ with $c(x_*) := 4L\|f'(x_*)^{-1}\|$,
  Thus $g(x) \in B(x_*, \delta)$ if $c(x_*)\delta\|x - x_*\| \leq \delta$, i.e., $\|x - x_*\| \leq \frac{1}{c(x_*)}$.
  Let $e := \min\left\{\delta, \frac{1}{c(x_*)}\right\}$. Then, $g(B(x_*, e)) \subseteq B(x_*, e)$.
  Thus, for all $x_0 \in B(x_*, e)$ and $i \in \mathbb{N}$, $\|x_i - x_*\| \leq \frac{1}{c(x_*)}\left(c(x_*)\|x_0 - x_*\|\right)^{2^i}$,
  hence $\|x_i - x_*\| \leq \varepsilon$ if $\frac{1}{c(x_*)}\left(c(x_*)\|x_0 - x_*\|\right)^{2^i} \leq \varepsilon$, i.e.,

  $$i \geq \log_2\left(\frac{\ln(c(x_*)\varepsilon)}{\ln(c(x_*)\|x_0 - x_*\|)}\right).$$

References:

- J. M. Ortega & W. C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, SIAM, 2000;
- E. Süli & D. F. Mayers, An Introduction to Numerical Analysis, Cambridge University Press, 2012.