# Computational methods-Lecture 5

## Roots of nonlinear equations of one variable

Next, we will move onto solving the scalar nonlinear equations (equations of one variables). This will be finished in two or three lectures (one or 1.5 weeks). Then, we move onto linear systems (the matrices), which will be the second main part of our course. The third part of our course will be about solutions or optimization of nonlinear systems, and a little bit about solving ordinary differential equations.

## 1 The bisection method

The simplest possible solving nonlinear equations will the binary search method, or **bisection method**.

Let $f \in C[a,b]$ with $f(a)f(b) < 0$, then there must be a root on $(a,b)$.

The method is like this:

1. Let $[a_1, b_1] = [a, b]$. Fix tolerance $\epsilon$.
2. For $i = 1, 2, \cdots,$

   Compute the midpoint

   $$p_i = a_i + \frac{b_i - a_i}{2}.$$

   If $f(p_i) = 0$ or $\frac{b_i - a_i}{2} < \epsilon$, then stop and output $p_i$.
   Esle If $f(a_i)f(p_i) > 0$, set $[a_{i+1}, b_{i+1}] = [p_i, b_i]$;
   otherwise $[a_{i+1}, b_{i+1}] = [a_i, p_i]$.

Clearly, the error satisfies

$$|p_n - p| \leq \frac{b-a}{2^n}.$$

Though the error decays exponentially in $n$, this is called the **linear convergence in optimization**, and one often desires superlinear convergence, as we will define later.

The bisection method cannot be generalized to functions of several variables. Also, the convergence is slow.

## 2 Fixed point iteration

The idea is like this: one rewrites the equation $f(x) = 0$ into

$$x = g(x)$$

for some suitable $g$. One example is $g = x - \lambda f(x)$ for some $\lambda > 0$. Of course, there are other ways to construct. Different construction gives different behaviors.

Then, one repeats the the process

$$x_{n+1} = g(x_n)$$

hoping that $x_n \to x^*$. If it has a limit, then $x^*$ will be a solution. *Exercise: Think about why.*

The point $x^*$ that satisfies

$$x^* = g(x^*)$$

is called a **fixed point** of $g$.

There are many fixed point theorems in mathematics. Here, we state two of them.

### Brouwer fixed point theorem

The first is the **Brouwer fixed point theorem** for topology. It says that if a function that sends a disk in some topology space **into** the same disk is continuous, then it has a fixed point.

In our special case, it becomes the following

**Theorem 1.** *Let $g \in C[a, b]$ such that $g(x) \in [a, b]$ for all $x \in [a, b]$, then $g$ has a fixed point on $[a, b]$.*

### Contraction mapping theorem

The general statement from mathematics is like this:

**Theorem 2.** *Let $(X, d)$ be a complete metric space and that $f : X \to X$ is a contraction mapping in the sense that there exists some $q \in (0, 1)$ such that*

$$d(f(x), f(y)) \leq q d(x, y), \ \forall x, y \in X,$$

*then $f$ has a unique fixed point $z$ and that the sequence $f^n(x_0) \to z$ for any $x_0$.*

In our case, the theorem is like this

**Theorem 3.** *Let $g \in C[a,b]$ satisfy that $g(x) \in [a,b]$ for $x \in [a,b]$. If there is a constant $q \in (0,1)$ such that*

$$|g(x) - g(y)| \leq q|x - y|, \ \forall x, y \in (a,b),$$

*then $g$ has a unique fixed point $z$ in $[a,b]$. Moreover, the sequence $x_n = g(x_{n-1})$ converges to $z$ with the rate*

$$|x_n - z| \leq q^n |x_0 - z|$$

*and that*

$$|x_n - z| \leq \frac{q^n}{1-q}|x_1 - x_0|$$

*Proof.* Consider the sequence $\{x_n\}$. One has

$$|x_n - x_m| = |g(x_{n-1}) - g(x_{m-1})| \leq q^{\min(m,n)}|b - a|.$$

Hence, $\{x_n\}$ is a Cauchy sequence, and it must have a limit $z$. Take the limit in

$$x_n = g(x_{n-1}),$$

one then has $z = g(z)$.

The uniqueness follows easily from the inequality $|z_1 - z_2| \leq |g(z_1) - g(z_2)| \leq q|z_1 - z_2|$. Hence, one must have $|z_1 - z_2| = 0$.

Then,

$$|x_n - z| \leq q|x_{n-1} - z| \leq \cdots \leq q^n |x_0 - z|.$$

For the posterior error bound:

$$|z - x_n| \leq \sum_{m=n}^{\infty} |x_{m+1} - x_m| \leq \sum_{m=n}^{\infty} q^m |x_1 - x_0| = \frac{q^n}{1-q}|x_1 - x_0|.$$

□

The advantage of the second estimate is that we know the value of $x_0$ and $x_1$. Such type of estimates using computed numerical values to bound errors are called **posterior error estimates**.

**Example** Find the root of $x^3 + 4x^2 - 10 = 0$ in [1, 2] using the fixed point iteration.

*Method 1* We set $g(x) = x - f(x) = x - x^3 - 4x^2 + 10$. You can see that $|g'| > 1$ for many values and the image of $[1,2]$ is not contained in $[1,2]$. The sequence constructed using $x_{n+1} = g(x_n)$ is likely to diverge.

*Method 2* Consider

$$x = \sqrt{\frac{10}{4+x}} = g(x)$$

It is easily verified that $|g'| < 0.15$ and $g[1,2] \subset [1,2]$. Then, $x_{n+1} = g(x_n)$ works.

# 3 Newton's method

The Newton's method is also to construct an iterative method, but it enjoys bettor **local convergence** property.

Consider a function $f(x)$ and some approximation $x_k$ for the root. We then approximate $f$ using its tangent line at $x_k$:

$$f(x_k) + f'(x_k)(x - x_k) \approx f(x).$$

Hence, the root is found to be

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

This is also an iterative method, which is called the **Newton's method**. The fixed point of $g(x) = x - \frac{f(x)}{f'(x)}$ is then a root of $f$.

**Draw the picture for showing how it works geometrically.**

**Example** Consider the example $f(x) = x^3+4x^2-10$ again. By Newton's method, the iteration is given by

$$x_{k+1} = x_k - \frac{x_k^3 + 4x_k^2 - 10}{3x_k^2 + 8x_k}.$$

In practice, the Newton's method either **diverges** or **converges very fast**, depending on whether your initial guess is good enough. If you initial guess is close to the true root, then often it converges fast. In next section, we will try to see why it converges fast if the initial guess is good.

In practice, computing $f'(x)$ may be hard. In this case, we note

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Hence, we can use the quotient difference to approximate:

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

Then,

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}.$$

This is called the secant method.

For the secant method, one must provides two initial guesses. The secant method converges slower than Netwon's method but still superlinearly. If you use three points to interpolate, you get a parabola, which then leads to the so-called Müller's method.

# 4 Error analysis for iterative methods

**Definition 1.** *If $g(x)$ has a fixed point at $x^*$, and there is a neighborhood of $x^*$: $R = (x^* - \delta, x^* + \delta)$ such that for any $x_0 \in R$, the sequence $x_k = g^k(x_0)$ converges to $x^*$, we say the iterative method converges* **locally**.

In mathematics, "local" is often related to "neighborhoods" or the so-called compact sets.

**Definition 2.** *Suppose $\{x_k\}$ is a sequence that converges to $x^*$. If there exists $p > 0$ and $\lambda > 0$ such that*

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} \to \lambda, \quad k \to \infty,$$

*then we say $\{x_k\}$ converges to $x^*$ with order $p$, and asymptotic error constant $\lambda$. If $p = 1$, we say it converges linearly, we say it converges superlinearly if $p > 1$, and we say it converges quadratically if $p = 2$.*

Using the contraction mapping theorem, we can show easily that

**Theorem 4.** *If $x^*$ is a fixed point of $g$ and $|g'(x)| \leq q < 1$ in some a neighborhood $R = (x^* - \delta, x^* + \delta)$, then the iterative method converges locally.*

This will be left as a homework problem. In fact, using this, you can show that Newton's method converges locally if $g \in C^2$ and $g'(x^*) \neq 0$.

**Example** Assume $g \in C^1$ in the interative method. Let $x^*$ be a fixed point and $|g'(x)| \leq q < 1$ but $g'(x^*) \neq 0$ in the neighborhood, then the method converges linearly, but not superlinearly.

First of all, by the contraction mapping theorem, we know $x_k \to x^*$ as $x \to \infty$. However,

$$x_{k+1} - x^* = g(x_k) - g(x^*) = g'(\xi_k)(x_k - x^*)$$

Hence,

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|} = |g'(\xi_k)| \to |g'(x^*)|.$$

Now, we consider that $g$ is $p$th order continuously differentiable and $x^*$ is a **zero of multiplicity** $p$ of $g(x) - g(x^*)$. In other words, $g^{(k)}$ is continuous around $x^*$ for all $k \leq p$ and

$$g^{(k)}(x^*) = 0, \ \forall 1 \leq k \leq p-1, \ g^{(p)}(x^*) \neq 0.$$

In this case,

$$g(x) - g(x^*) = (x - x^*)^p h(x)$$

where $h(x)$ is a continuous function around $x^*$ such that $h(x^*) \neq 0$.

We have the following theorem:

**Theorem 5.** *Let $g$ be $p$th order continuously differentiable around $x^*$ and $x^*$ is a* **zero of multiplicity** *$p$ of $g(x) - g(x^*)$. Then, the iterative method $x_{k+1} = g(x_k)$ converges with order $p$.*

*Proof.* Again, $|g'(x)| \leq q < 1$ for $\delta$ small enough. Then, the contraction mapping theorem guarantees the convergence of $\{x_k\}$ to $x^*$.

By the Lagrangian remainder theorem:

$$x_{k+1} - x^* = \frac{g^{(p)}(\xi_k)}{p!}(x_k - x^*)^p.$$

Hence,

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} \to \frac{|g^{(p)}(x^*)|}{p!}$$

□

Now, we check Newton's method.If $f'(x^*) \neq 0$ and $f''(x^*) \neq 0$, then we see that it converges quadratically:

$$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}, \quad g''(x^*) = \frac{f''(x^*)}{f'(x^*)}$$

Hence, we basically know that the error

$$e_{k+1} \leq Ce_k^2$$

for quadratically convergent method.

Hence, for the example we discussed, the Newton's method

$$x_{k+1} = x_k - \frac{x_k^3 + 4x_k^2 - 10}{3x_k^2 + 8x_k}.$$

can converge faster than

$$x_{n+1} = \sqrt{\frac{10}{x_n + 4}}$$

In general, quadratical convergence is hard. We expect $p \in [1, 2)$ and superlinear convergence is good enough. For example, for the secant method $p \approx 1.618$

## 5 Accelerating convergence*(Not required)

There are two methods that are often used to accelerate the convergence: Aitken's and Steffensen's methods. You can read the reference books.