17. Lecture 17

Nonlinear Equations

Essentially, the only way that one can solve nonlinear equations is by iteration.

The quadratic formula enables one to compute the roots of $p(x) = 0$ when $p \in \mathbb{P}^2$. Formulas were derived for finding the roots of $p \in \mathbb{P}^3$ and $p \in \mathbb{P}^4$ by expressions involving radicals ($\sim$ 1545). The case of quintics was studied unsuccessfully for almost 300 years until Abel (1824) proved *no such formula existed*. As we shall see, the roots of quintics (and other nonlinear equations) can be solved by iteration!

Example 17.1 (Eigenvalues)**.** *Let A be a $n \times n$ matrix with $n \geq 5$. We propose to compute the eigenvalues of A. The eigenvalues of A are the roots of the characteristic polynomial, i.e.*

$$p(\lambda) = \det(A - \lambda I).$$

The characteristic polynomial p has degree n and leading term is $(-1)^n \lambda^n$. This problem can only be solved by iteration when $n \geq 5$.

Bisection method. Suppose f is continuous on $[a, b]$ and satisfies $f(a)f(b) < 0$ (not the same sign). The intermediate value theorem tells us there is at least 1 root $f(x^*) = 0$, $x^* \in (a, b)$. The bisection method is used to finding such root.

Bisection Algorithm (mathematical)
  Set $a_0 = a$ and $b_0 = b$
  For $i = 0, 1, 2, ...$
    Set $a_{i+\frac{1}{2}} = \frac{a_i + b_i}{2}$
    If $f(a_{i+\frac{1}{2}}) = 0$
      STOP, $a_{i+\frac{1}{2}}$ is the desired root
    Else If $f(a_{i+\frac{1}{2}})f(a_i) > 0$
      Set $a_{i+1} = a_{i+\frac{1}{2}}$ and $b_{i+1} = b_i$
    Else
      Set $a_{i+1} = a_i$ and $b_{i+1} = a_{i+\frac{1}{2}}$
  End For

It is obvious that after j steps, either we have found a root or there is a root in (a_j, b_j). Note that in that case the root x^* is at most half of the interval length away from either a_j or b_j, i.e.

$$x^* = \frac{a_j + b_j}{2} + \varepsilon,$$

where

$$\varepsilon < b_j - \frac{a_j + b_j}{2} = \frac{b_j - a_j}{2} = \frac{b - a}{2^j}.$$

We now describe the *matlab* version of the mathematical bisection algorithm using minimal memory usage.

```
%%% R is the root approximation after N steps
%%% A,B are real numbers
%%% F is a continuous function
%%% F(A)F(B)<=0
Function R=BISECTION(A,B,N,F)

```

```
%% Preliminaries: A or B are a root, F(A)F(B)>0
    SA = sign(F(A));
    if (SA == 0)
        R=A;
        return;
    end

    SB = sign(F(B));
    if (SB == 0)
        R=B;
        return;
    end

    if (SA == SB)
      fprintf('Input error to bisection\n');
      R=NaN(1);  %return not a number
      return;
    end

%% all the preliminaries are done
   for I=1:N
      AV = (A_B)/2;
      FAV = F(AV);
      S=sign(FAV);
      if (S==0)
        R=AV;
        return;
      end
      if (S==SA)
        A=AV;
      else
        B=AV;
      end
   end
end
```

This algorithm will only get the real roots of a real valued function $f : \mathbb{R} \to \mathbb{R}$.

Example 17.2 (Cubic). *Let $f(x) = x^3 + x = x(x^2+1)$. The roots of f are $(0, i, -i)$ and bisection will only get the real roots, i.e. $x = 0$ To find the complex roots, we need to treat f as a complex valued functions. We define $F : \mathbb{R}^2 \simeq \mathbb{C} \to \mathbb{C} \simeq \mathbb{R}^2$ by*

$$F\begin{pmatrix} R \\ I \end{pmatrix} \sim f(R+iI) = (R+iI)^3 + (R+iI) = R^3 + 3iR^2I - 3RI^2 - iI^3 + (R+iI)$$
$$= (R^3 - 3RI^2 + R) + (3R^2I - I^3 + I)i \sim \begin{pmatrix} R^3 - 3RI^2 + R \\ 3R^2I - I^3 + I \end{pmatrix}.$$

for real number R, I. Find the complex roots correspond to a system

$$F\begin{pmatrix} R \\ I \end{pmatrix} := \begin{pmatrix} R^3 - 3RI^2 + R \\ 3R^2 I - I^3 + I \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The previous illustrates the need of *iterative techniques for systems* (but is far from the only reason).

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a vector values function on $\mathbb{R}^n$. We want to find roots $x^* \in \mathbb{R}^n$ satisfying

$$F(x^*) = 0 \in \mathbb{R}^n.$$

This gives n equations and $x_1^*, ..., x_n^*$ are the n unknown.

Example 17.3 ($n = 2$)**.**

$$F\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{pmatrix} x_1^2 + 4x_2^2 - 1 \\ 4x_1^2 + x_2^2 - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Refer to Figure 6 for an illustration of the situation.

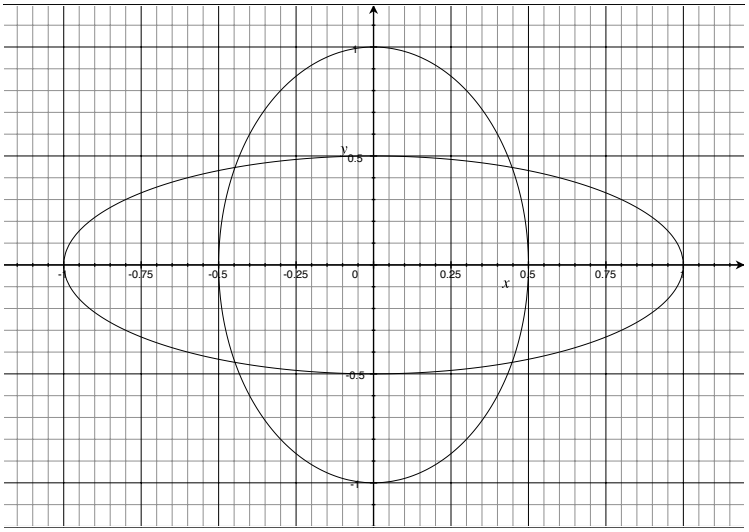

FIGURE 6. There are 4 roots of the system $x_1^2 + 4x_2^2 - 1 = 0$ and $4x_1^2 + x_2^2 - 1 = 0$.

Example 17.4 ($n = 1$)**.** *Consider*

$$f(x) = \sin(x) = 0.$$

The roots are $x = j\pi$, $j \in \mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$.

Definition 17.1 (Fixed Point Equation)**.** *A* fixed point equation *is one of the form*

$$x = G(x),$$

with $x \in \mathbb{R}^n$ and $G : \mathbb{R}^n \to \mathbb{R}^n$.

A solution to a fixed point equations is $x^* \in \mathbb{R}^n$ satisfies

$$x^* = G(x^*).$$

We can turn $F(x) = 0$ into a fixed point problem, i.e.

$$x = x - F(x), \qquad \text{i.e. } G(x) := x - F(x).$$

This means that x^* solves $F(x^*) = 0$ if and only if $x^* = G(x^*)$. Obviously fixed point problems can be turned into $F(x) = 0$ by setting $F(x) = x - G(x)$. We could have also used

$$x = x - BF(x) =: G(x)$$

for any non singular $n \times n$ matrix B. We will see the importance of the matrix B soon.

18. Lecture 18

We considered in the last lecture the fixed point formulation: Find $x^* \in \mathbb{R}^n$ satisfying

$$x^* = G(x^*),$$

where $G : \mathbb{R}^n \to \mathbb{R}^n$. We now discuss an algorithm approximating such x^*.

Fixed Point Iteration or Picard Iteration. Start with an initial iterate (guess) $x^0 \in \mathbb{R}^n$. Then, for $i = 0, 1, 2, ...$ set

$$x^{i+1} = G(x^i).$$

We now want to understand when $x^i \to x^*$.

Definition 18.1 (Lipschitz)**.** *Let $\Omega \subset \mathbb{R}^n$. The vector-valued function $G : \Omega \to \mathbb{R}^n$ is called* Lipschitz conitnuous *if there is a $M \geq 0$ with*

$$\|G(x) - G(y)\|_\infty \leq M\|x - y\|_\infty$$

for all $x, y \in \Omega$. Here for $w \in \mathbb{R}^n$

$$\|w\|_\infty := \max_{i=1,...,n} |w_i|.$$

Definition 18.2 (Contraction Mapping)**.** *Let $\Omega \subset \mathbb{R}^n$. The vector-valued function $G : \Omega \to \Omega$ is a* contraction mapping *if G is Lipschitz continuous with constant $M = \rho < 1$.*

Theorem 18.1 (Contraction Mapping Theorem)**.** *Let Ω be a closed subset of $\mathbb{R}^n$ and G be a contraction mapping of Ω into Ω with constant ρ. Then, there is a unique fixed point $x^* \in \Omega$ (satisfying $x^* = G(x^*)$) and the Picard iteration $\{x^j\}_{j=0}^\infty$, starting with any $x^0 \in \Omega$, converges to x^* and satisfies*

$$\|x^j - x^*\|_\infty \leq \frac{\rho^j}{1-\rho}\|x^0 - x^*\|_\infty.$$

This is often called linear convergence. We postpone the proof for later.

To understand the contraction mapping hypothesis, we consider the equation $F(x) = 0$, where $F : \mathbb{R} \to \mathbb{R}$, and assume that it has as solution $x^* \in \mathbb{R}$, i.e. $F(x^*) = 0$. Furthermore, we assume that $F \in C^2$ in a neighborhood $B_{\delta_1}(x^*) := [x^* - \delta_1, x^* + \delta_1]$ and that $F'(x^*) \neq 0$. Notice that the latter condition, guarantees that there is a neighborhood $B_{\delta_2}(x^*)$ ($\delta_2 \leq \delta_1$) such that

$$|F'(\xi)| \geq \frac{1}{2}|F'(x^*)|$$

for every $\xi \in B_{\delta_2}(x^*)$. This implies that

$$|F'(\xi)|^{-1} \leq 2|F'(x^*)|^{-1} \qquad \xi \in B_{\delta_2}(x^*).$$

Now for $\delta \leq \delta_2$ (to be determined), we pick $w \in B_\delta(x^*)$ and set

$$G(x) = x - (F'(w))^{-1}F(x). \tag{9}$$

Clearly x^* is a fixed point of G and for $x, y \in B_\delta(x^*)$

$$G(x) - G(y) = G'(y)(x - y) + \frac{G''(\xi)}{2}(x - y)^2,$$

for some ξ between x and y. Therefore,

$$|G(x) - G(y)| \leq |G'(y)||x - y| + \frac{|G''(\xi)|}{2}(x - y)^2.$$

However,

$$G'(y) = 1-(F'(w))^{-1}F'(y) = (F'(w)-F'(y))(F'(w))^{-1} = F''(\xi_1)(w-y)(F'(w))^{-1}$$

for ξ_1 between w and y. Now by the C^2 assumption, there exists a constant M_0 such that $|F''(\theta)| \leq M$ for every $\theta \in B_{\delta_1}(x^*)$. Thus,

$$|G'(y)| \leq 4M|F'(x^*)|^{-1}\delta$$

so

$$|G(x) - G(y)| \leq \left\{4M|F'(x^*)|^{-1}\delta + M\delta\right\}|x-y|, \tag{10}$$

using the fact that $|w-y| \leq 2\delta$ and $|x-y| \leq 2\delta$.

Given $0 < \rho < 1$, we chose δ so that

$$\left\{4M|F'(x^*)|^{-1} + M\right\}\delta < \rho.$$

This implies that G is a contraction mapping and so the Picard iterates

$$x^{i+1} = x^i - (F'(w))^{-1}F(x^i)$$

converges to x^* provided $|x^0-x^*| \leq \delta$ and $|w-x^*| \leq \delta$. The next theorem generalize this argumentation to $\mathbb{R}^n$.

Theorem 18.2 (Secant Algorithm)**.** *Assume $F : \mathbb{R}^n \to \mathbb{R}^n$ and $F(x^*) = 0$ for some $x^* \in \mathbb{R}^n$ with*

(1) *$F \in C^2$ in a neighborhood of x^*;*
(2) *$DF(x^*) =$ the derivative matrix at x^* given by*

$$(DF(x^*))_{ij} = \frac{\partial}{\partial x_j}F_i(x^*)$$

is nonsingular.

Then there is a $\delta > 0$ such that if $w \in B_\delta(x^*)$ and $x^0 \in B_\delta(x^*)$, the iteration

$$x^{i+1} = x^i - (DF(x^*))^{-1}F(x^i)$$

converges to x^*.

The proof is as in the case of scalar functions discussed before the theorem but more complicated due to matrix notations. The major obstacle to apply this algorithm is getting close enough to the root to come up with w (then we can always take $x^0 = w$).

19. Lecture 19

We start with the proof of the contraction mapping theorem (Theorem 18.1).

Proof of Theorem 18.1. Let $j \leq k$ and $l \geq 0$, $x^{j+1} = G(x^j)$ with $x^0 \in \Omega \subset \mathbb{R}^n$ (closed) and G a contraction (Lipschitz constant $\rho < 1$) on Ω. Note that

$$x^{k+l+1} - x^k = x^{k+l+1} - x^{k+l} + x^{k+l} - x^{k+l-1} + ... + x^{k+1} - x^k.$$

Now,

$$x^{m+1} - x^m = G(x^m) - G(x^{m-1})$$

and so

$$\|x^{m+1} - x^m\|_\infty = \|G(x^m) - G(x^{m-1})\|_\infty \leq \rho \|x^m - x^{m-1}\|_\infty.$$

Repeating

$$\|x^{m+k+1} - x^{m+k}\|_\infty \leq \rho^k \|x^{m+1} - x^m\|_\infty$$

and thus

$$\begin{aligned}
\|x^{k+l+1} - x^k\|_\infty &\leq \|x^{k+l+1} - x^{k+l}\|_\infty + \|x^{k+l} - x^{k+l-1}\|_\infty + ... + \|x^{k+1} - x^k\|_\infty \\
&\leq (\rho^l + \rho^{l-1} + ... + 1)\|x^{k+1} - x^k\|_\infty \\
&\leq (\rho^l + \rho^{l-1} + ... + 1)\rho^k \|x^1 - x^0\|_\infty \\
&\leq \underbrace{\rho^{k-j}}_{\leq 1} \frac{\rho^j}{1-\rho} \|x^1 - x^0\|_\infty \leq \frac{\rho^j}{1-\rho} \|x^1 - x^0\|_\infty.
\end{aligned}$$

This means that if $m, l > j$

$$\|x^m - x^l\|_\infty \leq \frac{\rho^j}{1-\rho} \|x^1 - x^0\|_\infty.$$

The quantity on the right side can be made as small as we want by taking j large. This implies that the sequence $\{x^j\}$ is a Cauchy sequence and so converges to some $x^* \in \Omega$. (Recall that $\mathbb{R}^n$ is complete and Ω closed implies that Ω is complete). Moreover,

$$\|x^* - G(x^*)\|_\infty \leq \|x^* - x^j + G(x^{j-1}) - G(x^*)\|_\infty \leq \|x^* - x^j\|_\infty + \rho \|x^* - x^{j-1}\|_\infty.$$

As x^j converges to x^*, the quantity on the right can be made as small as desired buy taking j large, i.e.

$$x^* = G(x^*).$$

This shows that every Picard iteration converges to a fixed point. These fixed points are unique. Indeed, if $x_1^* = G(x_1^*)$ is another fixed point, then

$$\|x_1^* - x^*\|_\infty = \|G(x_1^*) - G(x^*)\|_\infty \leq \rho \|x_1^* - x^*\|_\infty,$$

i.e.

$$(1-\rho)\|x_1^* - x^*\|_\infty \leq 0$$

and therefore $\|x_1^* - x^*\|_\infty$ or $x_1^* = x^*$. □

Newton's Method. We start with a motivation. We look for zero of $F(x) = 0$, where $F : \mathbb{R}^n \to \mathbb{R}^n$. Assume $F \in C^2$ and that there is a zero x^* of F such that $DF(x^*)$ is nonsingular. As in the previous lecture, there is a neighborhood $B_\delta(x^*)$ such that $DF(x)$ is nonsingular when $x \in B_\delta(x^*)$. Assume that $x^j \in B_\delta(x^*)$. The Taylor theorem for vector fields guarantees that

$$F(x) = F(x^i) + DF(x^j)(x - x^j) + \varepsilon(x),$$

where $\|\varepsilon(x)\| = O(\|x - x^j\|^2)$. Ignoring the error term and recalling that we want $F(x^*) = 0$, we chose x^{j+1} by

$$0 = F(x^j) + DF(x^j)(x^{j+1} - x^j), \textit{i.e.}$$
$$x^{j+1} = x^j - DF(x^j)^{-1}F(x^j).$$

This is the *Newton's* iterative method.

Note that this iteration is of the form

$$x^{j+1} = G_j(x^j), \qquad G_j(x) = x - DF(x^j)^{-1}F(x)$$

so it is not quite a fixed point iteration because G changes at each step!

The mathematical analysis of this iterative scheme follows that provided before Theorem 18.2. For simplicity, we again consider the case $\Omega \subset \mathbb{R}$. The main difference is that w in (9) is replaced by x^j. As in the previous analysis, for

(1) x^* satisfying $F(x^*) = 0$;
(2) $F'(x^*) \neq 0$;
(3) $F \in C^2$ in a neighborhood of x^*

we assume that $x^j \in B_\delta(x^*)$ with $\delta \leq \delta_2$ (δ to be determined).

Now

$$x^* - x^{j+1} = G_j(x^*) - G_j(x^j)$$

so that

$$|x^* - x^{j+1}| \leq |G_j(x^*) - G_j(x^j)|.$$

Since $x^j \in B_\delta(x^*)$, we have as in (10)

$$|G_j(x^*) - G_j(x^j)| \leq \left\{4M|F'(x^*)|^{-1} + M\right\} \delta |x^* - x^j|.$$

As long as $|x^* - x^j| \leq \delta_2$, we may take $\delta = |x^* - x^j|$ to conclude

$$|x^* - x^{j+1}| \leq M\{4|F'(x^*)|^{-1} + 1\}|x^* - x^j|^2.$$

This is called *quadratic convergence.*

Once Newton's iterates get close to the solution so that quadratic convergence kicks in, the convergence is extremely fast.

For a geometric interpretation of the Newton's method, we refer to Figure 7.

Example 19.1 (Newton)**.** *Consider the function $f(x) = x^3 e^{-x^2}$, which has only one root (at $x = 0$), see Figure 8. We compute*

$$f'(x) = (3x^2 - 2x^4)e^{-x^2}$$

so

$$f'(x) > 0 \qquad \textit{for} \qquad |x| < \sqrt{3/2}$$

and

$$f'(x) < 0 \qquad \textit{for} \qquad |x| > \sqrt{3/2}$$

The geometric interpretation of the Newton method implies that

(1) *if $x^0 > \sqrt{3/2}$, Newton's method diverge to ∞*

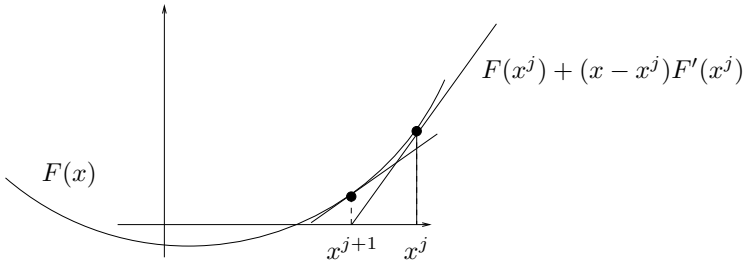

FIGURE 7. Geometric Interpretation of the Newton's method.

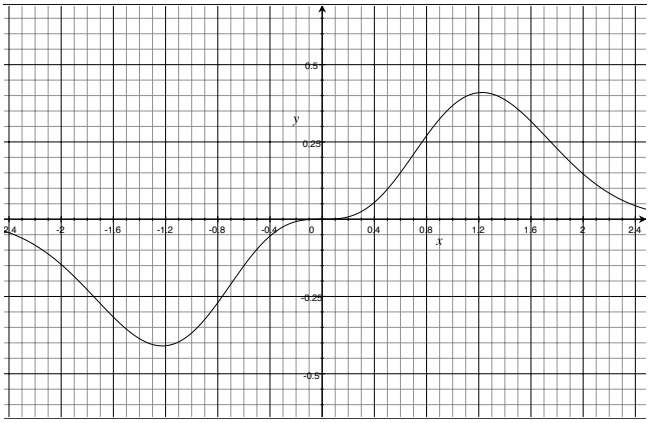

FIGURE 8. $f(x) = x^3 e^{-x^2}$.

(2) *if* $x^0 < -\sqrt{3/2}$, *Newton's method diverge to* $-\infty$
(3) *it diverge in a neighborhood of both* $-\sqrt{3/2}$ *and* $\sqrt{3/2}$
(4) *it converges to* 0 *otherwise.*