Lecture Notes

Kyle Chui

2022-01-04

Contents

1 Lecture 1

The goal of this class is to solve *mathematical* problems with the help of *computers*.

1.1 Chapter Summaries

1. Computations in computers
   - How to store (real) numbers in the computer

     > **Note.** If the number has finitely many digits, then it is simple. What about numbers with infinitely many digits, i.e. $\frac{1}{3}$? We have to truncate or round, and store an approximation with finitely many digits.

   - How to perform computations

     > **Note.** From regular math, we know that $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$. However, in the computer, due to errors, we have $\frac{1}{3} \oplus \frac{1}{3} = ? \oplus ? = ??$.

   - Errors
2. - Find roots of $f(x) = 0$ using bisection, Newton's method, ...
   - Convergence
   - Convergence order (how fast it converges)
3. Polynomial interpolation
   - Approximate a function $f(x)$ by a polynomial $P(x)$, where $f(x_i) = P(x_i)$ for finitely many x_i
   - *Accuracy* of the polynomial approximations
4. Numerical differentiations and numerical integrations
   - Using the approximations from chapter 3, we can approximate using

     $$f'(x^*) \approx P'(x^*) = \sum_{i=0}^{k} f(x_k)c_k,$$

     and

     $$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b P(x)\,\mathrm{d}x = \sum_{i=1}^{k} f(\overline{x_k})\overline{c_k}.$$

   - Error analysis

6.7. Solving linear systems of equations
   - Direct methods: Gaussian elimination (computationally expensive)
   - Iterative methods: (faster and cheaper)
   - Solution stability

1.2 Round-off errors and computer arithmetics

There are three kinds of errors:

- Modeling Error: Occurs when we convert a problem from the real world into the mathematical world.
- Method Error: Occurs when we try to solve the mathematical problem numerically.
- Round-off error: Occurs when the computer gives an incorrect result with the correct algorithm (comes from storage and computation).

1.2.1 Storage

Infinite digit real numbers are *stored* as finite digit numbers, using the normalized decimal form of real numbers.

Definition. *Normalized decimal form of a real number*
For any $y \in \mathbb{R}$, we may write

$$y = \pm 0.d_1 d_2 d_3 \dots d_k d_{k+1} \dots \cdot 10^n,$$

where $0 < d_1 \leq 9$, $0 \leq d_i \leq 9$, n are integers. For the particular case where $y = 0$, we write $y = 0.0 \cdot 10^0$.

Definition. *Normalized machine numbers (Floating-point form)*
Any machine number y can be written as

$$y = \pm 0.d_1 d_2 \cdots d_k \cdot 10^n,$$

where $0 < d_1 \leq 9$, $0 \leq d_i \leq 9$, n are integers.

We can think of the storage process as mapping normalized real numbers to normalized machine numbers. We do this via rounding or truncating.

Consider some $y \in \mathbb{R} \setminus \{0\}$.

- Truncating (k-digit truncation of $y = \pm 0.d_1 d_2 d_3 \dots d_k d_{k+1} d_{k+2} \dots \cdot 10^n$) Simply omit the digits from d_{k+1} and onwards, in other words

  $$\text{fl}(\pm 0.d_1 d_2 d_3 \dots d_k d_{k+1} d_{k+2} \dots \cdot 10^n) = \pm 0.d_1 d_2 \dots d_k \cdot 10^n.$$

  Thus we have $y \approx \text{fl}(y)$.

- Rounding (k-digit rounding of $y = \pm 0.d_1 d_2 d_3 \dots d_k d_{k+1} d_{k+2} \dots \cdot 10^n$)

  If $d_{k+1} < 5$, then we drop $d_{k+1} d_{k+2} \dots$ (same with truncating)

  If $d_{k+1} \geq 5$, then add 1 to d_k and drop $d_{k+1} d_{k+2} \dots$

  $$\text{fl}(\pm 0.d_1 d_2 d_3 \dots d_k d_{k+1} d_{k+2} \dots \cdot 10^n) = \pm \delta_1 \delta_2 \dots \delta_k \cdot 10^m.$$

2 Lecture 2

Note. Since we use the notation fl to denote both k-digit truncation as well as k-digit rounding, be sure not to mix the two up.

Definition. *Errors*
Suppose p^* is an approximation of p. Then the *actual error* is $p - p^*$, the *absolute error* is $|p - p^*|$, and the relative error is $\frac{|p-p^*|}{|p|}$, where $p \neq 0$.

Definition. *Significant Digits*
The number p^* is said to approximate p to "t" significant digits if "t" is the largest non-negative integer for which

$$\frac{|p - p^*|}{|p|} \leq 5 \cdot 10^{-t}.$$

2.1 Computer Arithmetic

Assume that x, y are real numbers, then

$$\begin{aligned} x \oplus y &= \text{fl}(\text{fl}(x) + \text{fl}(y)) \\ x \ominus y &= \text{fl}(\text{fl}(x) - \text{fl}(y)) \\ x \otimes y &= \text{fl}(\text{fl}(x) \cdot \text{fl}(y)) \\ x \oslash y &= \text{fl}(\text{fl}(x) + \text{fl}(y)) \end{aligned}$$

3 Lecture 3

Note. When computing relative error, keep 2 non-zero digits.

If after performing our operation with k-digit chopping, we still have k significant digits, then our operation is pretty good (because it did not lose precision). In general, the $\oplus$ operator will lose at most one significant digit. Unlike addition, the $\ominus$ operator can lose multiple significant digits (when you subtract almost identical numbers).

We see errors introduced in computations:

- Subtraction of almost identical numbers results in loss of significant digits
- More operations $\to$ more errors $\to$ loss of significant digits

To avoid loss of *accuracy*, we can perform some manipulations to avoid scenarios that involve inaccuracies:

Example. *Quadratic formula*
For a given quadratic $ax^2 + bx + c = 0$, we know that a solution is

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}.$$

Suppose $b > 0$ and $b^2 \gg 4ac$. Then $b^2 - 4ac \approx b^2$, so $b \approx \sqrt{b^2 - 4ac}$. This can lead to inaccuracies because we are subtracting nearly identical numbers. To avoid this, we can perform some algebra as follows:

$$\begin{aligned} x_1 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \\ &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \cdot \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} \\ &= \frac{b^2 - (b^2 - 4ac)}{2a(-b - \sqrt{b^2 - 4ac})} \\ &= \frac{4ac}{2a(-b - \sqrt{b^2 - 4ac})} \\ &= \frac{2c}{-b - \sqrt{b^2 - 4ac}}. \end{aligned}$$

Since this solution avoids the subtraction of two nearly identical numbers, it is more accurate. For the other root, we are subtracting two negative numbers that are nearly identical, so there is no issue.

4 Lecture 4

4.1 Nested Algorithm

Goal. We want to reduce our error by reducing the number of operations done.

Example. Consider the polynomial

$$p(x) = 4x^4 + 5x^2 + 2x^2 + 5x - 10.$$

We need to perform 4 multiplications for the first term, 3 for the second, etc., for a total of 10 multiplication operations and 4 addition operations. Notice that we can factor out a common x from the first few terms, to get

$$p(x) = x \cdot (4x^3 + 5x^1 + 2x + 5) - 10.$$

Continuing this pattern we get

$$p(x) = x(x(x(4x + 5) + 2) + 5) - 10.$$

We now only have 4 multiplications and 4 additions, which can improve the accuracy of our algorithm.

4.2 Convergence Order

Definition. *Convergence Order*
Assume $\{P_n\}_{n=0}^{\infty}$ converges to P with $P_n \neq P$. If there exists $0 < \lambda < \infty$ and $\alpha > 0$ such that

$$\lim_{n\to\infty} \frac{|P_{n+1} - P|}{|P_n - P|^{\alpha}} = \lambda.$$

Then we say $\{P_n\}_{n=0}^{\infty}$ converges to P with order α.

Note. Larger α implies a faster convergence.

Definition. *Linear and Quadratic Convergence*
If $\alpha = 1$, then $\{P_n\}_{n=0}^{\infty}$ converges to P *linearly*. If $\alpha = 2$, then $\{P_n\}_{n=0}^{\infty}$ converges to P *quadratically*.

5 Lecture 5

5.1 Root Finding (Single Variable)

Given a function $f(x) = 0$ with a root $p \in [a, b]$, we want to find a sequence $p_1, p_2, \ldots$ such that $p_n \to p$. We want to find the root using the Intermediate Value Theorem. Suppose f is a continuous function on $[a, b]$ with $f(a) \cdot f(b) < 0$. Then f has at least one root in (a, b).

Note. This statement holds true because it means that $f(a)$ and $f(b)$ have opposite sign, or that 0 is between $f(a)$ and $f(b)$. Hence there must exist some $p \in (a, b)$ such that $f(p) = 0$.

5.1.1 Bisection Method

Idea. Use binary search to find the root.

You first find the midpoint of your interval (let's call this m), and check whether the sign of $f(m)$ is the same as $f(a)$ or $f(b)$. If the former, then we may "shrink" the interval to $[m, b]$, otherwise we may "shrink" the interval to $[a, m]$. If $f(m) = 0$, then we have found our root. If we perform this algorithm recursively, then we may find a sequence of points that converges to our root.

Note.

- The bisection method *always* converges to a root.
- We stop the iteration if:
  - We reach the maximum iteration number.
  - For some small $\varepsilon > 0$,

$$|p_n - p_{n-1}| < \varepsilon \text{ or } \frac{|p_n - p_{n-1}|}{|p_n|} < \varepsilon \text{ or } |f(p_n)| < \varepsilon.$$

Theorem. Suppose $f \in C[a, b]$ and $f(a) - f(b) < 0$. The Bisection method generates a sequence $\{p_n\}_{n=1}^{\infty}$ approximating a zero p of $f(x)$ with

$$|p_n - p| \leq \frac{b - a}{2^n}, \text{ where } n \geq 1.$$

Proof. For $n \geq 1$, we have $p_n = \frac{a_n + b_n}{2^n}$, $p \in (a_n, b_n)$. Then we know that

$$\begin{aligned} |p_n - p| &\leq \frac{1}{2}(b_n - a_n) \\ &= \frac{1}{2}\left(\frac{1}{2}(b_{n-1} - a_{n-1})\right) \\ &= \cdots \\ &= \underbrace{\frac{1}{2} \cdot \cdots \cdot \frac{1}{2}}_{n \text{ times}}(b_1 - a_1) \\ &= \frac{1}{2^n}(b - a). \end{aligned}$$

□

6 Lecture 6

6.1 Fixed Point Method (Finding Roots)

Definition. *Fixed Point*
We say p is a *fixed point* of $g(x)$ if $g(p) = p$.

- We first need to translate our root-finding problem into a fixed point problem.
- We want to solve $f(x) = 0$, so we should find p such that $f(p) = 0$.
- We rewrite $f(x) = 0$ as $g(x) \coloneqq f(x) + x = x$.
- Consider the relation between f and g:
  - Assume p is a root of $f(x)$, so $g(p) = f(p) + p = 0 + p = p$.
  - Thus if p is a root of f, then it must be a fixed point of g.

Note. The fixed point function $g(x)$ is *not* unique! In general, we have that $g(x) \coloneqq cf(x) + x$ is a fixed point function of f, given that $c \neq 0$.

How can we find the fixed point p of $g(x)$?

- Choose an initial approximation p_0 (of p).
- We have $p_1 = g(p_0), p_2 = g(p_1), \ldots$, so compute $p_n = g(p_{n-1})$ for $n \geq 1$.
- If $|p_n - p_{n-1}| < \varepsilon$ or $\frac{|p_n - p_{n-1}|}{|p_n|} < \varepsilon$, then stop.

Note. Some fixed point functions can diverge! You should choose a fixed point that converges.

Theorem — *Fixed Point Theorem*

(1) If $g \in C[a, b]$ and $g(x) \in [a, b]$ for all $x \in [a, b]$, then $g(x)$ has at least one fixed point in $[a, b]$.

(2) If, in addition, $g'(x)$ exists on (a, b) and there exists a positive constant $k < 1$ such that

$$|g'(x)| \leq k, \quad \text{for all } x \in (a, b),$$

then there is exactly one fixed point in $[a, b]$.

(1) *Proof.* If $g(a) = a$ or $g(b) = b$ then we are done. Hence we may assume that $g(a) \neq a, g(b) \neq b$. Let us define a function $h(x) \coloneqq g(x) - x$, which is continuous. Then we know that since $g(a) > a$, then $h(a) > 0$. Similarly, as $g(b) < b$, we have $h(b) < 0$. Hence by Intermediate Value Theorem there exists a point $c \in (a, b)$ such that $h(c) = 0$. In other words, there exists a point such that $g(c) - c = 0$, or $g(c) = c$. □

(2) *Proof.* We have that $|g'(x)| \leq k < 1$. Suppose towards a contradiction that there are two fixed points p, q of $g(x)$ on $[a, b]$ where $p \neq q$. Thus we have

$$\begin{aligned} |p - q| &= |g(p) - g(q)| \\ &= |p - q| \cdot \frac{g(p) - g(q)}{|p - q|} \\ &= |p - q| \cdot g'(c) && \text{(MVT, } c \in (a, b)\text{)} \end{aligned}$$

Dividing both sides by $|p - q|$, we have $g'(c) = 1$, a contradiction. Hence $g(x)$ has a *unique* fixed point on $[a, b]$. □

6.1.1 Convergence of Fixed Point Iteration

A fixed point of $g(x)$ is the point where $g(x)$ intersects the line $y = x$.

7 Lecture 7

If we consider the error generated by the fixed point iteration method, $e_n = |p - p_n|$, we have that this is equal to $|g(p) - g(p_{n-1})|$, where $g \in C[p, p_{n-1}]$. By Mean Value Theorem we know that there exists some $\xi \in (p, p_{n-1})$ such that

$$\begin{aligned} |g(p) - g(p_{n-1})| &= |g'(\xi)|\,|p - p_{n-1}| \\ &= |g'(\xi)|\,e_{n-1}. \end{aligned}$$

Hence if $e_n < e_{n-1}$ then $|g'(\xi)| < 1$ and the fixed point iteration converges to p. On the other hand, if $e_n \geq e_{n-1}$, then $|g'(\xi)| \geq 1$ and the fixed point iteration diverges.

Theorem — *Convergence of the Fixed Point Iteration*
Let $g(x) \in C[a, b]$ such that $g(x) \in [a, b]$ for all $x \in [a, b]$. Furthermore, suppose that $g'(x)$ exists on (a, b) and there exists $0 < k < 1$ such that

$$|g'(x)| \leq k \quad \text{for all } x \in (a, b).$$

Then for any $p_0 \in [a, b]$, the sequence defined by $p_n = g(p_n - 1)$ for all $n \geq 1$ converges to the unique fixed-point p in $[a, b]$.

Proof. Let $p_0 \in [a, b]$. Then

$$\begin{aligned} |p - p_n| &= |g(p) - g(p_{n-1})| \\ &= |g'(\xi)(p - p_{n-1})| && (\xi \in (p, p_{n-1}) \subset [a, b]) \\ &= |g'(\xi)|\,|p - p_{n-1}| \\ &\leq k\,|p - p_{n-1}| \\ &\leq k^2\,|p - p_{n-2}| \\ &\leq \cdots \\ &\leq k^n\,|p - p_0|\,. \end{aligned}$$

Hence

$$\begin{aligned} \lim_{n\to\infty} |p - p_n| &\leq \lim_{n\to\infty} k^n\,|p - p_0| \\ &= |p - p_0| \lim_{n\to\infty} k^n \\ &= 0. \end{aligned}$$

Thus $\lim_{n\to\infty} |p - p_n| = 0$, so $p_n \to p$ for any $p_0 \in [a, b]$. □

Error Analysis. We want to bound $|p - p_n|$ by some known values. Since $p_0 \in [a, b]$ and $|p - p_n| \leq k^n\,|p - p_0|$, we have

$$|p - p_n| \leq k^n \max\{|a - p_0|\,, |b - p_0|\}\,.$$

We can then use this equation to find the number of iterations required to get to a certain error bound. If we want to have some error bound of ε, then

$$\begin{aligned} k^n \max\{|a - p_0|\,, |b - p_0|\} &\leq \varepsilon \\ k^n &\leq \frac{\varepsilon}{\max\{|a - p_0|\,, |b - p_0|\}} \\ n \log k &\leq \log \frac{\varepsilon}{\max\{|a - p_0|\,, |b - p_0|\}} \\ n &\geq \frac{1}{\log k} \cdot \log \frac{\varepsilon}{\max\{|a - p_0|\,, |b - p_0|\}}. && (\log k < 0) \end{aligned}$$

8 Lecture 8

8.1 Newton's Method

- Choose an initial approximation p_0 of p.
- Set p_n to be the x-intercept of the tangent line passing through $(p_{n-1}, f(p_{n-1}))$. If we wish to solve this equation we have:

$$y - f(p_{n-1}) = f'(p_{n-1}) \cdot (x - p_{n-1})$$
$$x = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}.$$

  We can only use this for $f'(p_{n-1}) \neq 0$, since if it were then the tangent line would have no x-intercept. We can treat this method as a form of a fixed point method, with $g(x) = x - \frac{f(x)}{f'(x)}$.

Note. Newton's method converges faster than the regular fixed point method when $p_0 \approx p$.

Theorem — *Convergence of Newton's Method*
Let $f(x) \in C^2[a, b]$. If $p \in (a, b)$ such that $f(p) = 0$ and $f'(p) \neq 0$, then there exists a $\delta > 0$ such that the Newton's method generates a sequence $\{p_n\}_{n=1}^{\infty}$ which converges to p for any initial approximation $p_0 \in [p - \delta, p + \delta]$.

Proof. Consider Newton's method to be a fixed-point method with

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

We wish to show:

(i) $g(x)$ exists on $(p - \delta, p + \delta)$.

(ii) $g(x)$ exists on $(p - \delta, p + \delta)$ and there exists $0 < k < 1$ such that $|g'(x) \leq k|$.

(iii) $g(x) \in [p - \delta, p + \delta]$ for $x \in [p - \delta, p + \delta]$.

Since $f(x) \in C^2[a, b]$, we know that f, f', and f'' are continuous on $[a, b]$. Since $f'(p) \neq 0$ and f' continuous, there exists $\delta_1 > 0$ such that $f'(x) \neq 0$ for all $x \in [p - \delta_1, p + \delta_1]$. Hence $g(x)$ is continuous on $[p - \delta_1, p + \delta_1]$. Observe that

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x) \cdot f''(x)}{(f'(x))^2}.$$

Since f, f', and f'' are continuous on $[p - \delta_1, p + \delta_1]$ and $f'(x) \neq 0$ on $[p - \delta_1, p + \delta_1]$, we have that $g'(x)$ is continuous on $[p - \delta_1, p + \delta_1]$. Hence

$$g'(p) = \frac{f(p)f''(p)}{(f'(p))^2} = 0,$$

so there must be some small interval $0 < \delta < \delta_1$ such that $|g'(x)| \leq k$ where $0 < k < 1$.

We must finally show that $g(x) \in [p - \delta, p + \delta]$ when $x \in [p - \delta, p + \delta]$. Let $x \in [p - \delta, p + \delta]$, so

$$\begin{aligned} |g(x) - p| &= |g(x) - g(p)| \\ &= |g'(\xi) \cdot (x - p)| && \text{(MVT)} \\ &= |g'(\xi)|\,|x - p| \\ &< |x - p| \\ &\leq \delta. \end{aligned}$$

Thus $g(x) \in (p - \delta, p + \delta)$.
Therefore there exists some $\delta > 0$ such that the above three statements hold, and so by the convergence of a fixed point iteration we have that any $p_0 \in [p - \delta, p + \delta]$ generates a sequence that converges to p. □

Note. This theorem only tells us the *existence* of such a δ, and doesn't tell us anything about δ. In practice, we usually run a few bisection iterations to get closer to the actual root (since it will converge towards the root), and then switch over to Newton's method.