Numerical Methods

Aaron Naiman
Jerusalem College of Technology
naiman@math.jct.ac.il
http://math.jct.ac.il/~naiman

based on: Numerical Mathematics and Computing
by Cheney & Kincaid, ©1994
Brooks/Cole Publishing Company
ISBN 0-534-20112-1

Taylor Series

$\Rightarrow$ Definitions and Theorems

- Examples
- Proximity of x to c
- Additional Notes

Motivation

- Sought: $\cos(0.1)$
- Missing: calculator or lookup table
- Known: cos for another (nearby) value, i.e., at 0
- Also known: lots of (all) derivatives at 0
- Can we use them to *approximate* $\cos(0.1)$?
- What will be the worst error of our approximation?

These techniques are used by computers, calculators, tables.

Taylor Series

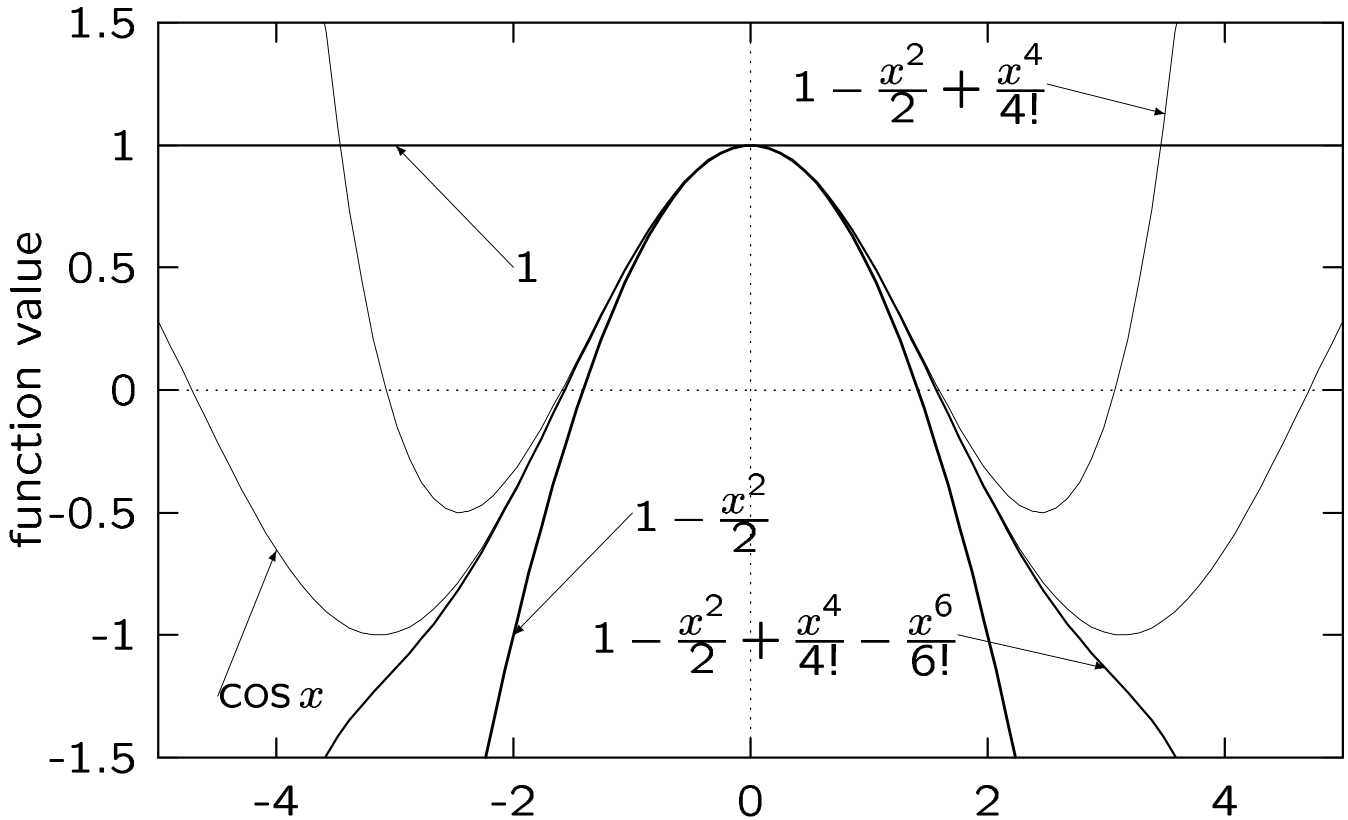
- Series definition: If $\exists f^{(k)}(c)$, $k = 0, 1, 2, \ldots$, then:

$$\begin{aligned} f(x) &\approx f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots \\ &= \sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!}(x-c)^k \end{aligned}$$

- c is a *constant* and much is known about it ($f^{(k)}(c)$)
- x a variable near c, and $f(x)$ is sought
- With $c = 0 \Rightarrow$ Maclaurin series
- What is the maximum error if we stop after n terms?
- Real life: crowd estimation: 100K ±10K vs. 100K ±1K

Key NM questions: What is estimate? What is its max error?

Taylor Series — $\cos x$



Better and better approximation, near c, and away.

Taylor's Theorem

- Theorem: If $f \in C^{n+1}[a,b]$ then

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-c)^{n+1},$$

where

$$x, c \in [a,b], \;\; \xi(x) \in \text{ open interval between } x \text{ and } c$$

- Notes:
  - * $f \in C(X)$ means f is continuous on X
  - * $f \in C^k(X)$ means $f, f', f'', f^{(3)}, \ldots, f^{(k)}$ are continuous on X
  - * $\xi = \xi(x)$, i.e., a point whose position is a function of x
  - * Error term is just like other terms, with $k := n+1$

ξ-term is "truncation error", due to series termination

Taylor Series—Procedure

- Writing it out, step-by-step:
  - ∗ write formula for $f^{(k)}(x)$
  - ∗ choose c (if not already specified)
  - ∗ write out summation and error term
    - ⋆ note: sometimes easier to write out a few terms

- Things to (possibly) prove — by analyzing worst case ξ
  - ∗ letting $n \to \infty$
    - ⋆ LHS remains $f(x)$
    - ⋆ summation becomes infinite Taylor series
    - ⋆ if error term $\to 0 \Rightarrow$

      infinite Taylor series represents $f(x)$
  - ∗ for given n, we can estimate max of error term

Taylor Series

- Definitions and Theorems

$\Rightarrow$ Examples

- Proximity of x to c
- Additional Notes

Taylor Series: e^x

- $f(x) = e^x,\ |x| < \infty \therefore f^{(k)}(x) = e^x, \forall k$
- Choose $c := 0$
- We have

$$e^x = \sum_{k=0}^{n} \frac{x^k}{k!} + \frac{e^{\xi(x)}}{(n+1)!} x^{n+1}$$

- As $n \to \infty$ — take worst case ξ (just less than x)

  error term $\to 0$ (why?) $\therefore$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Taylor Series: $\sin x$

- $f(x) = \sin x,\ |x| < \infty \therefore f^{(k)}(x) = \sin\left(x + \frac{\pi k}{2}\right), \forall k,\ c := 0$

- We have

$$\sin x = \sum_{k=0}^{n} \frac{\sin\left(\frac{\pi k}{2}\right)}{k!} x^k + \frac{\sin\left(\xi(x) + \frac{\pi(n+1)}{2}\right)}{(n+1)!} x^{n+1}$$

- Error term $\to 0$ as $n \to \infty$

- Even k terms are zero $\therefore \ell = 0, 1, 2, \ldots$, and $k \to 2\ell + 1$

$$\sin x = \sum_{\ell=0}^{\infty} \frac{\sin\left(\frac{\pi(2\ell+1)}{2}\right)}{(2\ell+1)!} x^{2\ell+1} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

Taylor Series: $\cos x$

- $f(x) = \cos x$, $|x| < \infty \therefore f^{(k)}(x) = \cos\left(x + \frac{\pi k}{2}\right), \forall k$, $c := 0$

- We have

$$\cos x = \sum_{k=0}^{n} \frac{\cos\left(\frac{\pi k}{2}\right)}{k!} x^k + \frac{\cos\left(\xi(x) + \frac{\pi(n+1)}{2}\right)}{(n+1)!} x^{n+1}$$

- Error term $\to 0$ as $n \to \infty$

- Odd k terms are zero $\therefore \ell = 0, 1, 2, \ldots$, and $k \to 2\ell$

$$\cos x = \sum_{\ell=0}^{\infty} \frac{\cos\left(\frac{\pi(2\ell)}{2}\right)}{(2\ell)!} x^{2\ell} = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$$

Numerical Example: $\cos(0.1)$

- We have [1] $f(x) = \cos x$ and [2] $c = 0$
  - $*$ obtain series: $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$

- Actual value: $\cos(0.1) = 0.99500416527803\ldots$

- With [3] $x = 0.1$ and [4] specific n's
  - $*$ from Taylor approximations:

n^*	approximation	$\lvert\text{error}\rvert \leq$
0, 1	1	0.01/2!
2, 3	0.995	0.0001/4!
4, 5	$0.9950041\overline{6}$	0.000001/6!
6, 7	0.99500416527778	0.00000001/8!
$\vdots$	$\vdots$	$\vdots$

*includes odd k

Obtain accurate approximation easily and quickly.

Taylor Series: $(1-x)^{-1}$

- $f(x) = \frac{1}{1-x}$, $|x| < 1 \therefore f^{(k)}(x) = \frac{k!}{(1-x)^{k+1}}, \forall k$, choose $c := 0$
- We have

$$\begin{aligned} \frac{1}{1-x} &= \sum_{k=0}^{n} x^k + \frac{(n+1)!}{(1-\xi(x))^{n+2}} \cdot \frac{x^{n+1}}{(n+1)!} \\ &= \sum_{k=0}^{n} x^k + \left(\frac{x}{1-\xi(x)}\right)^{n+1} \frac{1}{1-\xi(x)} \end{aligned}$$

- Why bother, with LHS so simple? Ideas?
- Sufficient: $\left|\frac{x}{1-\xi(x)}\right|^{n+1} \to 0$ as $n \to \infty$
- For what range of x is this satisfied?

Need to determine radius of convergence.

$(1-x)^{-1}$ — Range of Convergence

- Sufficient: $\left|\frac{x}{1-\xi(x)}\right| < 1$
- Approach:
  - get variable x in middle of sufficiency inequality
  - transform range of ξ inequality to LHS and RHS of sufficiency inequality
  - require restriction on x
    - but check if already satisfied
- $|\xi| < 1 \Rightarrow 1-\xi > 0 \Rightarrow$ sufficient: $-(1-\xi) < x < 1-\xi$

$(1-x)^{-1}$ — Range of Convergence (cont.)

- case $x < \xi < 0$:
  - LHS: $-(1-x) < -(1-\xi) < -1 \Rightarrow$ require: $-1 \leq x$ $\surd$
  - RHS: $1 < 1-\xi < 1-x \Rightarrow$ require: $x \leq 1$ $\surd$

- case $0 < \xi < x$:
  - LHS: $-1 < -(1-\xi) < -(1-x) \Rightarrow$ require: $-(1-x) \leq x$, or: $-1 < 0$ $\surd$
  - RHS: $1-x < 1-\xi < 1 \Rightarrow$ require: $x \leq 1-x$, or: $x \leq \frac{1}{2}$

- Therefore, for $-1 < x \leq \frac{1}{2}$

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \cdots \quad \Big(\text{Zeno: } x = \tfrac{1}{2}, \ldots\Big)$$

Need more analysis for the whole range $\lvert x \rvert < 1$.

Taylor Series: $\ln x$

- $f(x) = \ln x$, $0 < x \leq 2 \mathrel{.\,\cdot\,.} f^{(k)}(x) = (-1)^{k-1}\frac{(k-1)!}{x^k}, \forall k \geq 1$
- Choose $c := 1$
- We have

$$\ln x = \sum_{k=1}^{n} (-1)^{k-1}\frac{(x-1)^k}{k} + (-1)^n \frac{1}{n+1}\frac{(x-1)^{n+1}}{\xi^{n+1}(x)}$$

- Sufficient $\left|\frac{x-1}{\xi(x)}\right|^{n+1} \to 0$ as $n \to \infty$
- Again, for what range of x is this satisfied?

$\ln x$ — Range of Convergence

- Sufficient: $\left|\frac{x-1}{\xi(x)}\right| < 1 \ \ldots \ 1-\xi < x < 1+\xi$

- case $1 < \xi < x$:
  - ∗ LHS: $1 - x < 1 - \xi < 0 \Rightarrow$ require: $0 \le x \ \surd$
  - ∗ RHS: $2 < 1 + \xi < 1 + x \Rightarrow$ require: $x \le 2 \ \surd$

- case $x < \xi < 1$:
  - ∗ LHS: $0 < 1 - \xi < 1 - x \Rightarrow$ require: $1 - x \le x$, or: $\frac{1}{2} \le x$
  - ∗ RHS: $1 + x < 1 + \xi < 2 \Rightarrow$ require: $x \le 1 + x \ \surd$

- Therefore, for $\frac{1}{2} \le x \le 2$

$$\ln x = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{(x-1)^k}{k} = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \cdots$$

Again, need more analysis for entire range of x.

Ratio Test and ln x Revisited

- Theorem: $\left|\frac{a_{n+1}}{a_n}\right| \to (< 1) \Rightarrow$ partial sums converge
- ln x: ratio of adjacent summand terms (*not* the error term)

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|(x-1)\frac{n}{n+1}\right|$$

- Obtain convergence of partial sums for $0 < x < 2$
- Note: not looking at ξ and the error term
- $x = 2$: $1 - \frac{1}{2} + \frac{1}{3} - \cdots$, which is convergent (why?)
- $x = 0$: same series, all same sign $\Rightarrow$ divergent harmonic series
- $\therefore$ we have $0 < x \leq 2$

$(1-x)^{-1}$ Revisited

- Letting $x \to (1-x)$

$$\ln(1-x) = -\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right), \quad -1 \leq x < 1$$

- $\frac{d}{dx}$: lhs $= \frac{-1}{1-x}$ and rhs $= -\left(1 + x + x^2 + x^3 + \cdots\right)$

- ⚠: no "=" for $x = -1$ as rhs oscillates (note: correct avg value)

- $|x| < 1$ we have (also with ratio test)

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots$$

Taylor Series

- Definitions and Theorems
- Examples

$\Rightarrow$ Proximity of x to c

- Additional Notes

Proximity of x to c

Problem: Approximate $\ln 2$

- Solution 1: Taylor $\ln(1+x)$ around 0 with $x = 1$

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots$$

- Solution 2: Taylor $\ln\left(\frac{1+x}{1-x}\right)$ around 0 with $x = \frac{1}{3}$

$$\ln 2 = 2\left(3^{-1} + \frac{3^{-3}}{3} + \frac{3^{-5}}{5} + \frac{3^{-7}}{7} + \cdots\right)$$

Proximity of x to c (cont.)

- Approximated values, rounded:
  - Solution 1, first 8 terms: 0.63452
  - Solution 2, first 4 terms: 0.69313
- Actual value, rounded: 0.69315
- $\therefore$ importance of proximity of evaluation and expansion points

This error is in *addition* to the truncation error.

Taylor Series

- Definitions and Theorems
- Examples
- Proximity of x to c

$\Rightarrow$ Additional Notes

Polynomials and a Second Form

- Polynomials $\in C^{\infty}(-\infty, \infty)$
  - $*$ have finite number of non-zero derivatives, $\therefore$
  - $*$ Taylor *series* $\forall c$ ... original polynomial, i.e., error = 0

$$f(x) = 3x^2 - 1, \ \ldots \ f(x) = \sum_{k=0}^{2} \frac{f^{(k)}(0)}{k!} x^k = -1 + 0 + 3x^2$$

  - $*$ Taylor *Theorem* can be used for fewer terms
    - $\star$ e.g.: approximate a P_{17} near c by a P_3

- Taylor's Theorem, second form (x = *constant* expansion point, h = distance, $x + h$ = *variable* evaluation point): If $f \in C^{n+1}[a, b]$ then

$$f(x+h) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!} h^k + \frac{f^{(n+1)}(\xi(h))}{(n+1)!} h^{n+1},$$

$$x, x + h \in [a, b], \ \ \xi(h) \in \text{ open interval between } x \text{ and } x + h$$

Taylor Approximate: $(1-3h)^{\frac{4}{5}}$

- Define: $f(z) \equiv z^{\frac{4}{5}}$; $x = 1$ is the constant expansion point
- Derivs: $f'(z) = \frac{4}{5} z^{-\frac{1}{5}}$, $f''(z) = -\frac{4}{5^2} z^{-\frac{6}{5}}$, $f'''(z) = \frac{24}{5^3} z^{-\frac{11}{5}}$, ...
- $\therefore$:

$$(x+h)^{\frac{4}{5}} = x^{\frac{4}{5}} + \frac{4}{5} x^{-\frac{1}{5}} h - \frac{4}{2! \cdot 5^2} x^{-\frac{6}{5}} h^2 + \frac{24}{3! \cdot 5^3} x^{-\frac{11}{5}} h^3 + \ldots$$

$$(x-3h)^{\frac{4}{5}} = x^{\frac{4}{5}} - \frac{4}{5} x^{-\frac{1}{5}} 3h - \frac{4}{2! \cdot 5^2} x^{-\frac{6}{5}} 9h^2 - \frac{24}{3! \cdot 5^3} x^{-\frac{11}{5}} 27h^3 + \ldots$$

$$\begin{aligned} (1-3h)^{\frac{4}{5}} &= 1 - \frac{4}{5} 3h - \frac{4}{2! \cdot 5^2} 9h^2 - \frac{24}{3! \cdot 5^3} 27h^3 + \ldots \\ &= 1 - \frac{12}{5} h - \frac{18}{25} h^2 - \frac{108}{125} h^3 + \ldots \end{aligned}$$

Second Form — $\ln(e+h)$

- Evaluation of interest: $\ln(e+h)$
- Define: $f(z) \equiv \ln(z)$
- $x = e$ is the constant expansion point
- $\ln \Rightarrow z > 0$
- Derivatives

$$
\begin{array}{ll}
f(z) = \ln z & f(e) = 1 \\
f'(z) = z^{-1} & f'(e) = e^{-1} \\
f''(z) = -z^{-2} & f''(e) = -e^{-2} \\
f'''(z) = 2z^{-3} & f'''(e) = 2e^{-3} \\
f^{(n)}(z) = (-1)^{n-1}(n-1)!z^{-n} & f^{(n)}(e) = (-1)^{n-1}(n-1)!\, e^{-n}
\end{array}
$$

$\ln(e+h)$ — Expansion and Convergence

- Expansion (recall: $x = e$)

$$\ln(e+h) \equiv f(x+h) \;=\; 1 + \sum_{k=1}^{n} \frac{(-1)^{k-1}(k-1)!e^{-k}h^k}{k!} + \frac{(-1)^n\, n!\, \xi(h)^{-(n+1)} h^{n+1}}{(n+1)!}$$

or

$$\ln(e+h) = 1 + \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}\left(\frac{h}{e}\right)^k + \frac{(-1)^n}{n+1}\left(\frac{h}{\xi(h)}\right)^{n+1}$$

- Range of convergence, sufficient (for variable h): $-\xi < h < \xi$
  - case $e+h<\xi<e$: . . . $-\frac{e}{2} \leq h$
  - case $e<\xi<e+h$: . . . $h \leq e$

$O()$ Notation and MVT

- As $h \to 0$, we write the speed of $f(h) \to 0$

$$f(h) = O\left(h^k\right) \quad \equiv \quad |f(h)| \le C|h|^k$$

e.g., $f(h)$: h, $\frac{1}{1000}h$, h^2; let $h \to \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}, \ldots$

- Taylor truncation error $= O\left(h^{n+1}\right)$; if for a given n the max exists, then

$$C := \left|\max_{\xi(h)} f^{(n+1)}(\xi(h))\right| / (n+1)!$$

- Mean value theorem (Taylor, $n = 0$): If $f \in C^1[a, b]$ then

$$f(b) = f(a) + (b - a)f'(\xi), \;\; \xi \in (a, b)$$

or:

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}$$

Alternating Series Theorem

- Alternating series theorem: If $a_k > 0$, $a_k \geq a_{k+1}$, $\forall k \geq 0$, and $a_k \to 0$, then

$$\sum_{k=0}^{n} (-1)^k a_k \to S \text{ and } |S - S_n| \leq a_{n+1}$$

- Intuitively understood
- Note: *direction* of error is also know for specific n
- We had this with sin and cos
- Another useful method for max truncation error estimation

Max truncation error estimation without ξ-analysis

$\ln(e+h)$ — Max Trunc. Error Estimate

- What is the max error after $n+1$ terms?
- Max error estimate *also* depends on proximity—size of h
  - ∗ from Taylor: obtain $O\left(h^{n+1}\right)$

$$|\text{error}| \le \frac{1}{n+1}|h|^{n+1} \max_{\xi}\left|\frac{1}{\xi}\right|^{n+1}$$

  - ∗ from AST (check the conditions!): also obtain $O\left(h^{n+1}\right)$, with different constant

$$|\text{error}| \le \frac{1}{n+1}\left|\frac{h}{e}\right|^{n+1}$$

- E.g.: $h = -\frac{e}{2}$: $\ln\frac{e}{2} = 1 - \frac{1}{2} - \frac{1}{2}\cdot\frac{1}{2^2} - \frac{1}{3}\cdot\frac{1}{2^3} - \frac{1}{4}\cdot\frac{1}{2^4} - \cdots$
  - ∗ Taylor max error (occurs as $\xi \to \frac{e}{2}^+$): $\frac{1}{n+1}$
  - ∗ AST max error: $\frac{1}{n+1}\cdot\frac{1}{2^{n+1}}$
  - ∗ note the huge difference in *max error* estimate

Base Representations

- $\Rightarrow$ Definitions
- Conversions
- Computer Representation
- Loss of Significant Digits

Number Representation

- Simple representation in one base $\not\Rightarrow$ simple representation in another base, e.g.

$$(0.1)_{10} = (0.0\ 0011\ 0011\ 0011\ \ldots)_2$$

- Base 10:

$$\begin{aligned} 37294 &= 4 + 90 + 200 + 7000 + 30000 \\ &= 4 \times 10^0 + 9 \times 10^1 + 2 \times 10^2 + 7 \times 10^3 + 3 \times 10^4 \end{aligned}$$

in general: $a_n \ldots a_0 = \sum_{k=0}^{n} a_k 10^k$

Fractions and Irrationals

- Base 10 fraction:

$$0.7217 = 7 \times 10^{-1} + 2 \times 10^{-2} + 1 \times 10^{-3} + 7 \times 10^{-4}$$

- In general, for real numbers:

$$a_n \ldots a_0.b_1 \ldots = \sum_{k=0}^{n} a_k 10^k + \sum_{k=1}^{\infty} b_k 10^{-k}$$

- Note: $\exists$ numbers, i.e., irrationals, such that an infinite number of digits are required, in *any* rational base, e.g., $e, \pi, \sqrt{2}$

- Need infinite number of digits in *a* base $\not\Rightarrow$ irrational

$$(0.333\ldots)_{10} \text{ but } \frac{1}{3} \text{ is not irrational}$$

Other Bases

- Base 8, $\nexists$ '8' or '9', using octal digits

$$(21467)_8 = \cdots = (9015)_{10}$$

$$(0.36207)_8 = 8^{-5}\Big(3 \times 8^4 + \cdots\Big) = \frac{15495}{32768} = (0.47286\ \ldots)_{10}$$

- Base 16: '0', '1', . . . , '9', 'A' (10), 'B' (11), 'C' (12), 'D' (13), 'E' (14), 'F' (15)

- Base β

$$(a_n \ldots a_0.b_1 \ldots)_\beta = \sum_{k=0}^{n} a_k \beta^k + \sum_{k=1}^{\infty} b_k \beta^{-k}$$

- Base 2: just '0' and '1', or for computers: "off" and "on", "bit" = binary digit

Base Representations

- Definitions

$\Rightarrow$ Conversions

- Computer Representation
- Loss of Significant Digits

Conversion: Base 10 $\rightarrow$ Base 2

- Basic idea:

$$\begin{aligned} 3781 &= 1 + \underbrace{10}_{(1010)_2}\left(\underbrace{8}_{(1000)_2} + 10(7 + 10(3))\right) = \cdots \\ &= (111\ 011\ 000\ 101)_2 \end{aligned}$$

- Easy for computer, but by hand: $(3781.372)_{10}$

$$\begin{array}{rll} & \text{remainder} & \\ 2\overline{)3781} & & \\ 2\overline{)1890} & \boxed{1} = a_0 & \downarrow \\ 2\overline{)945} & \boxed{0} = a_1 & \\ \vdots & & \end{array} \qquad \begin{array}{rrl} & 0.372 & \\ & \underline{\quad 2} & \\ \downarrow\ b_1 = & \boxed{0}.744 & \\ & \underline{\quad 2} & \\ b_2 = & \boxed{1}.488 & (\text{drop } \boxed{1}) \\ & \vdots & \end{array}$$

Base 8 Shortcut

- Base 2 $\leftrightarrow$ base 8, trivial

$$(551.624)_8 = (101\ \ 101\ \ 001.110\ \ 010\ \ 100)_2$$

- $\approx$ 3 bits for every 1 octal digit
- One digit produced for every step in (hand) conversion
- $\therefore$ base 10 $\rightarrow$ base 8 $\rightarrow$ base 2

Base Representations

- Definitions
- Conversions

⇒ Computer Representation

- Loss of Significant Digits

Computer Representation

- Scientific notation:

$$32.213 \to 0.32213 \times 10^2$$

- In general

$$x = \pm 0.d_1 d_2 \ldots \times 10^n, \;\; d_1 \neq 0, \text{ or: } x = \pm r \times 10^n, \; \frac{1}{10} \leq r < 1$$

we have sign, “mantissa” r and “exponent” n

- On the computer, base 2 is represented

$$x = \pm 0.b_1 b_2 \ldots \times 2^n, \;\; b_1 \neq 0, \text{ or: } x = \pm r \times 2^n, \; \frac{1}{2} \leq r < 1$$

- Finite number of mantissa digits, therefore “roundoff” or “truncation” error

Base Representations

- Definitions
- Conversions
- Computer Representation

$\Rightarrow$ Loss of Significant Digits

LSD—Addition

- $(a+b)+c = a+(b+c)$ on the computer?

- Six decimal digits for mantissa

$$1,000,000. + \underbrace{1. + \cdots + 1.}_{\text{million times}} = 1,000,000.$$

because

$$0.100000 \times 10^7 + 0.100000 \times 10^1 = 0.100000 \times 10^7$$

but

$$\underbrace{1. + \cdots + 1.}_{\text{million times}} + 1,000,000. = 2,000,000.$$

Add numbers in size order.

LSD—Subtraction

- E.g.: $x - \sin x$ for x's close to zero

$$\begin{aligned} x &= \frac{1}{15} \text{ (radians)} \\ x &= 0.66666\ \ 66667 \times 10^{-1} \\ \sin x &= 0.66617\ \ 29492 \times 10^{-1} \\ x - \sin x &= 0.00049\ \ 37175 \times 10^{-1} \\ &= 0.49371\ \ 75000 \times 10^{-4} \end{aligned}$$

- Note
  - ∗ still have 10^{-10} precision, but
  - ∗ can we retain 3 "lost" digits for 10^{-13} precision?

Avoid subtraction of close numbers.

LSD Avoidance for Subtraction

- $x - \sin x$ for $x \approx 0 \rightarrow$ use Taylor series
  - no subtraction of *close* numbers
  - e.g., 3 terms: $0.49371\ 74328 \times 10^{-4}$
    actual: $0.49371\ 74327 \times 10^{-4}$

- $e^x - e^{-2x}$ for $x \approx 0 \rightarrow$ use Taylor series twice and add common powers

- $\sqrt{x^2 + 1} - 1$ for $x \approx 0 \rightarrow \frac{x^2}{\sqrt{x^2+1}+1}$

- $\cos^2 x - \sin^2 x$ for $x \approx \frac{\pi}{4} \rightarrow \cos 2x$

- $\ln x - 1$ for $x \approx e \rightarrow \ln \frac{x}{e}$

Nonlinear Equations

- $\Rightarrow$ Motivation
- Bisection Method
- Newton's Method
- Secant Method
- Summary

Motivation

- For a given function $f(x)$, find its root(s), i.e.:

  $\Rightarrow$ find x (or r = root) such that $f(x) = 0$

- BVP: dipping of suspended power cable. What is λ?

$$\lambda \cosh \frac{50}{\lambda} - \lambda - 10 = 0$$

- (Some) simple equations $\Rightarrow$ solve analytically

$$\begin{aligned} 6x^2 - 7x + 2 &= 0 \\ (3x-2)(2x-1) &= 0 \\ x &= \frac{2}{3},\ \frac{1}{2} \end{aligned}$$

$$\begin{aligned} \cos 3x - \cos 7x &= 0 \\ 2 \sin 5x \sin 2x &= 0 \\ x &= \frac{n\pi}{5}, \frac{n\pi}{2},\ \ n \in \mathbb{Z} \end{aligned}$$

Motivation (cont.)

- In general, we cannot exploit the function, e.g.:

$$2^{x^2} - 10x + 1 = 0$$

  and

$$\cosh\left(\sqrt{x^2+1} - e^x\right) + \log|\sin x| = 0$$

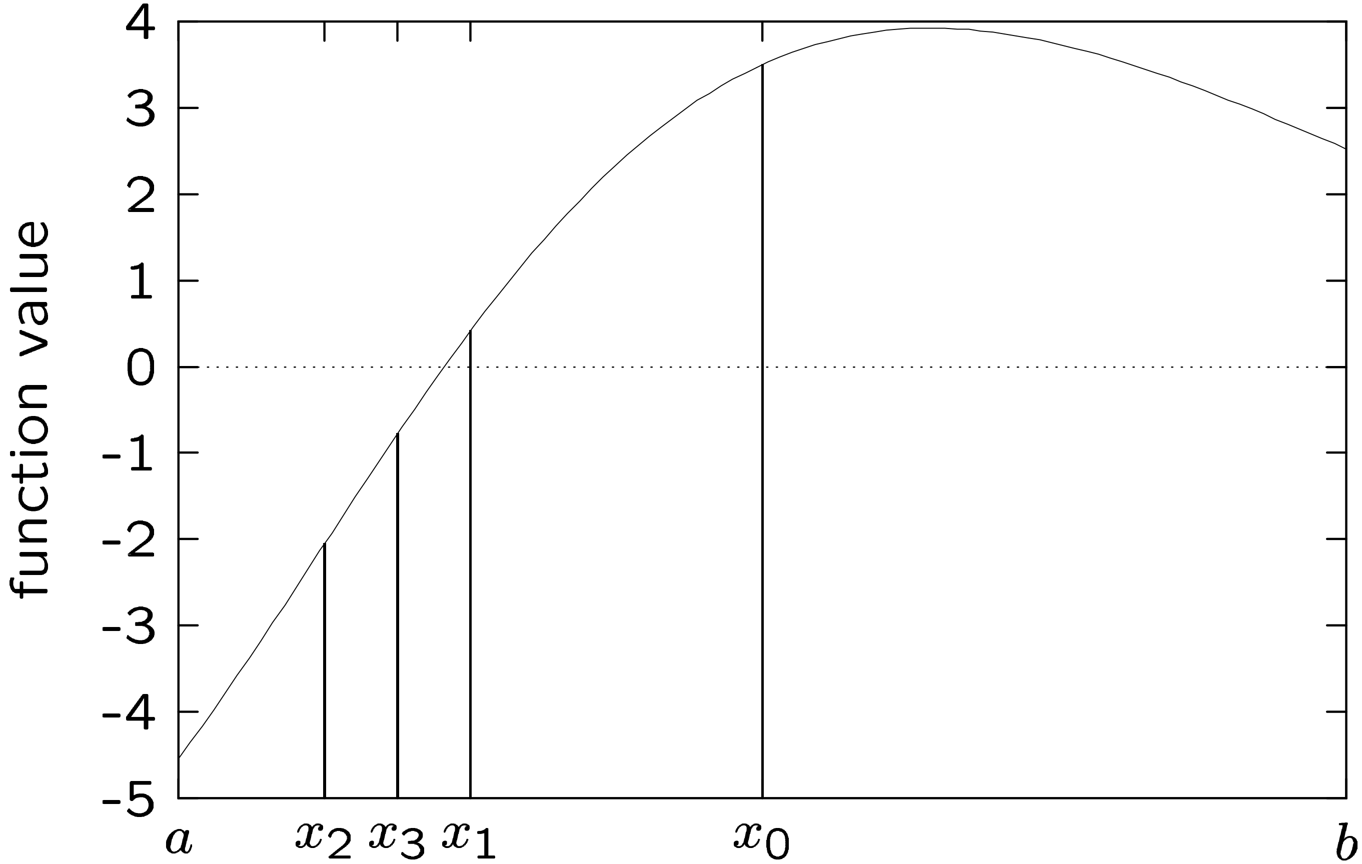
- Note: at times $\exists$ multiple roots
  - ∗ e.g., previous parabola and cosine
  - ∗ we want at least one
  - ∗ we may only get one (for each search)

Need a general, function-independent algorithm.

Nonlinear Equations

- Motivation
- ⇒ Bisection Method
- Newton's Method
- Secant Method
- Summary

Bisection Method—Example



Intuitive, like guessing a number $\in [0, 100]$.

Restrictions and Max Error Estimate

- Restrictions
  - $*$ function slices x-axis at root
    - $\star$ start with two points a and $b \ni f(a)f(b) < 0$
    - $\star$ graphing tool (e.g., Matlab) can help to find a and b
  - $*$ require $C^0[a, b]$ (why? note: not a big deal)

- Max error estimate
  - $*$ after n steps, guess midpoint of current range
  - $*$ error: $\epsilon \leq \frac{b-a}{2^{n+1}}$ (think of $n = 0, 1, 2$)
  - $*$ note: error is in x; can also look at error in $f(x)$ or combination
    - $\star$ enters entire world of stopping criteria

Question: Given tolerance (in x), what is n? . . .

Convergence Rate

- Given tolerance τ (e.g., 10^{-6}), how many steps are needed?
- Tolerance restriction (ϵ from before):

$$\left(\epsilon \le \frac{b-a}{2^{n+1}}\right) < \tau$$

- $\therefore$ 1) $\times$ 2, 2) log (any base)

$$\log(b-a) - n\log 2 < \log 2\tau$$

or

$$n > \frac{\log(b-a) - \log 2\tau}{\log 2}$$

Rate is independent of function.

Convergence Rate (cont.)

- Base 2 (i.e., bits of accuracy)

$$n > \log_2 (b - a) - 1 - \log_2 \tau$$

  i.e., number of steps is a constant plus one step per bit

- Linear convergence rate: $\exists C \in [0, 1)$

$$\left|x_{n+1} - r\right| \le C|x_n - r|, \quad n \ge 0$$

  i.e., monotonic decreasing error at *every* step, and

$$\left|x_{n+1} - r\right| \le C^{n+1}|x_0 - r|$$

- Bisection convergence
  - * *not* linear (examples?), but compared to init. *max* error:
  - * similar form: $\left|x_{n+1} - r\right| \le C^{n+1}(b - a)$, with $C = \frac{1}{2}$

Okay, but restrictive and slow.

Nonlinear Equations

- Motivation
- Bisection Method

$\Rightarrow$ Newton's Method

- Secant Method
- Summary

Newton's Method—Definition

- Approximate $f(x)$ near x_0 by tangent $\ell(x)$

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) \equiv \ell(x)$$

Want $\ell(r) = 0 \Rightarrow r = x_0 - \frac{f(x_0)}{f'(x_0)}$, $\therefore x_1 := r$, likewise:

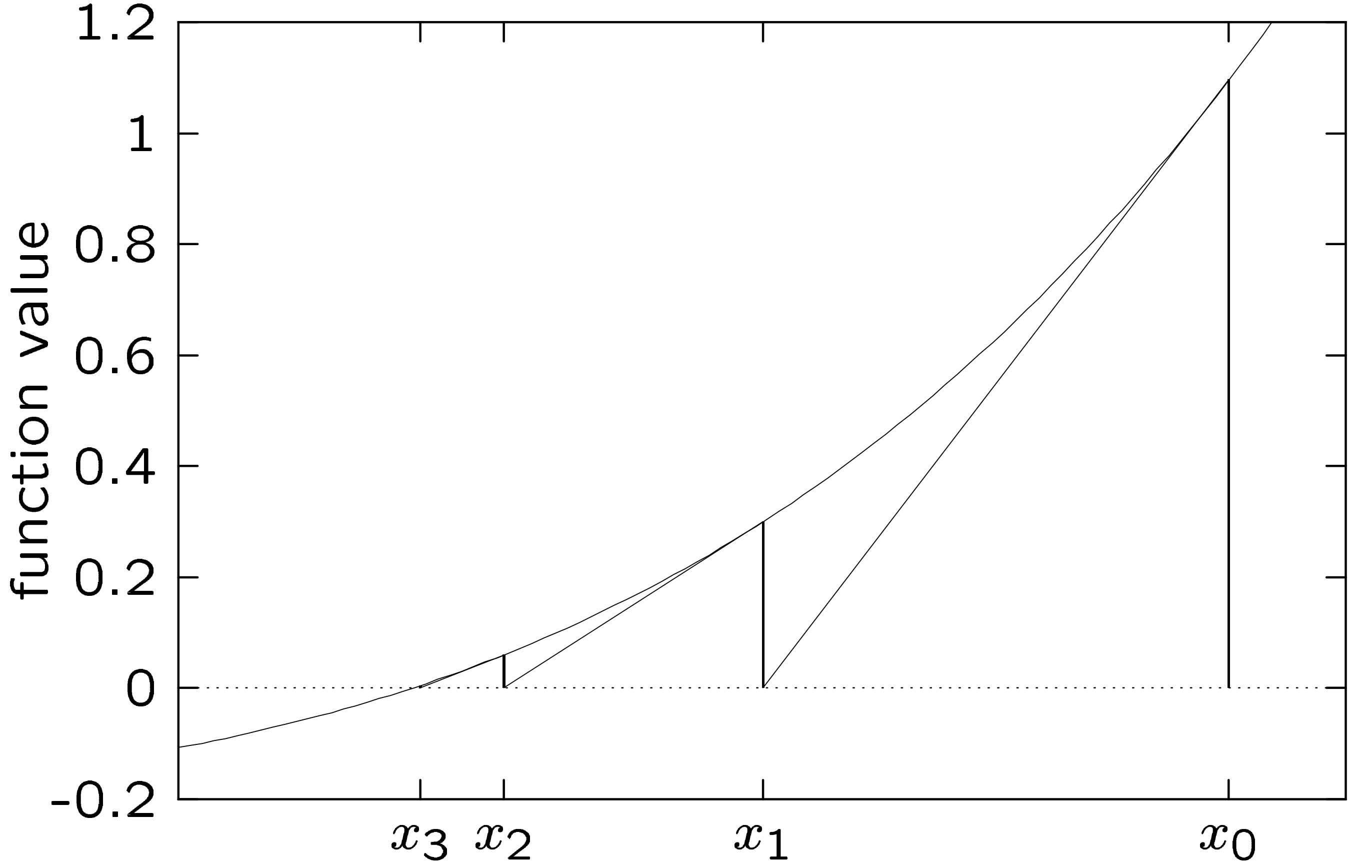
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

- Alternatively (Taylor's): have x_0, for what h is

$$f\Bigg(\underbrace{x_0 + h}_{\equiv x_1}\Bigg) = 0$$

$f(x_0 + h) \approx f(x_0) + hf'(x_0)$ or $h = -\frac{f(x_0)}{f'(x_0)}$

Newton's Method—Example



Convergence Rate

- Theorem: With the following three conditions:

  $^{1)} f(r) = 0$, $^{2)} f'(r) \neq 0$, $^{3)} f \in C^2\Big(B\big(r, \hat{\delta}\big)\Big) \Rightarrow$

  $\exists \delta \ni \forall x_0 \in B(r, \delta)$ and $\forall n$ we have $\Big|x_{n+1} - r\Big| \leq C(\delta)|x_n - r|^2$

  $*$ for a given δ, C is a constant (not necessarily < 1)

- English: With enough continuity and proximity $\Rightarrow$ quadratic convergence!

- Note: again, use graphing tool to seed x_0

Newton's method can be very fast.

Convergence Rate Example

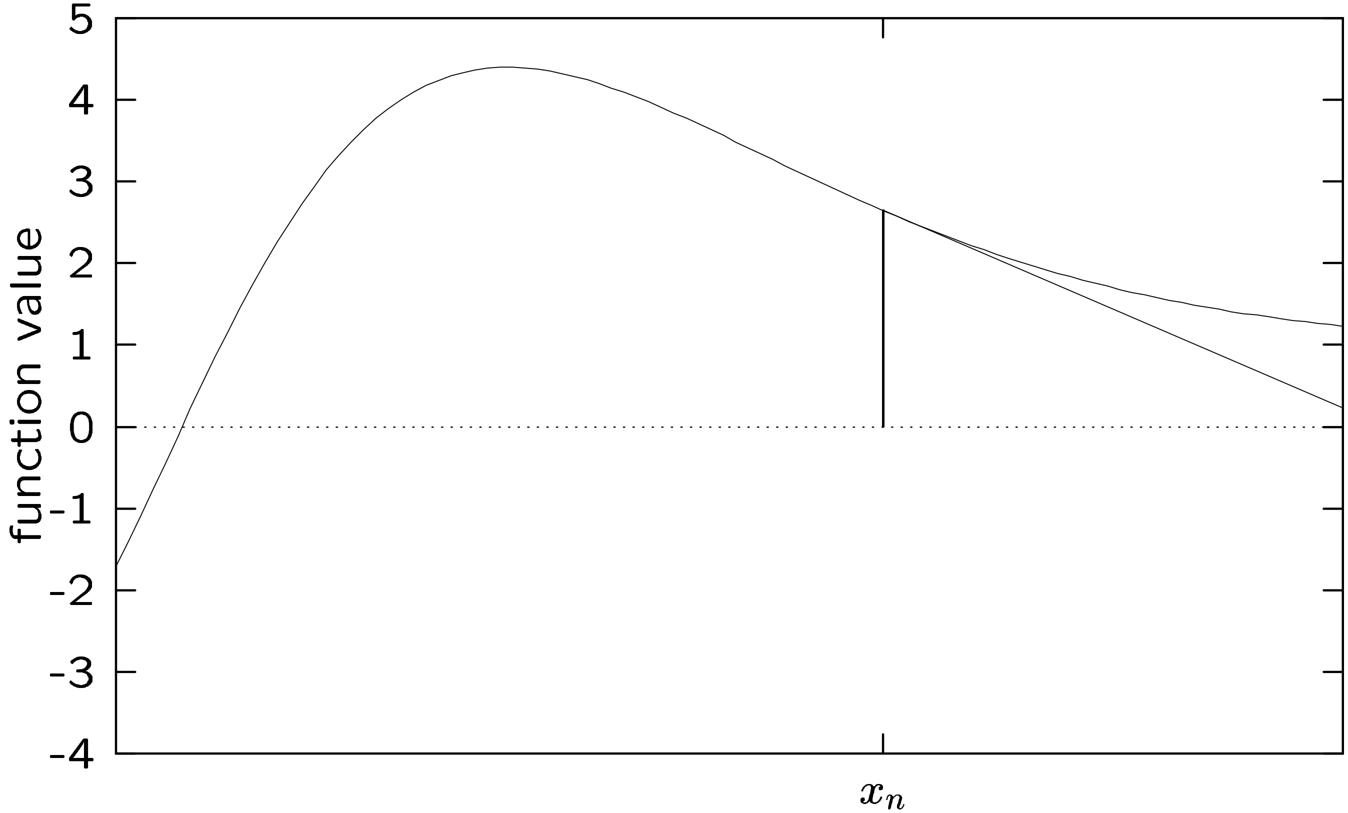
$f(x) = x^3 - 2x^2 + x - 3, \;\; x_0 = 4$

n	x_n	$f(x_n)$
0	4	33
1	3	9
2	2.4375	2.036865234375
3	2.21303271631511	0.256363385061418
4	2.17555493872149	0.00646336148881306
5	2.17456010066645	$4.47906804996122e-06$
6	2.17455941029331	$2.15717547991101e-12$

- Stopping criteria
  - ∗ theorem: uses x; above: uses $f(x)$—often all we have
  - ∗ possibilities: absolute/relative, size/change, x or $f(x)$ (combos, . . .)

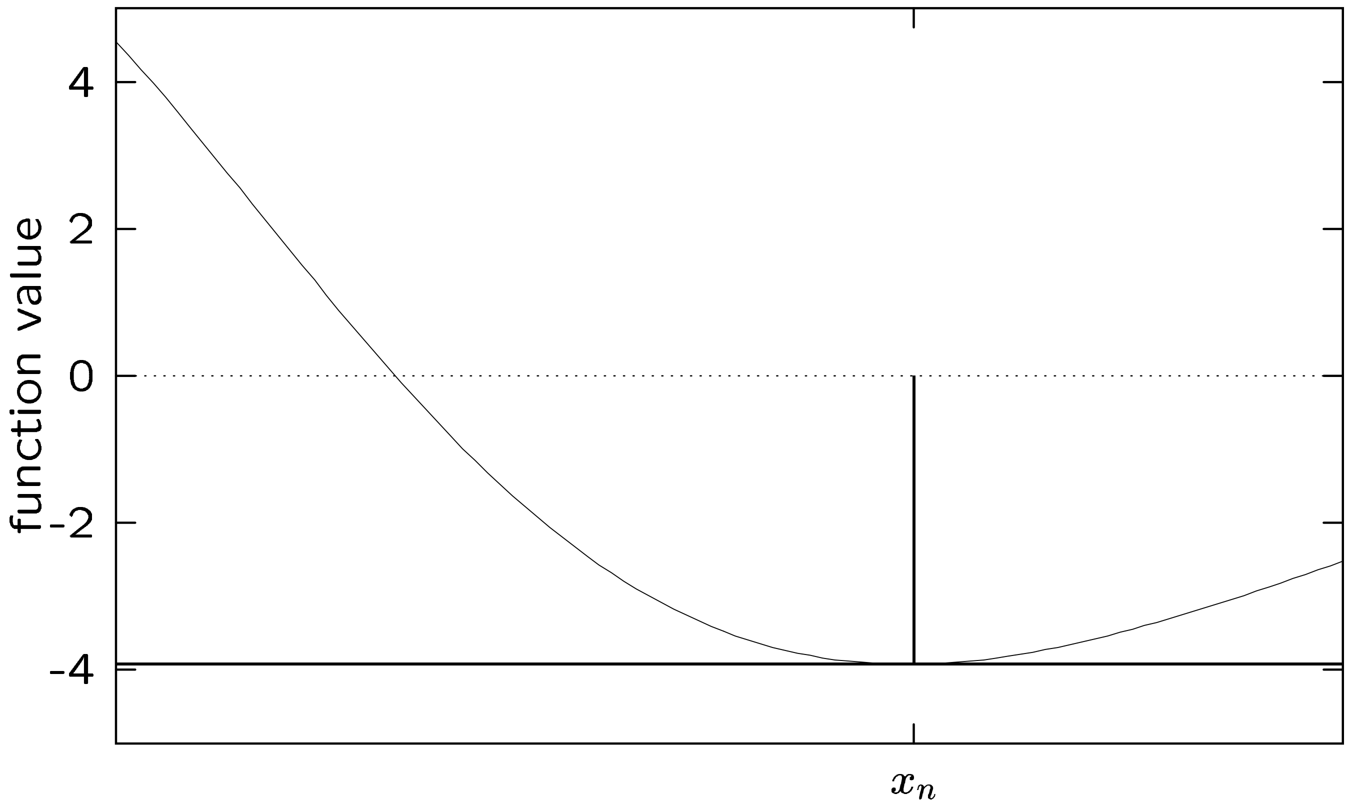
But proximity issue can bite, ...

Sample Newton Failure #1



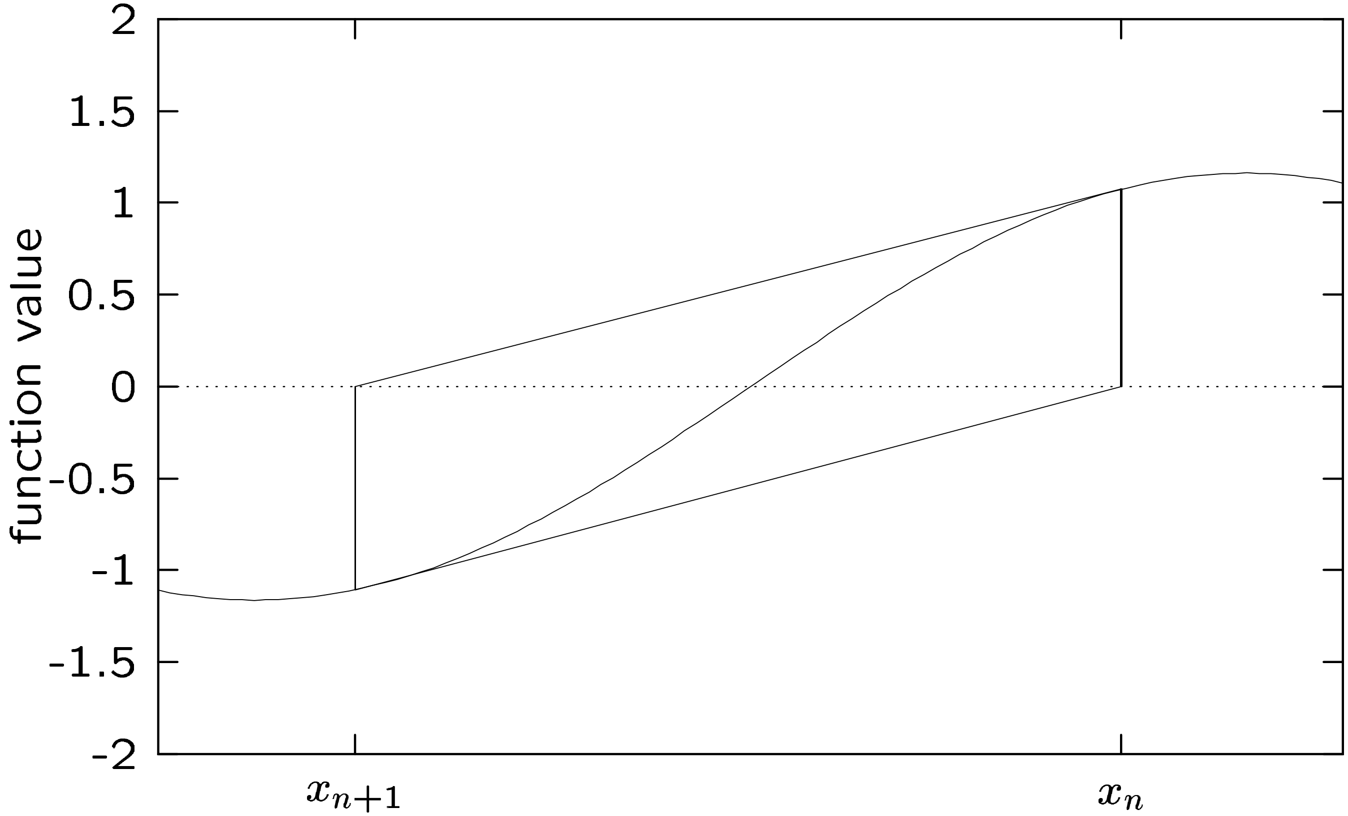
Runaway process

Sample Newton Failure #2



Division by zero derivative—recall algorithm

Sample Newton Failure #3



Loop-d-loop (can happen over m points)

Nonlinear Equations

- Motivation
- Bisection Method
- Newton's Method
- $\Rightarrow$ Secant Method
- Summary

Secant Method—Definition

- Motivation: avoid derivatives
- Taylor (or derivative): $f'(x_n) \approx \frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}$
- $\therefore x_{n+1} = x_n - f(x_n)\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}$
- Bisection requirements comparison:
  - $\boxed{\surd}$ 2 previous points
  - $\boxed{\times}$ $f(a)f(b) < 0$
- Additional advantage vs. Newton:
  - only one function evaluation per iteration
- Superlinear convergence: $\left|x_{n+1} - r\right| \leq C|x_n - r|^{1.618\ldots}$ (recognize the exponent?)

Nonlinear Equations

- Motivation
- Bisection Method
- Newton's Method
- Secant Method

$\Rightarrow$ Summary

Root Finding—Summary

- Performance and requirements

	$f \in C^2$	nbhd(r)	init. pts.	□	■	speedy
bisection	$\times$	$\times$	2	$\surd$	1	$\times$
Newton	$\surd$	$\surd$	1	$\times$	2	$\surd$
secant	$\times$	$\surd$	2	$\times$	1	$\surd\!\!\times$

□ requirement that $f(a)f(b) < 0$

■ function evaluations per iteration

- Often methods are combined (how?), with restarts for divergence or cycles

- Recall: use graphing tool to seed x_0 (and x_1)

Interpolation and Approximation

- ⇒ Motivation
- Polynomial Interpolation
- Numerical Differentiation
- Additional Notes

Motivation

- Three sample problems
  - $\{(x_i, y_i)|i = 0, \ldots, n\}$, ($x_i$ distinct), want simple (e.g., polynomial) $p(x) \ni y_i = p(x_i), i = 0, \ldots, n \quad \equiv$ "interpolation"
  - Assume data includes errors, relax equality but still close, . . . least squares
  - Replace complicated $f(x)$ with simple $p(x) \approx f(x)$

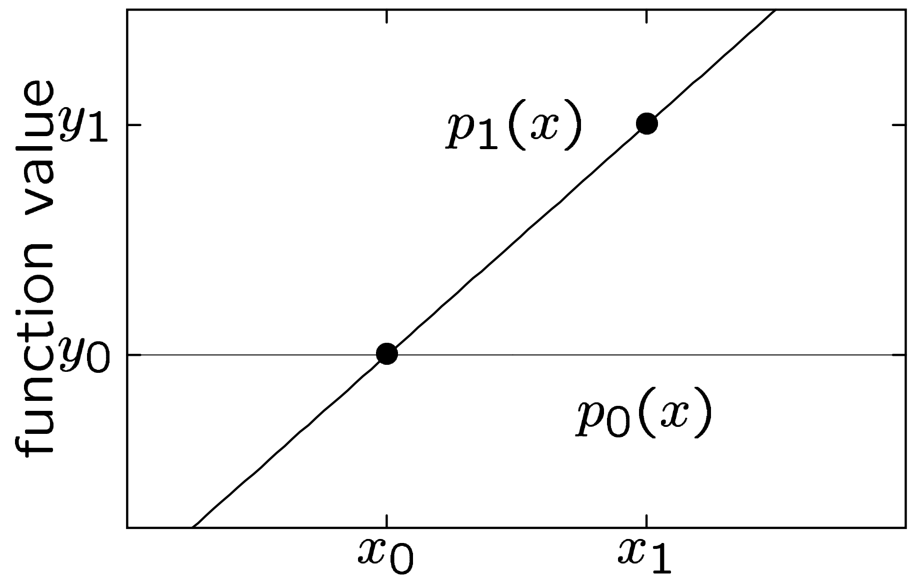
- Interpolation
  - similar to English term (contrast: extrapolation)
  - for now: polynomial
  - later: splines

Use $p(x)$ for $p(x_{\text{new}})$, $\int p(x)\,dx$, ...

Interpolation and Approximation

- Motivation
- ⇒ Polynomial Interpolation
- Numerical Differentiation
- Additional Notes

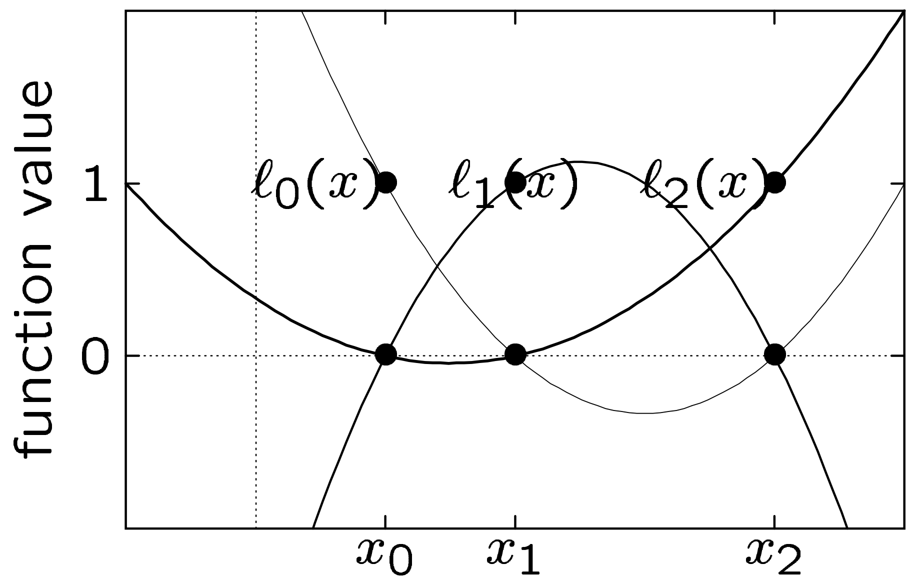
Constant and Linear Interpolation



- $n = 0$: $p(x) = y_0$
- $n = 1$: $p(x) = y_0 + g(x)(y_1 - y_0)$, $g(x) \in P_1$, and $g(x) = \begin{cases} 0, & x = x_0, \\ 1, & x = x_1 \end{cases}$ $\therefore g(x) = \frac{x - x_0}{x_1 - x_0}$
- $n = 2$: more complicated, ...

Lagrange Polynomials

- Given: x_i, $i = 0, \ldots, n$; "Kronecker delta": $\delta_{i\,j} = \begin{cases} 0, & i \neq j, \\ 1, & i = j \end{cases}$
- Lagrange polynomials: $\ell_i(x) \in P_n$, $\ell_i\big(x_j\big) = \delta_{i\,j}$, $i = 0, \ldots, n$
  - $*$ independent of any y_i values
- E.g., $n = 2$:



Lagrange Interpolation

- We have

$$\ell_0(x) = \frac{x-x_1}{x_0-x_1}\cdot\frac{x-x_2}{x_0-x_2}, \quad y_0\ell_0\big(x_j\big) = y_0\delta_{0j} = \begin{cases} 0, & j \neq 0, \\ y_0, & j = 0 \end{cases}$$

$$\ell_1(x) = \frac{x-x_0}{x_1-x_0}\cdot\frac{x-x_2}{x_1-x_2}, \quad y_1\ell_1\big(x_j\big) = y_1\delta_{1j} = \begin{cases} 0, & j \neq 1, \\ y_1, & j = 1 \end{cases}$$

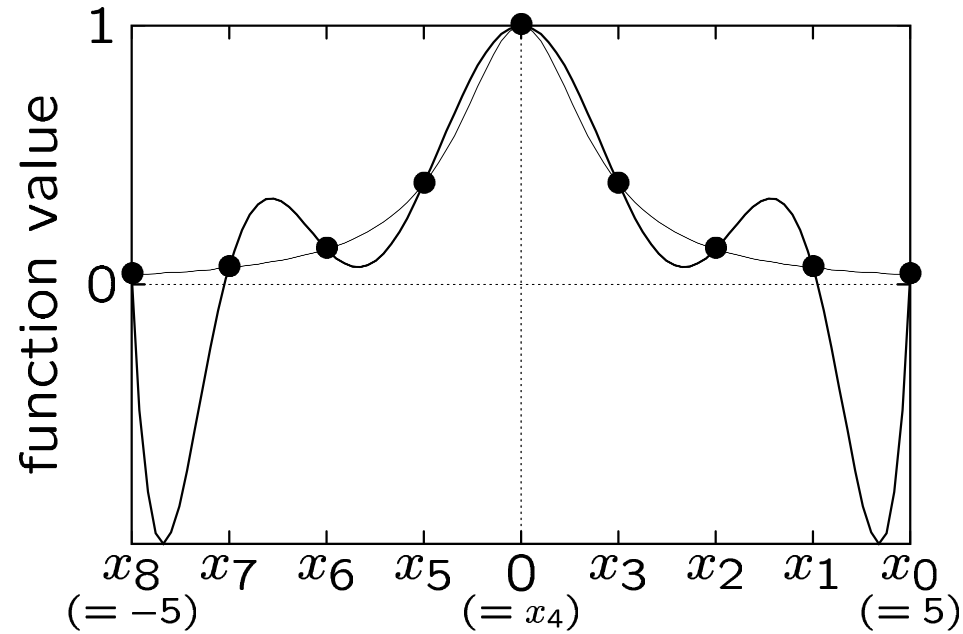
$$\ell_2(x) = \frac{x-x_0}{x_2-x_0}\cdot\frac{x-x_1}{x_2-x_1}, \quad y_2\ell_2\big(x_j\big) = y_2\delta_{2j} = \begin{cases} 0, & j \neq 2, \\ y_2, & j = 2 \end{cases}$$

- $\therefore \exists! p(x) \in P_2$, with $p\big(x_j\big) = y_j$, $j = 0, 1, 2$: $p(x) = \sum_{i=0}^{2} y_i \ell_i(x)$
- In general: $\ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x-x_j}{x_i-x_j}, \quad i = 0, \ldots, n$
- Great! What could be wrong? Easy functions (polynomials), interpolation ($\therefore$ error $= 0$ at x_i) ... but what about $p(x_{\text{new}})$?

Interpolation Error & the Runge Function

- $\{(x_i, f(x_i))|i = 0, \ldots, n\}$, $|f(x) - p(x)| \leq$?
- Runge function: $f_R(x) = \left(1 + x^2\right)^{-1}, x \in [-5, 5]$ and *uniform* mesh: ⚠ $p(x)$'s wrong shape and high oscillations

$$\lim_{n\to\infty} \max_{-5\leq x\leq 5} |f_R(x) - p_n(x)| = \infty$$



Error Theorem

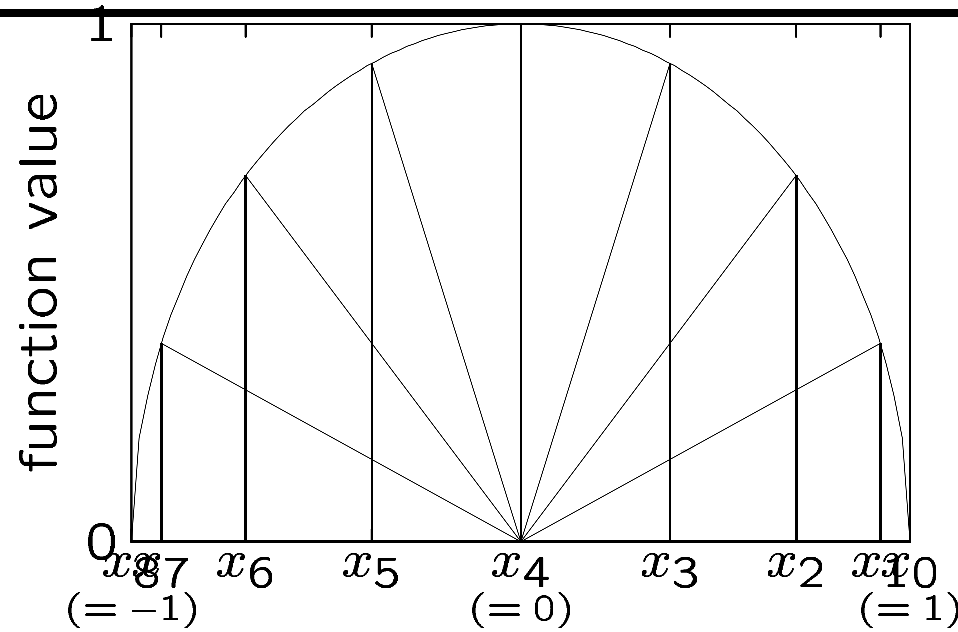
- Theorem: . . . , $f \in C^{n+1}[a,b]$, $\forall x \in [a,b]$, $\exists \xi \in (a,b) \ni$

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) \prod_{i=0}^{n} (x - x_i)$$

- Max error
  - ∗ with x_i and x, still need $\max_{(a,b)} f^{(n+1)}(\xi)$
  - ∗ with x_i only, also need max of $\prod$
  - ∗ without x_i:

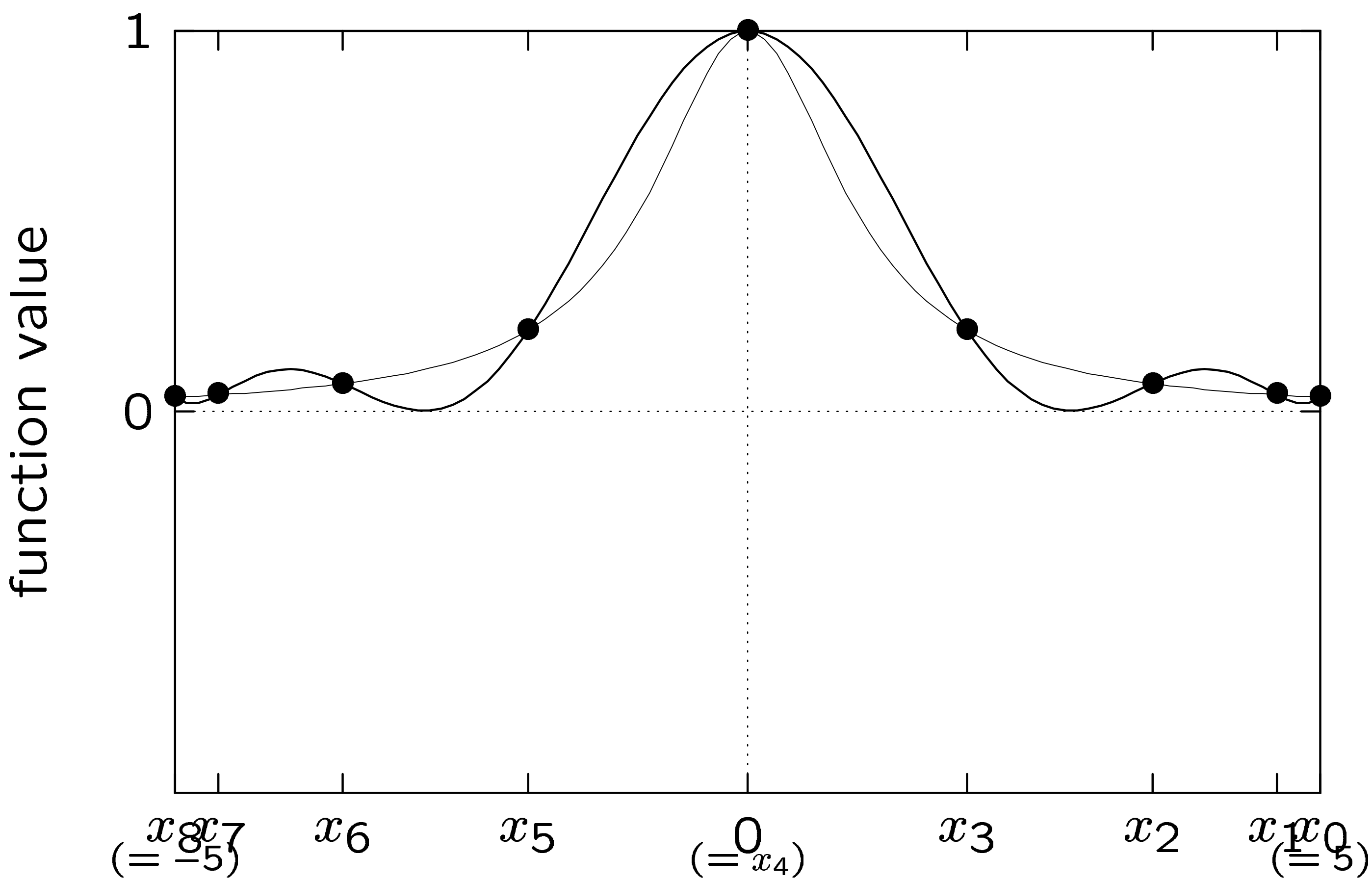
$$\max_{(a,b)} \prod_{i=0}^{n} (x - x_i) = (b-a)^{n+1}$$

Chebyshev Points



- Chebyshev points on $[-1,1]$: $x_i = \cos\left[\left(\frac{i}{n}\right)\pi\right]$, $i = 0,\ldots,n$
- In general on $[a,b]$: $x_i = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\cos\left[\left(\frac{i}{n}\right)\pi\right]$, $i = 0,\ldots,n$
- Points concentrated at edges

Runge Function with Chebyshev Points



Is this good interpolation?

Chebyshev Interpolation

- Same interpolation method

- Different interpolation points

- Minimizes

$$\left| \prod_{i=0}^{n} (x - x_i) \right|$$

- Periodic behavior $\Rightarrow$ interpolate with sins/coss instead of P_n
  - $*$ uniform mesh minimizes max error

- Note: uniform partition with spacing $= \text{cheb}_1 - \text{cheb}_0$
  - $*$ num. points $\uparrow \therefore$ polynomial degree $\uparrow \therefore$ oscillations $\uparrow$

- Note: shape is still wrong . . . see splines later

Interpolation and Approximation

- Motivation
- Polynomial Interpolation
- ⇒ Numerical Differentiation
- Additional Notes

Numerical Differentiation

- Note: until now, approximating $f(x)$, now $f'(x)$
- $f'(x) \approx \frac{f(x+h)-f(x)}{h}$
- Error = ?
- Taylor: $f(x+h) = f(x) + hf'(x) + h^2\frac{f''(\xi)}{2}$
- $\therefore f'(x) = \frac{f(x+h)-f(x)}{h} - \frac{1}{2}hf''(\xi)$
- I.e., truncation error: $O(h)$

Can we do better?

Numerical Differentiation—Take Two

- Taylor for $+h$ and $-h$:

$$f(x \pm h) = f(x) \pm hf'(x) + h^2\frac{f''(x)}{2!} \pm h^3\frac{f'''(x)}{3!} + h^4\frac{f^{(4)}(x)}{4!} \pm h^5\frac{f^{(5)}(x)}{5!} + \cdots$$

- Subtracting:

$$f(x+h) - f(x-h) = 2hf'(x) + 2h^3\frac{f'''(x)}{3!} + 2h^5\frac{f^{(5)}(x)}{5!} + \cdots$$

- $\therefore$

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{1}{6}h^2 f'''(x) - \cdots$$

We gained $O(h)$ to $O\left(h^2\right)$. However, . . .

Richardson Extrapolation—Take Three

- We have

$$f'(x) = \underbrace{\frac{f(x+h) - f(x-h)}{2h}}_{\equiv \phi(h)} + a_2 h^2 + a_4 h^4 + a_6 h^6 + \cdots$$

- Halving the stepsize, ∴

$$\begin{aligned}
\phi(h) &= f'(x) - a_2 h^2 - a_4 h^4 - a_6 h^6 - \cdots \\
\phi\left(\frac{h}{2}\right) &= f'(x) - a_2 \left(\frac{h}{2}\right)^2 - a_4 \left(\frac{h}{2}\right)^4 - a_6 \left(\frac{h}{2}\right)^6 - \cdots \\
\phi(h) - 4\phi\left(\frac{h}{2}\right) &= -3f'(x) - \frac{3}{4} a_4 h^4 - \frac{15}{16} a_6 h^6 - \cdots
\end{aligned}$$

Q: So what? *A*: The h^2 term disappeared!

Richardson—Take Three (cont.)

- Divide by 3 and write $f'(x)$

$$\begin{aligned} f'(x) &= \frac{4}{3}\phi\left(\frac{h}{2}\right) - \frac{1}{3}\phi(h) - \frac{1}{4}a_4h^4 - \frac{5}{16}a_6h^6 - \cdots \\ &= \phi\left(\frac{h}{2}\right) + \underbrace{\frac{1}{3}\left[\phi\left(\frac{h}{2}\right) - \phi(h)\right]}_{\equiv(*)} + O\left(h^4\right) \end{aligned}$$

- $(*)$ only uses old and current information

We gained $O\left(h^2\right)$ to $O\left(h^4\right)$!!

Interpolation and Approximation

- Motivation
- Polynomial Interpolation
- Numerical Differentiation

$\Rightarrow$ Additional Notes

Additional Notes

- Three $f'(x)$ formulae used additional points $\Rightarrow$
  vs. Taylor, more derivatives in *same* point

- Similar for $f''(x)$:

$$f(x \pm h) = f(x) \pm hf'(x) + h^2 \frac{f''(x)}{2!} \pm h^3 \frac{f'''(x)}{3!} + h^4 \frac{f^{(4)}(x)}{4!} \pm h^5 \frac{f^{(5)}(x)}{5!} + \cdots$$

Adding:

$$f(x+h) + f(x-h) = 2f(x) + h^2 f''(x) + \frac{1}{12} h^4 f^{(4)}(x) + \cdots$$

or:

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + \frac{1}{12} h^2 f^{(4)}(x) + \cdots$$

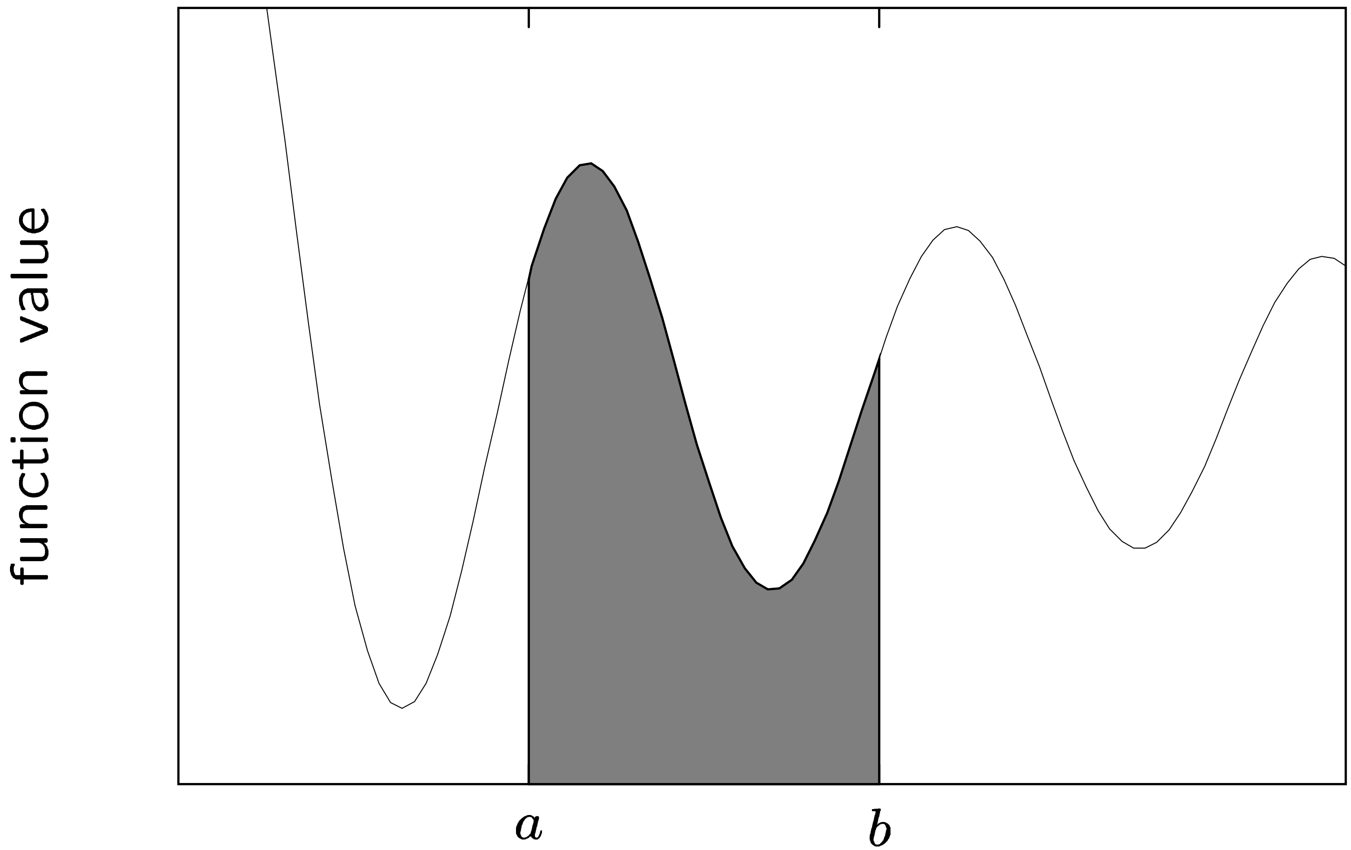
$\therefore$ error $= O\left(h^2\right)$

Numerical Quadrature

- ⇒ Introduction
- Riemann Integration
- Composite Trapezoid Rule
- Composite Simpson's Rule
- Gaussian Quadrature

Numerical Quadrature—Interpretation

- $f(x) \geq 0$ on $[a,b]$ bounded $\Rightarrow \int_a^b f(x)\,dx$ is area under $f(x)$



Numerical Quadrature—Motivation

- Analytical solutions—rare:

$$\int_0^{\frac{\pi}{2}} \sin x \, dx = -\cos x|_0^{\frac{\pi}{2}} = -(0-1) = 1$$

- In general:

$$\int_0^{\frac{\pi}{2}} \left(1 - a^2 \sin^2 \theta\right)^{\frac{1}{3}} d\theta$$

Need general numerical technique.

Definitions

- Mesh: $P \equiv \{a = x_0 < x_1 < \cdots < x_n = b\}$, n subintervals ($n+1$ points)

- Infima and suprema:

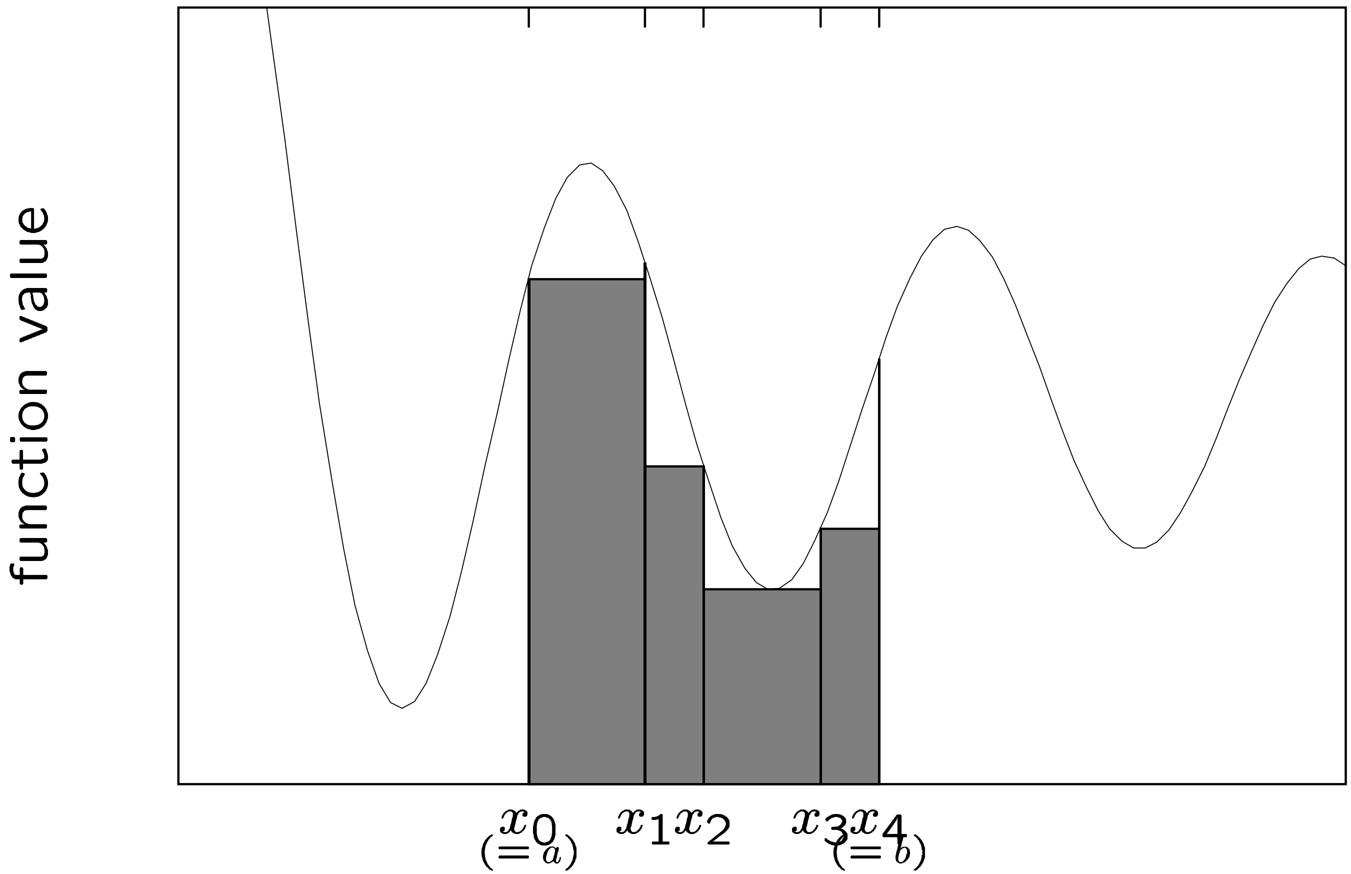
$$m_i \equiv \inf\left\{f(x) : x_i \le x \le x_{i+1}\right\}$$
$$M_i \equiv \sup\left\{f(x) : x_i \le x \le x_{i+1}\right\}$$

- Two methods (i.e., integral estimates): lower and upper sums

$$L(f;P) \equiv \sum_{i=0}^{n-1} m_i\left(x_{i+1} - x_i\right)$$
$$U(f;P) \equiv \sum_{i=0}^{n-1} M_i\left(x_{i+1} - x_i\right)$$

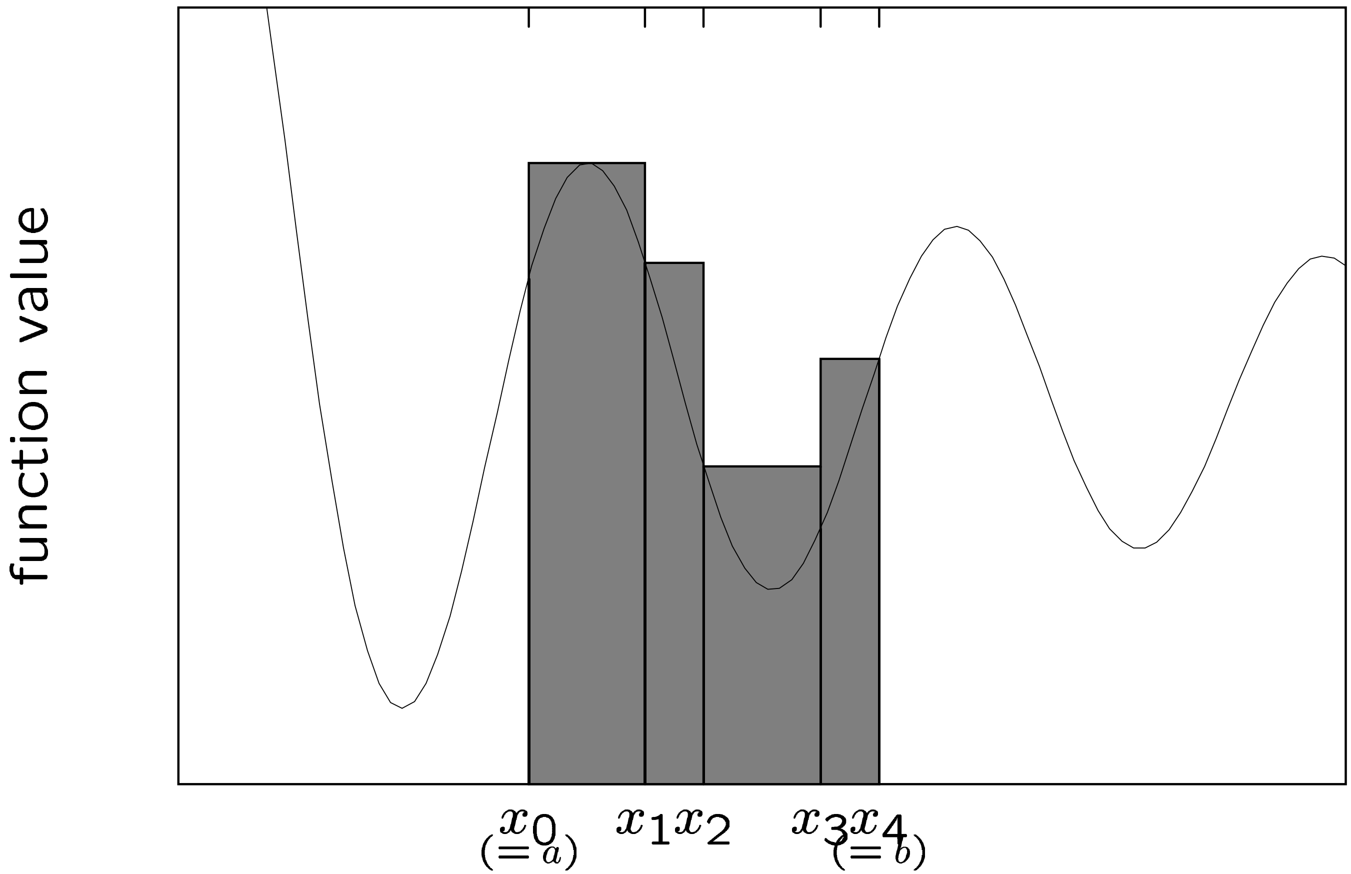
- For example, ...

Lower Sum—Interpretation



Clearly a lower bound of integral estimate, and . . .

Upper Sum—Interpretation



. . . an upper bound. What is the max error?

Lower and Upper Sums—Example

- Third method, use lower and upper sums: $(L + U)/2$
- $f(x) = x^2$, $[a, b] = [0, 1]$ and $P = \left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}$
- $\ldots$, $L = \frac{7}{32}$, $U = \frac{15}{32}$
- Split the difference: estimate $\frac{11}{32}$ (actual $\frac{1}{3}$)
- Bottom line
  - naive approach
  - low n
  - still error of $\frac{1}{96}$. (!)
- Max error: $(U - L)/2 = \frac{1}{8}$

Is this good enough?

Numerical Quadrature—Rethinking

- Perhaps lower and upper sums are enough?
  - Error seems small
  - Work seems small as well

- But: estimate of *max* error was not small ($\frac{1}{8}$)

- Do they converge to integral as $n \to \infty$?

- Will the extrema always be easy to calculate? Accurately? (Probably not!)

Proceed in theoretical and practical directions.

Numerical Quadrature

- Introduction

$\Rightarrow$ Riemann Integration

- Composite Trapezoid Rule
- Composite Simpson's Rule
- Gaussian Quadrature

Riemann Integrability

- $f \in C^0[a,b]$, $[a,b]$ bdd $\Rightarrow f$ is Riemann integrable
- When integrable, and max subinterval in $P \to 0$ ($|P| \to 0$):

$$\lim_{|P| \to 0} L(f;P) = \int_a^b f(x)\,dx = \lim_{|P| \to 0} U(f;P)$$

- Counter example: Dirichlet function $d(x) \equiv \begin{cases} 0, & x \text{ rational}, \\ 1, & x \text{ irrational} \end{cases}$

  $\Rightarrow L = 0,\; U = b - a$

Challenge: Estimate n for Third Method

- Current restrictions for n estimate:
  - $*$ Monotone functions
  - $*$ Uniform partition

- Challenge:
  - $*$ estimate $\int_0^{\pi} e^{\cos x}\, dx$
  - $*$ error tolerance $= \frac{1}{2} \times 10^{-3}$
  - $*$ using L and U
  - $*$ $n = ?$

Estimate n—Solution

- $f(x) = e^{\cos x} \searrow$ on $[0, \pi]$ $\therefore m_i = f\big(x_{i+1}\big)$ and $M_i = f(x_i)$
- $\therefore L(f; P) = h \sum_{i=0}^{n-1} f\big(x_{i+1}\big)$ and $U(f; P) = h \sum_{i=0}^{n-1} f(x_i)$, $h = \frac{\pi}{n}$
- Want $\frac{1}{2}(U - L) < \frac{1}{2} \times 10^{-3}$ or $\frac{\pi}{n}\big(e^1 - e^{-1}\big) < 10^{-3}$
- ... $n \geq 7385$ (!!) (note for later: max error estimate $= O(h)$)
- Number of $f(x)$ evaluations
  - ∗ 2 for $(U - L)$ max error calculation
  - ∗ > 7000 for either L or U

We need something better.

Numerical Quadrature

- Introduction
- Riemann Integration

$\Rightarrow$ Composite Trapezoid Rule

- Composite Simpson's Rule
- Gaussian Quadrature

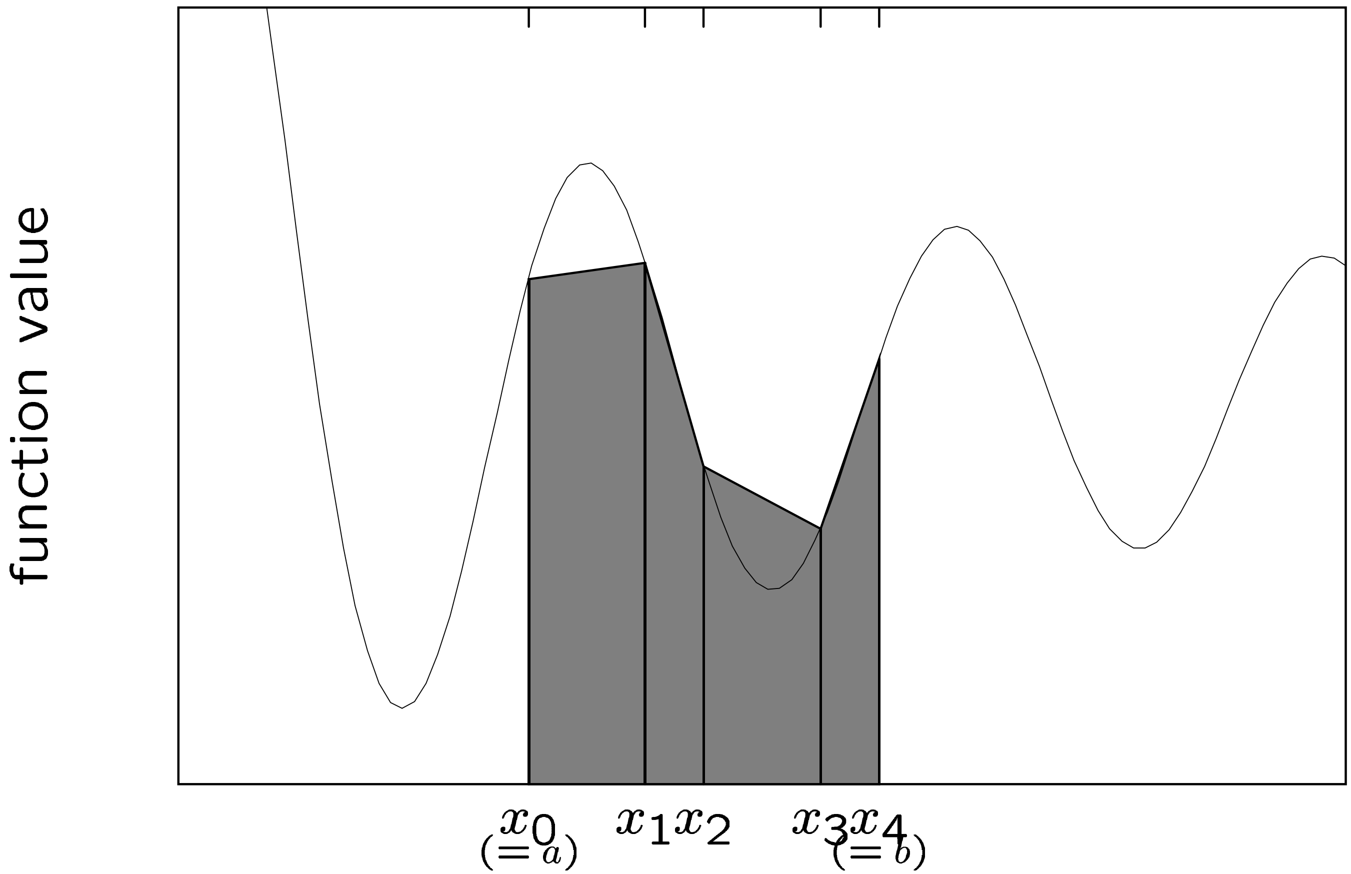
Composite Trapezoid Rule (CTR)

- Each area: $\frac{1}{2}\left(x_{i+1} - x_i\right)\left[f(x_i) + f\left(x_{i+1}\right)\right]$
- Rule: $T(f;P) \equiv \frac{1}{2}\sum_{i=0}^{n-1}\left(x_{i+1} - x_i\right)\left[f(x_i) + f\left(x_{i+1}\right)\right]$
- Note: for monotone functions and any given mesh (why?):

$$T = (L + U)/2$$

- Pro: no need for extrema calculations
- Con: adding new points to existing ones (for a non-monotonic function)
  - T can land on "bad point" $\Rightarrow$ no monotonic improvement (necessarily)
  - L and U look for extremum on $\left[x_i, x_{i+1}\right]$ $\Rightarrow$ monotonic improvement

CTR—Interpretation



Almost always better than L or U. (When not?)

Uniform Mesh and Associated Error

- Constant stepsize $h = \frac{b-a}{n}$

$$T(f;P) \equiv h\left\{\sum_{i=1}^{n-1} f(x_i) + \frac{1}{2}[f(x_0) + f(x_n)]\right\}$$

- Theorem: $f \in C^2[a,b] \to \exists \xi \in (a,b) \ni$

$$\int_a^b f(x)\,dx - T(f;P) = -\frac{1}{12}(b-a)h^2 f''(\xi) = O\big(h^2\big)$$

- Note: leads to popular Romberg algorithm (built on Richardson extrapolation)

How many steps does $T(f;P)$ require?

$e^{\cos x}$ Revisited—Using CTR

- Challenge: $\int_0^\pi e^{\cos x}\,dx$, error tolerance $= \frac{1}{2} \times 10^{-3}$, $n = ?$
- $f(x) = e^{\cos x} \Rightarrow f'(x) = -e^{\cos x}\sin x \ \ldots \ |f''(x)| \le e$ on $[0, \pi]$
- $\therefore$ |error| $\le \frac{1}{12}\pi(\pi/n)^2 e \le \frac{1}{2} \times 10^{-3}$
- $\ldots \ n \ge 119$
- Recall perennial two questions/calculations of NM
  - ∗ monotonic $\therefore$ estimate of T produces same $(L+U)/2$
  - ∗ but previous *max error* estimate was less exact $(O(h))$

Better estimate of *max error* $\therefore$ better estimate of n

Another CTR Example

- Challenge: $\int_0^1 e^{-x^2}\,dx$, error tolerance $= \frac{1}{2} \times 10^{-4}$, $n = ?$
- $f(x) = e^{-x^2}$, $\Rightarrow$ $f'(x) = -2xe^{-x^2}$ and $f''(x) = \left(4x^2 - 2\right)e^{-x^2}$
- $\therefore\ |f''(x)| \le 2$ on $[0, 1]$
- $\Rightarrow$ $|\text{error}| \le \frac{1}{6}h^2 \le \frac{1}{2} \times 10^{-4}$
- We have: $n^2 \ge \frac{1}{3} \times 10^4$ or $n \ge 58$ subintervals

How can we do better?

Numerical Quadrature

- Introduction
- Riemann Integration
- Composite Trapezoid Rule

$\Rightarrow$ Composite Simpson's Rule

- Gaussian Quadrature

Trapezoid Rule as $\int$ Linear Interpolant

Linear interpolant, one subinterval: $p_1(x) = \frac{x-b}{a-b} f(a) + \frac{x-a}{b-a} f(b)$, intuitively:

$$
\begin{aligned}
\int_a^b p_1(x)\,dx &= \frac{f(a)}{a-b}\int_a^b (x-b)\,dx + \frac{f(b)}{b-a}\int_a^b (x-a)\,dx \\
&= \frac{f(a)}{a-b}\left[\frac{b^2-a^2}{2} - b(b-a)\right] + \frac{f(b)}{b-a}\left[\frac{b^2-a^2}{2} - a(b-a)\right] \\
&= -f(a)\left[\frac{a+b}{2} - b\right] + f(b)\left[\frac{a+b}{2} - a\right] \\
&= -f(a)\left(\frac{a-b}{2}\right) + f(b)\left(\frac{b-a}{2}\right) \\
&= \frac{b-a}{2}(f(a)+f(b))
\end{aligned}
$$

CTR is integral of composite linear interpolant.

CTR for Two Equal Subintervals

- $n = 2$ (i.e., 3 points):

$$
\begin{aligned}
T(f) &= \frac{b-a}{2}\left\{f\left(\frac{a+b}{2}\right) + \frac{1}{2}[f(a)+f(b)]\right\} \\
&= \frac{b-a}{4}\left[f(a) + 2f\left(\frac{a+b}{2}\right) + f(b)\right]
\end{aligned}
$$

  with error $= O\left(\left(\frac{b-a}{2}\right)^3\right)$

- (Previously, CTR error $= O\left(h^2\right) =$ TR error $\times\ n$ subintervals $= O\left(h^3\right) \times O\left(\frac{1}{h}\right)$)

- Deficiency: each subinterval ignores the other

How can we take the entire picture into account?

Simpson's Rule

- Motivation: use $p_2(x)$ over the two equal subintervals
- Similar analysis actually loses $O(h)$, but ... $\exists \xi \in (a,b) \ni$

$$\int_a^b f(x)\,dx = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] - \frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi)$$

- Similar to CTR, but weights midpoint more
- Note: for each method, denominator $= \sum$ coefficients

Each method multiplies width by weighted average of height.

Composite Simpson's Rule (CSR)

- For an even number of subintervals n, $h = \frac{b-a}{n}$, $\exists \xi \in (a,b) \ni$

$$\int_a^b f(x)\,dx = \frac{h}{3}\left\{[f(a)+f(b)] + 4\sum_{i=1}^{n/2} \underbrace{f[a+(2i-1)h]}_{\text{odd nodes}} + 2\sum_{i=1}^{(n-2)/2} \underbrace{f(a+2ih)}_{\text{even nodes}}\right\} - \frac{b-a}{180}h^4 f^{(4)}(\xi)$$

- Note: denominator $= \sum$ coefficients $= 3n$
  - *but* only $n+1$ function evaluations

Can we do better than $O\left(h^4\right)$?

Evaluating the Error

- Another important accuracy angle
  - until now: error $= O(h^\alpha)$
  - now on, looking at $f^{(\beta)}$: error $= 0\ \forall f \in P_{\beta-1}$

- With higher β, $p_\beta(x)$ can approximate any $f(x)$ better

- Define $\epsilon(x) \equiv f(x) - p_\beta(x)$

- $\int f = \int (p_\beta + \epsilon) = \int p_\beta + \int \epsilon = \text{method}\big(p_\beta\big) + \int \epsilon =$ $\text{method}(f) - \text{method}(\epsilon) + \int \epsilon$

- As $\beta\uparrow$: $\epsilon(x)\downarrow$, $\left(\int \epsilon - \text{method}(\epsilon)\right)\downarrow \therefore \text{method}(f) \to \int f$

Can we do better than Simpson's P_3?

Integration Introspection

- Simpson beat CTR because heavier weighted midpoint
- But CSR similarly suffers at subinterval-pair boundaries (weight $=2$ vs. 4 for no reason)
- All composite rules
  - ∗ ignore other areas
  - ∗ patch together local calculations
  - ∗ $\therefore$ will suffer from this
- What about using all nodes and higher degree interpolation?
- Also note: we can choose
  - ∗ weights
  - ∗ location of calculation nodes

Numerical Quadrature

- Introduction
- Riemann Integration
- Composite Trapezoid Rule
- Composite Simpson's Rule

$\Rightarrow$ Gaussian Quadrature

Interpolatory Quadrature

- x_i, $\ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$, $i = 0, \ldots, n$; $p(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x)$

- If $f(x) \approx p(x) \Rightarrow$ hopefully $\int_a^b f(x)\,dx \approx \int_a^b p(x)\,dx$

- $\int_a^b p(x)\,dx = \int_a^b \sum_{i=0}^{n} f(x_i)\ell_i(x)\,dx = \sum_{i=0}^{n} f(x_i) \underbrace{\int_a^b \ell_i(x)\,dx}_{\equiv A_i}$

- $A_i = A_i\left(a, b; \left\{x_j\right\}_{j=0}^{n}\right)$, but $A_i \neq A_i(f)$!

(Endpoints, nodes) $\Rightarrow A_i \Rightarrow \int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i)$.

Interp. Quad.—Error Analysis

- $\forall f \in P_n \Rightarrow f(x) = p(x)$, and $\therefore$
  $\forall f \in P_n \Rightarrow \int_a^b f(x)\,dx = \sum_{i=0}^{n} A_i f(x_i)$, i.e., error = 0
- $n+1$ weights determined by nodes x_i (and a and b)
- True for *any* choice of $n+1$ nodes x_i
- What if we choose $n+1$ *specific* nodes (with weights, total: $2(n+1)$ choices)?

Can we get error $= 0$ $\forall f \in P_{2n+1}$?

Gaussian Quadrature (GQ)—Theorem

- Let
  - $q(x) \in P_{n+1} \ni \int_a^b x^k q(x)\,dx = 0, \;\; k = 0, \ldots, n$
    i.e., $q(x) \perp$ all polynomials of lower degree
  - note: $n+2$ coefficients, $n+1$ conditions
    - ⋆ unique to a constant multiplier
  - x_i, $i = 0, \ldots, n$, $\ni$ $q(x_i) = 0$
    i.e., x_i are zeros of $q(x)$

- Then $\forall f \in P_{2n+1}$, even though $f(x) \neq p(x)$ ($\forall f \in P_m$, $m > n$)

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} A_i f(x_i)$$

We jumped from P_n to P_{2n+1}!

Gaussian Quadrature—Proof

- Let $f \in P_{2n+1}$, and divide by $q \ni f = sq + r \therefore s, r \in P_n$
- We have (note: until last step, x_i can be arbitrary)

$$\begin{aligned}
\int_a^b f(x)\,dx &= \int_a^b s(x)q(x)\,dx + \int_a^b r(x)\,dx && \text{(division above)} \\
&= \int_a^b r(x)\,dx && (\perp' \text{ ity of } q(x)) \\
&= \sum_{i=0}^{n} A_i r(x_i) && (r \in P_n) \\
&= \sum_{i=0}^{n} A_i[f(x_i) - s(x_i)q(x_i)] && \text{(division above)} \\
&= \sum_{i=0}^{n} A_i f(x_i) && (x_i \text{ are zeros of } q(x))
\end{aligned}$$

■

GQ—Additional Notes

- Example $q_n(x)$: Legendre Polynomials: for $[a, b] = [-1, 1]$ and $q_n(1) = 1$ ($\exists$ a 3-term recurrence formula)

$$q_0(x) = 1, \;\; q_1(x) = x, \;\; q_2(x) = \frac{3}{2}x^2 - \frac{1}{2}, \;\; q_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x, \ldots$$

- Use $q_{n+1}(x)$ (why?), depends only on a, b and n

- Gaussian nodes $\in (a, b) \Rightarrow$

  good if $f(a) = \infty$ and/or $f(b) = \infty$ (e.g., $\displaystyle\int_0^1 \frac{1}{\sqrt{x}}\,dx$)

- More general: with weight function $w(x)$ in
  - original integral
  - $q(x)$ orthogonality
  - weights A_i

Numerical Quadrature—Summary

- $n + 1$ function evaluations

	composite?	node placement	error $= 0 \; \forall P_{\square}$
CTR	$\surd$	uniform (usually)*	1
CSR	$\surd$	uniform (usually)*	3
interp.	$\times$	any (distinct)	n
GQ	$\times$	zeros of $q(x)$	$2n + 1$

*P.S. There are also powerful adaptive quadrature methods

Linear Systems

- ⇒ Introduction
- Naive Gaussian Elimination
- Limitations
- Operation Counts
- Additional Notes

What Are Linear Systems (LS)?

$$\begin{array}{ccccccccc} a_{1\,1}\,x_1 & + & a_{1\,2}\,x_2 & + & \cdots & + & a_{1\,n}\,x_n & = & b_1 \\ a_{2\,1}\,x_1 & + & a_{2\,2}\,x_2 & + & \cdots & + & a_{2\,n}\,x_n & = & b_2 \\ \vdots & + & \vdots & + & \ddots & + & \vdots & = & \vdots \\ a_{m\,1}\,x_1 & + & a_{m\,2}\,x_2 & + & \cdots & + & a_{m\,n}\,x_n & = & b_m \end{array}$$

- Dependence on unknowns: powers of degree ≤ 1
- Summation form: $\sum_{j=1}^{n} a_{i\,j}\,x_j = b_i$, $1 \leq i \leq m$, i.e., m equations
- Presently: $m = n$, i.e., square systems (later: $m \neq n$)

Q: How to solve for $[x_1 \;\; x_2 \;\; \ldots \;\; x_n]^T$? *A*: ...

Linear Systems

- Introduction
- ⇒ Naive Gaussian Elimination
- Limitations
- Operation Counts
- Additional Notes

Overall Algorithm and Definitions

- Currently: direct methods only (later: iterative methods)
- General idea:
  - ∗ Generate upper triangular system ("forward elimination")
  - ∗ Easily calculate unknowns in reverse order ("backward substitution")
- "Pivot row" = current one being processed
  "pivot" = diagonal element of pivot row

Steps applied to RHS as well.

Forward Elimination

- Generate zero columns below diagonal
- Process rows downward

  for each row $i := 1, n-1$ $\{$ // the pivot row
      for each row $k := i+1, n$ $\{$ // $\forall$ rows below pivot
          multiply pivot row $\ni a_{i\,i} = a_{k\,i}$
          subtract pivot row from row$_k$ // now $a_{k\,i} = 0$
      $\}$ // now column below $a_{i\,i}$ is zero
  $\}$ // now $a_{i\,j} = 0$, $\forall i > j$

- Obtain triangular system

Let's work an example, . . .

Compact Form of LS

$$\left.\begin{array}{rrcrcrcrcr} & 6\,x_1 & - & 2\,x_2 & + & 2\,x_3 & + & 4\,x_4 & = & 16 \\ & 12\,x_1 & - & 8\,x_2 & + & 6\,x_3 & + & 10\,x_4 & = & 26 \\ & 3\,x_1 & - & 13\,x_2 & + & 9\,x_3 & + & 3\,x_4 & = & -19 \\ - & 6\,x_1 & + & 4\,x_2 & + & 1\,x_3 & - & 18\,x_4 & = & -34 \end{array}\right\} \rightarrow$$

$$\left(\begin{array}{rrrr|r} 6 & -2 & 2 & 4 & 16 \\ 12 & -8 & 6 & 10 & 26 \\ 3 & -13 & 9 & 3 & -19 \\ -6 & 4 & 1 & -18 & -34 \end{array}\right)$$

Proceeding with the forward elimination, . . .

Forward Elimination—Example

$$\left(\begin{array}{rrrr|r} 6 & -2 & 2 & 4 & 16 \\ 12 & -8 & 6 & 10 & 26 \\ 3 & -13 & 9 & 3 & -19 \\ -6 & 4 & 1 & -18 & -34 \end{array}\right) \rightarrow \left(\begin{array}{rrrr|r} 6 & -2 & 2 & 4 & 16 \\ 0 & -4 & 2 & 2 & -6 \\ 0 & -12 & 8 & 1 & -27 \\ 0 & 2 & 3 & -14 & -18 \end{array}\right) \rightarrow$$

$$\left(\begin{array}{rrrr|r} 6 & -2 & 2 & 4 & 16 \\ 0 & -4 & 2 & 2 & -6 \\ 0 & 0 & 2 & -5 & -9 \\ 0 & 0 & 4 & -13 & -21 \end{array}\right) \rightarrow \left(\begin{array}{rrrr|r} 6 & -2 & 2 & 4 & 16 \\ 0 & -4 & 2 & 2 & -6 \\ 0 & 0 & 2 & -5 & -9 \\ 0 & 0 & 0 & -3 & -3 \end{array}\right)$$

Matrix is upper triangular.

Backward Substitution

$$\left(\begin{array}{rrrr|r} 6 & -2 & 2 & 4 & 16 \\ 0 & -4 & 2 & 2 & -6 \\ 0 & 0 & 2 & -5 & -9 \\ 0 & 0 & 0 & -3 & -3 \end{array}\right)$$

- Last equation: $-3x_4 = -3 \Rightarrow x_4 = 1$
- Second to last equation: $2x_3 - 5\underbrace{x_4}_{=1} = 2x_3 - 5 = -9 \Rightarrow x_3 = -2$
- ... second equation ... $x_2 = \ldots$
- ... $[x_1\ \ x_2\ \ x_3\ \ x_4]^T = [3\ \ 1\ \ -2\ \ 1]^T$

For small problems, check solution in original system.

Linear Systems

- Introduction
- Naive Gaussian Elimination
- $\Rightarrow$ Limitations
- Operation Counts
- Additional Notes

Zero Pivots

- Clearly, zero pivots prevent forward elimination
- ⚠ zero pivots can appear along the way
- Later: When guaranteed no zero pivots?
- All pivots $\neq 0 \overset{?}{\Rightarrow}$ we are safe

Experiment with system with known solution.

Vandermonde Matrix

$$\begin{pmatrix} 1 & 2 & 4 & 8 & \cdots & 2^{n-1} \\ 1 & 3 & 9 & 27 & \cdots & 3^{n-1} \\ 1 & 4 & 16 & 64 & \cdots & 4^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & n+1 & (n+1)^2 & (n+1)^3 & \cdots & (n+1)^{n-1} \end{pmatrix}$$

- Want row sums on RHS $\Rightarrow x_i = 1,\ i = 1, \ldots, n$
- Geometric series:

$$1 + t + t^2 + \cdots + t^{n-1} = \frac{t^n - 1}{t - 1}$$

- We obtain b_i, for row $i = 1, \ldots, n$

$$\sum_{j=1}^{n} \underbrace{(1+i)^{j-1}}_{a_{ij}} \cdot \underbrace{1}_{x_j} = \frac{(1+i)^n - 1}{(1+i) - 1} = \underbrace{\frac{1}{i}[(1+i)^n - 1]}_{b_i}$$

System is ready to be tested.

Vandermonde Test

- Platform with 7 significant (decimal) digits
  - $n = 1, \ldots, 8 \Rightarrow$ expected results
  - $n = 9$: error $>$ 16,000% !!

- Questions:
  - What happened?
  - Why so sudden?
  - Can anything be done?

- Answer: matrix is "ill-conditioned"
  - Sensitivity to roundoff errors
  - Leads to error propagation and magnification

First, how to assess vector errors.

Errors

- Given system: $Ax = b$ and solution estimate $\tilde{x}$
- Residual (error): $r \equiv A\tilde{x} - b$
- Absolute error (if x is known): $e \equiv x - \tilde{x}$
- Norm taken of r or e: vector $\rightarrow$ scalar quantity
  (more on norms later)
- Relative errors: $||r||/||b||$ and $||e||/||x||$

Back to ill-conditioning, . . .

Ill-conditioning

- $$\left.\begin{array}{rcrcl} 0\cdot x_1 & + & x_2 & = & 1 \\ x_1 & + & x_2 & = & 2 \end{array}\right\} \Rightarrow 0 \text{ pivot}$$

- General rule: if 0 is problematic $\Rightarrow$
  numbers near 0 are problematic

- $$\left.\begin{array}{rcrcl} \epsilon\, x_1 & + & x_2 & = & 1 \\ x_1 & + & x_2 & = & 2 \end{array}\right\} \ldots\ x_2 = \tfrac{2-1/\epsilon}{1-1/\epsilon} \text{ and } x_1 = \tfrac{1-x_2}{\epsilon}$$

- ϵ small (e.g., $\epsilon = 10^{-9}$ with 8 significant digits) $\Rightarrow x_2 = 1$ and $x_1 = 0$—wrong!

What can be done?

Pivoting

- Switch order of equations, moving offending element off diagonal
- $\left.\begin{array}{rcrcl} x_1 & + & x_2 & = & 2 \\ \epsilon\, x_1 & + & x_2 & = & 1 \end{array}\right\} \Rightarrow$, $x_2 = \frac{1-2\epsilon}{1-\epsilon}$ and $x_1 = 2 - x_2 = \frac{1}{1-\epsilon}$
- This is correct, even for small ϵ (or even $\epsilon = 0$)
- Compare size of diagonal (pivot) elements to ϵ
- Ratio of first row of Vandermonde matrix $= 1 : 2^{n-1}$

Issue is relative size, not absolute size.

Scaled Partial Pivoting

- Also called row pivoting (vs. column pivoting)
- Instability source: subtracting large values: $a_{k\,j}$ -= $a_{i\,j}\frac{a_{k\,i}}{a_{i\,i}}$
- W|o l.o.g.: n rows, and choosing first row
- Find $i \ni \forall$ rows $k \neq i$, $\forall$ columns $j > 1$: minimize $\left|a_{i\,j}\frac{a_{k\,1}}{a_{i\,1}}\right|$
- $O\left(n^3\right)$ calculations! $\therefore$ simplify (remove k), imagine: $a_{k\,1} = 1$
- $\therefore$ find $i \ni \forall$ columns $j > 1$: $\min_i \left|\frac{a_{i\,j}}{a_{i\,1}}\right|$
- Still $^{1)}O\left(n^2\right)$ calculations, $^{2)}$how to minimize each row?
- Find i: $\min_i \frac{\max_j |a_{i\,j}|}{|a_{i\,1}|}$, or: $\max_i \frac{|a_{i\,1}|}{\max_j |a_{i\,j}|}$

Linear Systems

- Introduction
- Naive Gaussian Elimination
- Limitations

$\Rightarrow$ Operation Counts

- Additional Notes

How Much Work on A?

- Real life: crowd estimation costs? (will depend on accuracy)
- Counting $\times$ and $\div$ (i.e., long operations) only
- Pivoting: row decision amongst k rows $= k$ ratios
- First row:
    - n ratios (for choice of pivot row)
    - $n - 1$ multipliers
    - $(n-1)^2$ multiplications

  total: n^2 operations
- $\therefore$ forward elimination operations (for large n)

$$\sum_{k=2}^{n} k^2 = \frac{n}{6}(n+1)(2n+1) - 1 \approx \frac{n^3}{3}$$

How about the work on b?

Rest of the Work

- Forward elimination work on RHS: $\sum_{k=2}^{n} (k-1) = \frac{n(n-1)}{2}$
- Backward substitution: $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$
- Total: n^2 operations
- $O(n)$ fewer operations than forward elimination on A
- Important for multiple RHSs known from the start
  * do not repeat $O\left(n^3\right)$ work for each
  * rather, line them up, and process simultaneously

Can we do better at times?

Sparse Systems

$$\begin{pmatrix} \times & \times & 0 & \cdots & & \cdots & 0 \\ \times & \times & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & & & & \\ \vdots & \ddots & & & & \ddots & \vdots \\ & & & & & \ddots & 0 \\ \vdots & & & \ddots & \ddots & \ddots & \times \\ 0 & \cdots & & \cdots & 0 & \times & \times \end{pmatrix}$$

- Above, e.g., tridiagonal system (half bandwidth = 1)
- Opportunities for savings
  - ∗ storage
  - ∗ computations
- Both are $O(n)$

Linear Systems

- Introduction
- Naive Gaussian Elimination
- Limitations
- Operation Counts

$\Rightarrow$ Additional Notes

Pivot-Free Guarantee

- When are we guaranteed non-zero pivots?
- Diagonal dominance (just like it sounds):

$$|a_{i\,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|a_{i\,j}\right|, \quad i = 1, \ldots, n$$

- (Or "$>$" in one row, and "$\geq$" in remaining)
- Many finite difference and finite element problems $\Rightarrow$ diagonally dominant systems

Occurs often enough to justify individual study.

LU Decomposition

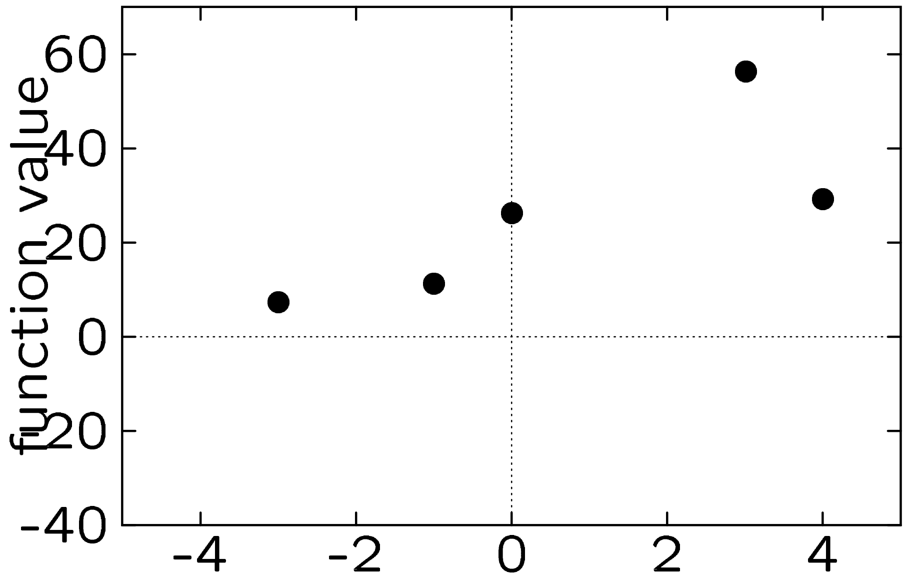
- E.g.: same A, many b's of time-dependent problem
  - $*$ not all b's are known from the start
- Want $A = LU$ for decreased work later
- Then define y: $L \underbrace{Ux}_{\equiv y} = b$
  - $*$ solve $Ly = b$ for y
  - $*$ solve $Ux = y$ for x
- U is upper triangular, result of Gaussian elimination
- L is unit lower triangular, 1's on diagonal and Gaussian multipliers below
- For small systems, verify (even by hand): $A = LU$

Each new RHS is n^2 work, instead of $O\left(n^3\right)$

Approximation by Splines

- $\Rightarrow$ Motivation
- Linear Splines
- Quadratic Splines
- Cubic Splines
- Summary

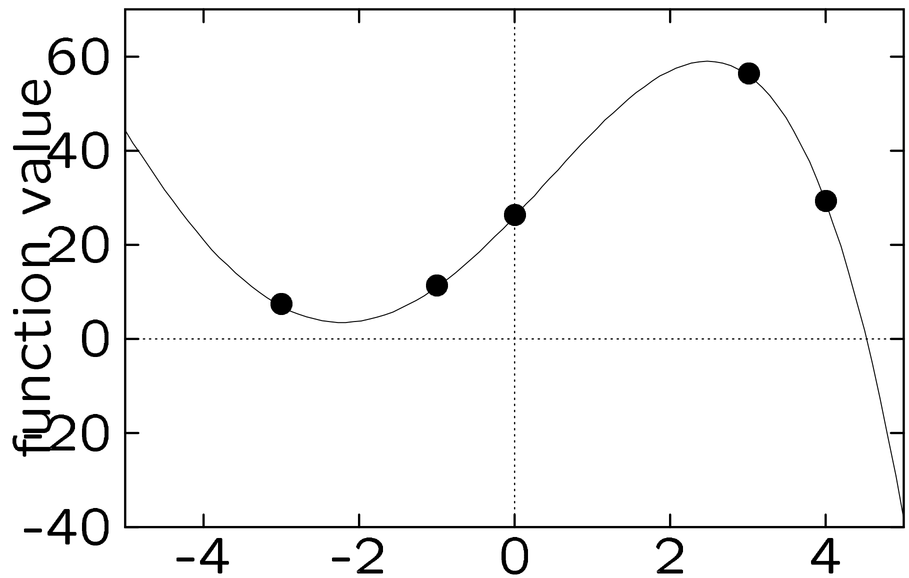
Motivation



- Given: set of *many* points, or perhaps very involved function
- Want: simple representative function for analysis or manufacturing

Any suggestions?

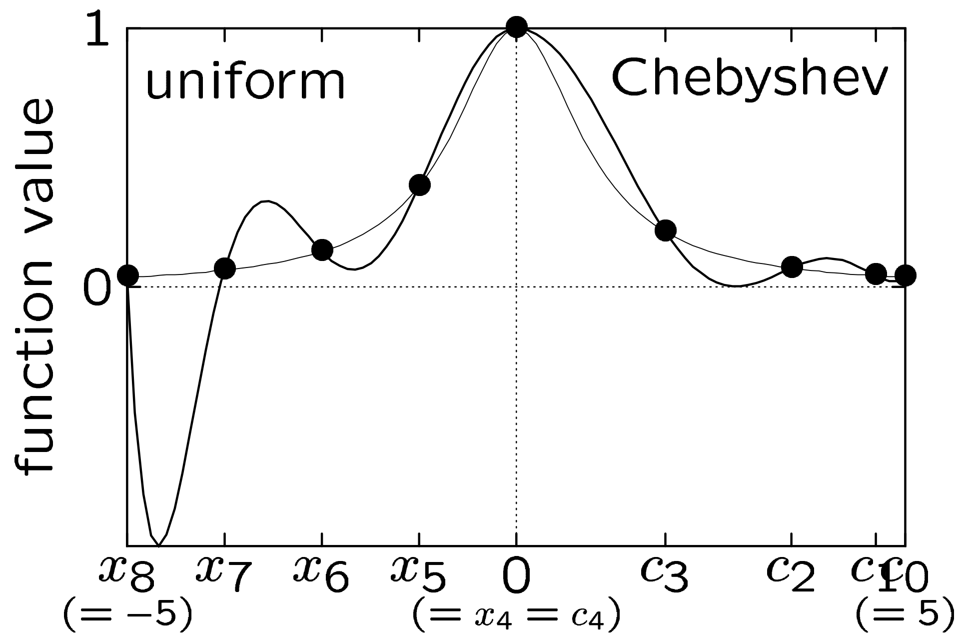
Let's Try Interpolation



Disadvantages:

- Values outside x-range diverge quickly (interp(10) $= -1592$)
- Numerical instabilities of high-degree polynomials

Runge Function—Two Interpolations



More disadvantages:

- Within x-range, often high oscillations
- Even Chebyshev points $\Rightarrow$ often uncharacteristic oscillations

Splines

Given domain $[a, b]$, a spline $S(x)$

- Is defined on entire domain

- Provides a certain amount of smoothness

- $\exists$ partition of "knots" (= where spline can change form)

$$\{a = t_0, t_1, t_2, \ldots, t_n = b\}$$

  such that

$$S(x) = \begin{cases} S_0(x), & x \in [t_0, t_1], \\ S_1(x), & x \in [t_1, t_2], \\ \vdots & \vdots \\ S_{n-1}(x), & x \in [t_{n-1}, t_n] \end{cases}$$

  is *piecewise* polynomial

Interpolatory Splines

- Note: splines *split up* range $[a, b]$
  - $*$ opposite of CTR $\rightarrow$ CSR $\rightarrow$ GQ development
- "Spline" implies no interpolation, not even any y-values
- If given points

$$\{(t_0, y_0), (t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)\}$$

"interpolatory spline" traverses these as well

Splines = nice, analytical functions

Approximation by Splines

- Motivation
- ⇒ Linear Splines
- Quadratic Splines
- Cubic Splines
- Summary

Linear Splines

Given domain $[a, b]$, a linear spline $S(x)$

- Is defined on entire domain

- Provides continuity, i.e., is $C^0[a, b]$

- $\exists$ partition of "knots"

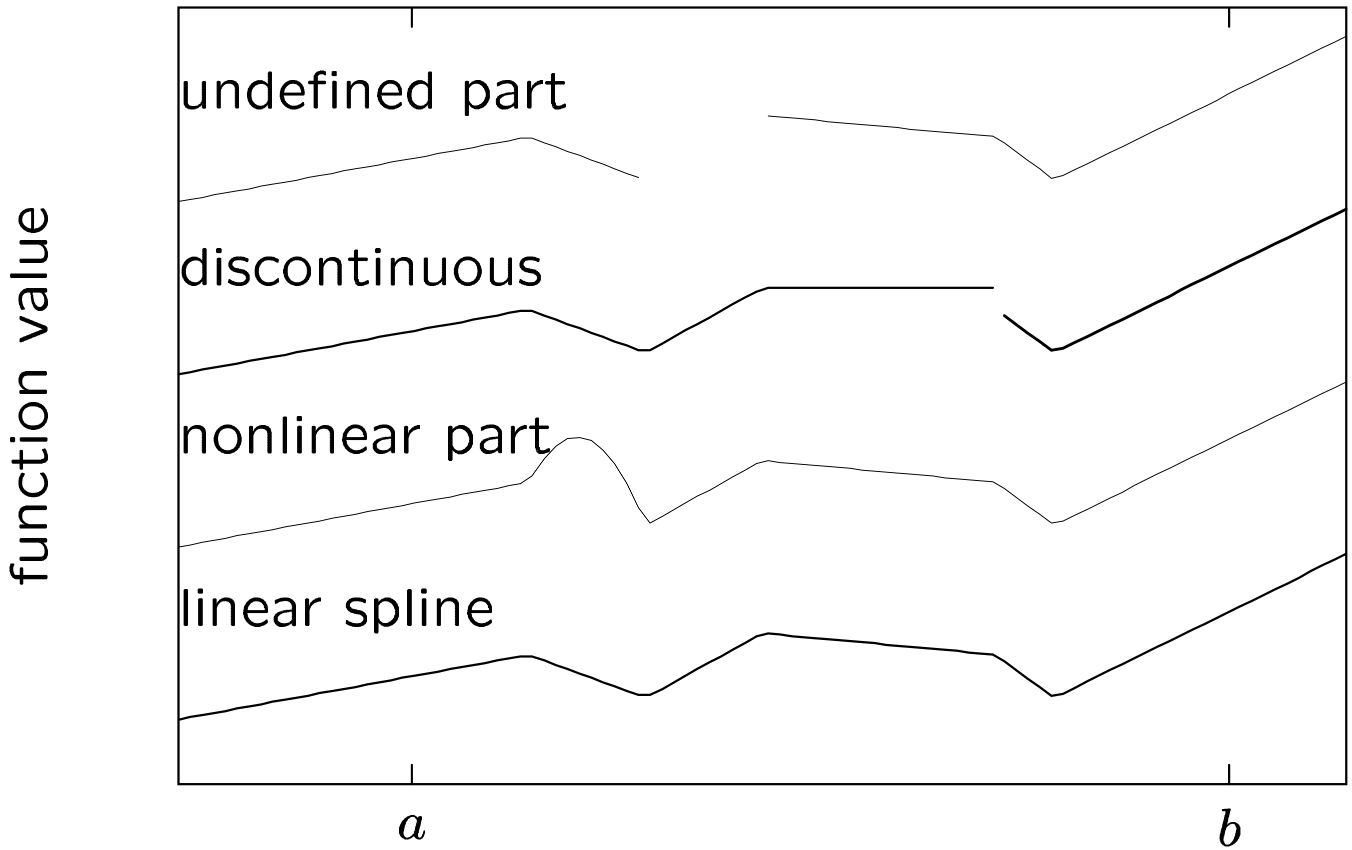
$$\{a = t_0, t_1, t_2, \ldots, t_n = b\}$$

  such that

$$S_i(x) = a_i\, x + b_i \in P_1\Big(\big[t_i, t_{i+1}\big]\Big), \quad i = 0, \ldots, n-1$$

Recall: no y-values or interpolation yet

Linear Spline—Examples



- Definition outside of $[a, b]$ is arbitrary

Interpolatory Linear Splines

- Given points

$$\{(t_0, y_0), (t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)\}$$

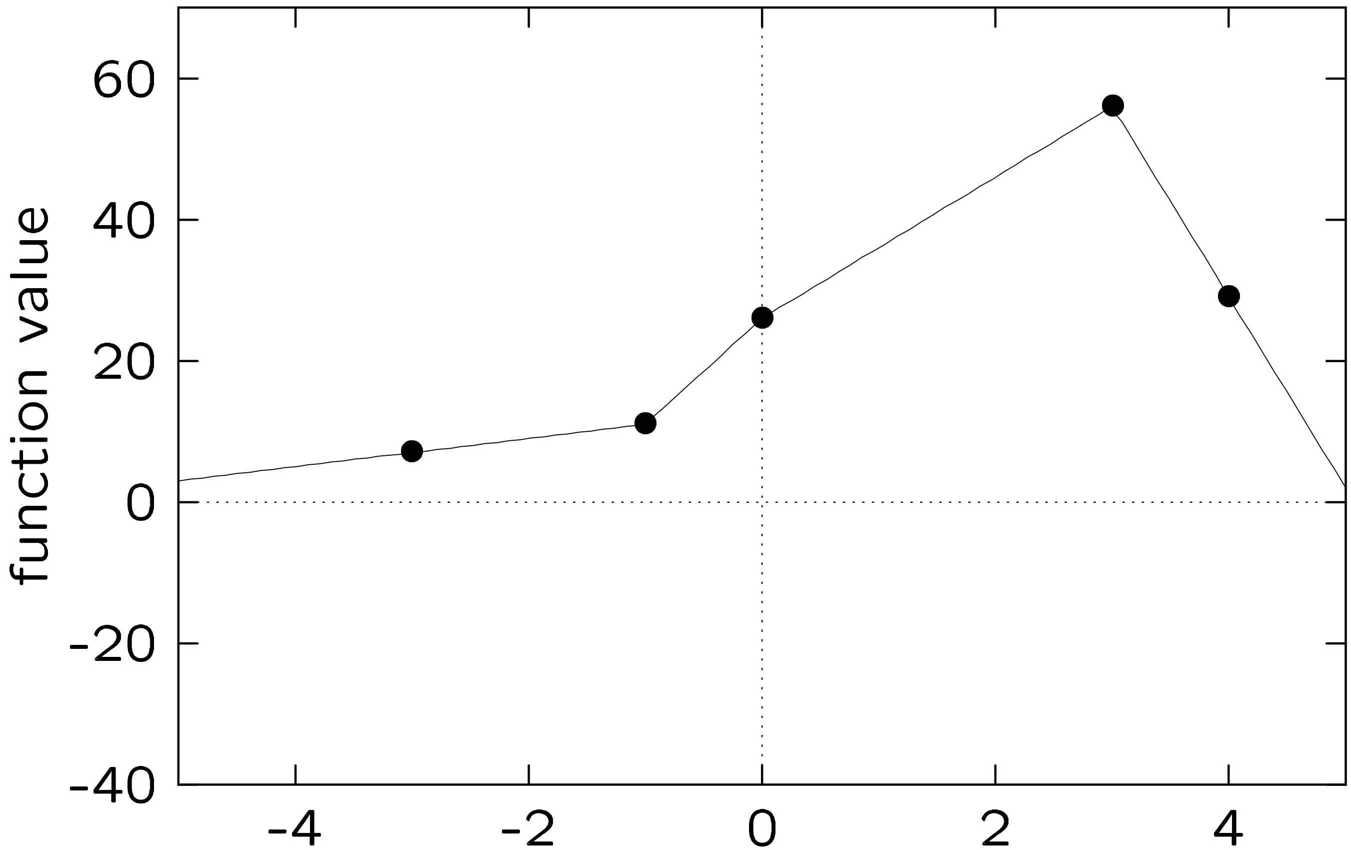
  spline must interpolate as well

- Are the $S_i(x)$ (with no additional knots) unique?
  * Coefficients: $a_i\, x + b_i$, $i = 0, \ldots, n-1 \Rightarrow$ $\boxed{\text{total} = 2n}$
  * Conditions: 2 prescribed interpolation points for $S_i(x)$, $i = 0, \ldots, n-1$ (includes continuity condition) $\Rightarrow$ $\boxed{\text{total} = 2n}$

- Obtain

$$S_i(x) = a_i\, x + (y_i - a_i\, t_i), \quad a_i = \frac{y_{i+1} - y_i}{t_{i+1} - t_i}, \quad i = 0, \ldots, n-1$$

Interpolatory Linear Splines—Example



Discontinuous derivatives at knots are unpleasing, . . .

Approximation by Splines

- Motivation
- Linear Splines
- ⇒ Quadratic Splines
- Cubic Splines
- Summary

Quadratic Splines

Given domain $[a, b]$, a quadratic spline $S(x)$

- Is defined on entire domain
- Provides continuity of zeroth and first derivatives, i.e., is $C^1[a, b]$
- $\exists$ partition of "knots"

$$\{a = t_0, t_1, t_2, \ldots, t_n = b\}$$

  such that

$$S_i(x) = a_i\, x^2 + b_i\, x + c_i \in P_2\Big(\big[t_i, t_{i+1}\big]\Big), \quad i = 0, \ldots, n-1$$

Again no y-values or interpolation yet

Quadratic Spline—Example

$$f(x) = \begin{cases} x^2, & x \le 0, \\ -x^2, & 0 \le x \le 1, \\ 1-2x, & x \ge 1, \end{cases} \qquad f(x) \stackrel{?}{=} \text{quadratic spline}$$

- Defined on domain $(-\infty, \infty)$ $\surd$
- Continuity (clearly okay away from $x = 0$ and 1):
  - ∗ Zeroth derivative:
    - ⋆ $f\left(0^-\right) = f\left(0^+\right) = 0$
    - ⋆ $f\left(1^-\right) = f\left(1^+\right) = -1$
  - ∗ First derivative:
    - ⋆ $f'\left(0^-\right) = f'\left(0^+\right) = 0$
    - ⋆ $f'\left(1^-\right) = f'\left(1^+\right) = -2$ $\surd$
- Each part of $f(x)$ is $\in P_2$ $\surd$

Interpolatory Quadratic Splines

- Given points

$$\{(t_0, y_0), (t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)\}$$

  spline must interpolate as well

- Are the $S_i(x)$ unique (same knots)?
  - Coefficients: $a_i\,x^2 + b_i\,x + c_i,\ \ i = 0, \ldots, n-1 \Rightarrow$ $\boxed{\text{total} = 3n}$
  - Conditions:
    - ⋆ 2 prescribed interpolation points for $S_i(x)$, $i = 0, \ldots, n-1$ (includes continuity of function condition)
    - ⋆ $(n-1)$ C^1 continuities

  $\Rightarrow$ $\boxed{\text{total} = 3n - 1}$

Interpolatory Quadratic Splines (cont.)

- Underdetermined system $\Rightarrow$ need to add one condition
- Define (as yet to be determined) $z_i = S'(t_i), \;\; i = 0, \ldots, n$
- Write

$$S_i(x) = \frac{z_{i+1} - z_i}{2\left(t_{i+1} - t_i\right)}(x - t_i)^2 + z_i(x - t_i) + y_i$$

therefore

$$S_i'(x) = \frac{z_{i+1} - z_i}{t_{i+1} - t_i}(x - t_i) + z_i$$

- Need to
  - ∗ verify continuity and interpolatory conditions
  - ∗ determine z_i

Checking Interpolatory Quadratic Splines

Check four continuity (and interpolatory) conditions:

$$\text{(i)}\quad S_i(t_i) \overset{\surd}{=} y_i \qquad \text{(iii)}\quad S_i'(t_i) \overset{\surd}{=} z_i$$

$$\text{(ii)}\quad S_i\big(t_{i+1}\big) = \text{(below)} \qquad \text{(iv)}\quad S_i'\big(t_{i+1}\big) \overset{\surd}{=} z_{i+1}$$

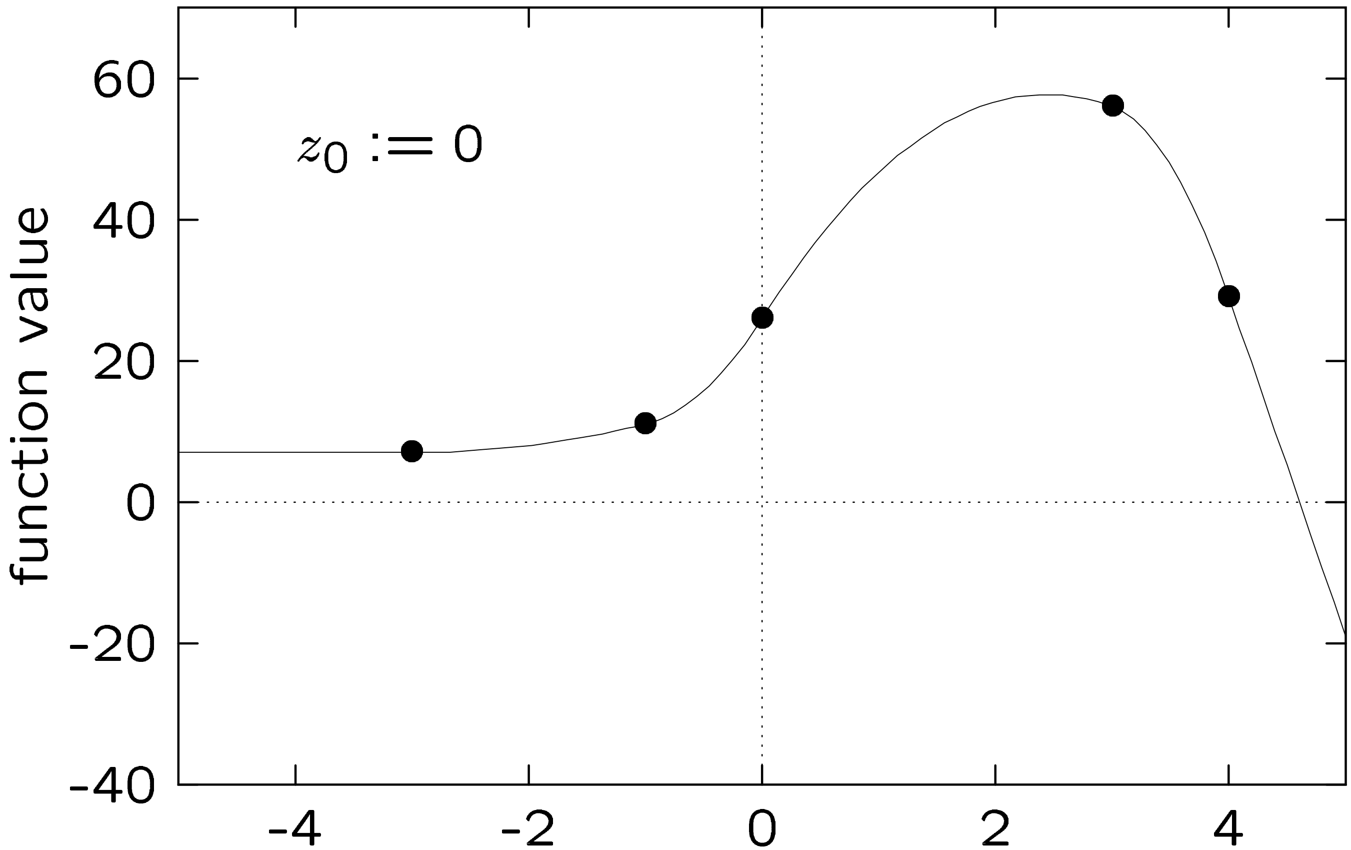
$$\begin{aligned}
\text{(ii)}\quad S_i\big(t_{i+1}\big) &= \frac{z_{i+1} - z_i}{2}\big(t_{i+1} - t_i\big) + z_i\big(t_{i+1} - t_i\big) + y_i \\
&= \frac{z_{i+1} + z_i}{2}\big(t_{i+1} - t_i\big) + y_i \\
&\overset{\text{set}}{=} y_{i+1}
\end{aligned}$$

therefore (n equations, $n+1$ unknowns)

$$z_{i+1} = 2\frac{y_{i+1} - y_i}{t_{i+1} - t_i} - z_i, \quad i = 0, \ldots, n-1$$

Choose any 1 z_i and the remaining n are determined.

Interpolatory Quadratic Splines—Example



Okay, but discontinuous curvature at knots, . . .

Approximation by Splines

- Motivation
- Linear Splines
- Quadratic Splines
- ⇒ Cubic Splines
- Summary

Cubic Splines

Given domain $[a, b]$, a cubic spline $S(x)$

- Is defined on entire domain

- Provides continuity of zeroth, first and second derivatives, i.e., is $C^2[a, b]$

- $\exists$ partition of "knots"

$$\{a = t_0, t_1, t_2, \ldots, t_n = b\}$$

  such that for $i = 0, \ldots, n-1$

$$S_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i \in P_3\Big(\big[t_i, t_{i+1}\big]\Big),$$

In general: spline of degree k ... C^{k-1} ... P_k ...

Why Stop at $k = 3$?

- Continuous curvature is visually pleasing
- Usually little numerical advantage to $k > 3$
- Technically, odd k's are better for interpolating splines
- *Natural* (defined later) cubic splines
  - ∗ "best" in an analytical sense (stated later)

Interpolatory Cubic Splines

- Given points

$$\{(t_0, y_0), (t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)\}$$

  spline must interpolate as well

- Are the $S_i(x)$ unique (same knots)?
  - ∗ Coefficients: $a_i\,x^3 + b_i\,x^2 + c_i\,x + d_i, \;\; i = 0, \ldots, n-1 \Rightarrow$ $\boxed{\text{total} = 4n}$
  - ∗ Conditions:
    - ⋆ 2 prescribed interpolation points for $S_i(x)$, $i = 0, \ldots, n-1$ (includes continuity of function condition)
    - ⋆ $(n-1)\ C^1 + (n-1)\ C^2$ continuities

  $\Rightarrow \boxed{\text{total} = 4n - 2}$

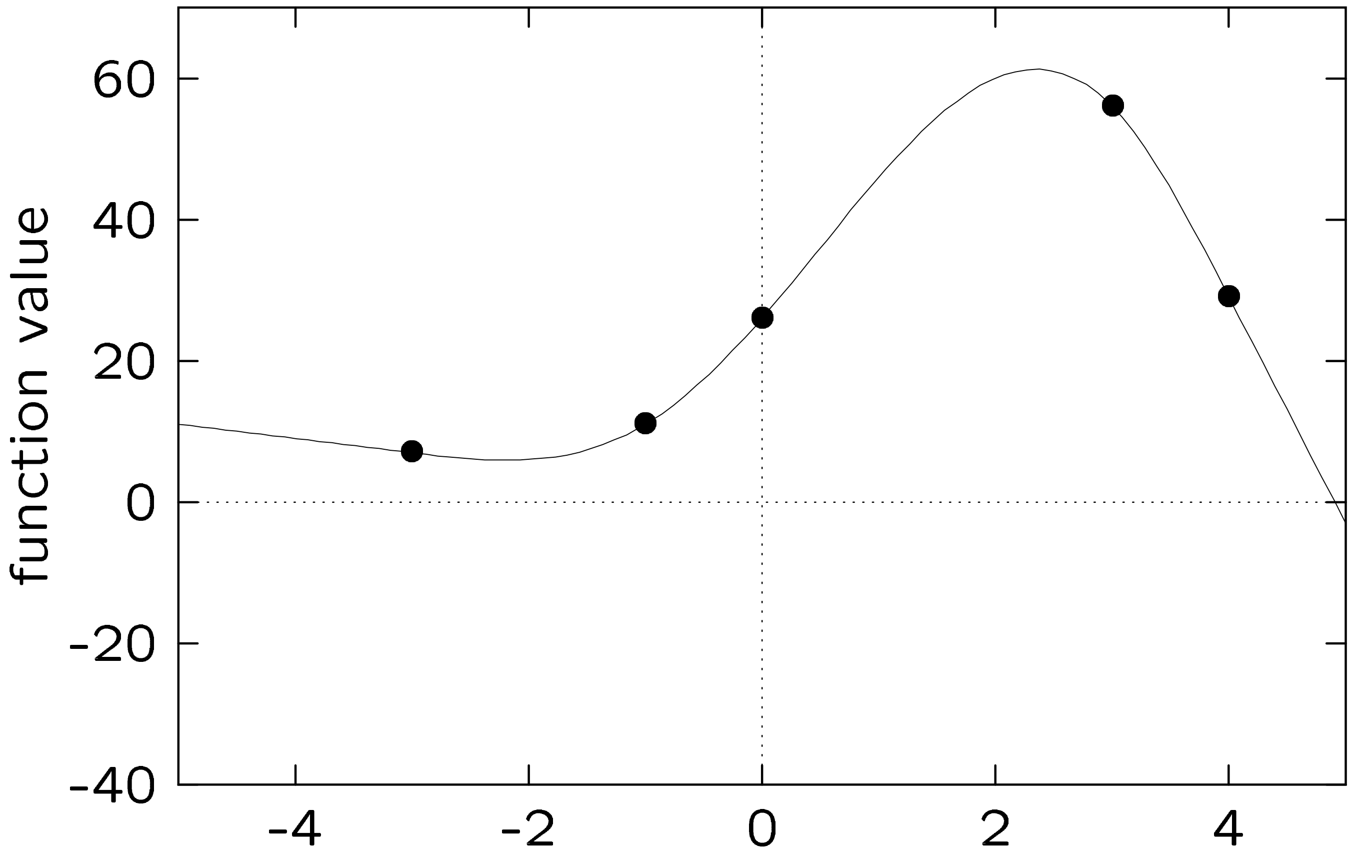
Interpolatory Cubic Splines (cont.)

- Underdetermined system $\Rightarrow$ need to add two conditions

- Natural cubic spline
  - add: $S''(a) = S''(b) = 0$
  - Assumes straight lines (i.e., no more constraints) outside of $[a, b]$
  - Imagine bent beam of ship hull
  - Defined for non-interpolatory case as well

- Required matrix calculation for S_i definitions
  - Linear: independent $a_i = \frac{y_{i+1} - y_i}{t_{i+1} - t_i}$ $\Rightarrow$ diagonal
  - Quadratic: two-term z_i definition $\Rightarrow$ bidiagonal
  - Cubic: ... $\Rightarrow$ tridiagonal

Interp. Natural Cubic Splines—Example



Now the curvature is continuous as well.

Optimality of Natural Cubic Spline

- Theorem: If
  - $f \in C^2[a,b]$,
  - knots: $\{a = t_0, t_1, t_2, \ldots, t_n = b\}$
  - interpolation points: $(t_i, y_i) : y_i = f(t_i),\ \ i = 0, \ldots, n$
  - $S(x)$ is the natural cubic spline which interpolates $f(x)$

  then

$$\int_a^b \left[S''(x)\right]^2 dx \leq \int_a^b \left[f''(x)\right]^2 dx$$

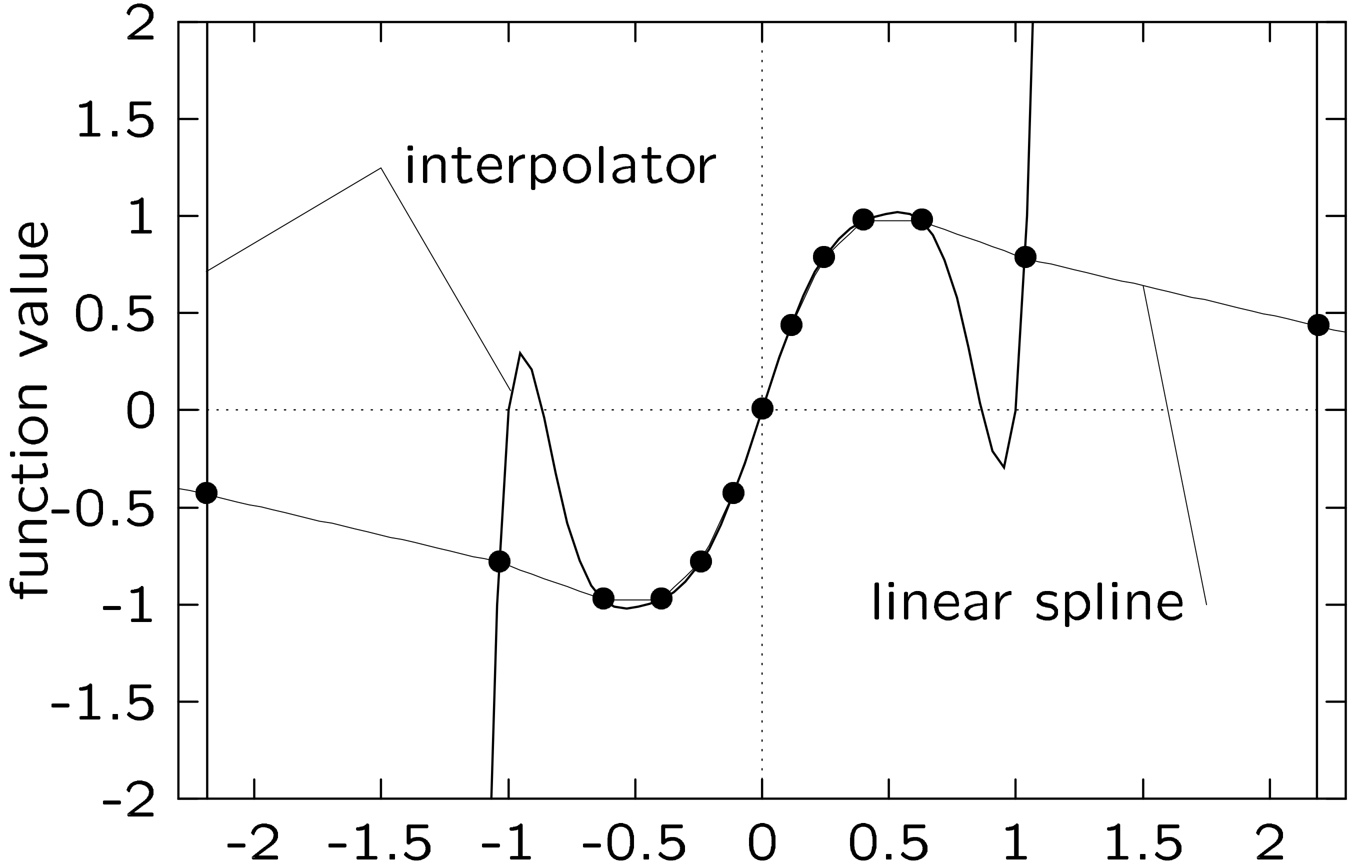
- Bottom line
  - average curvature of $S \leq$ that of f
  - compare with interpolating polynomial

Approximation by Splines

- Motivation
- Linear Splines
- Quadratic Splines
- Cubic Splines

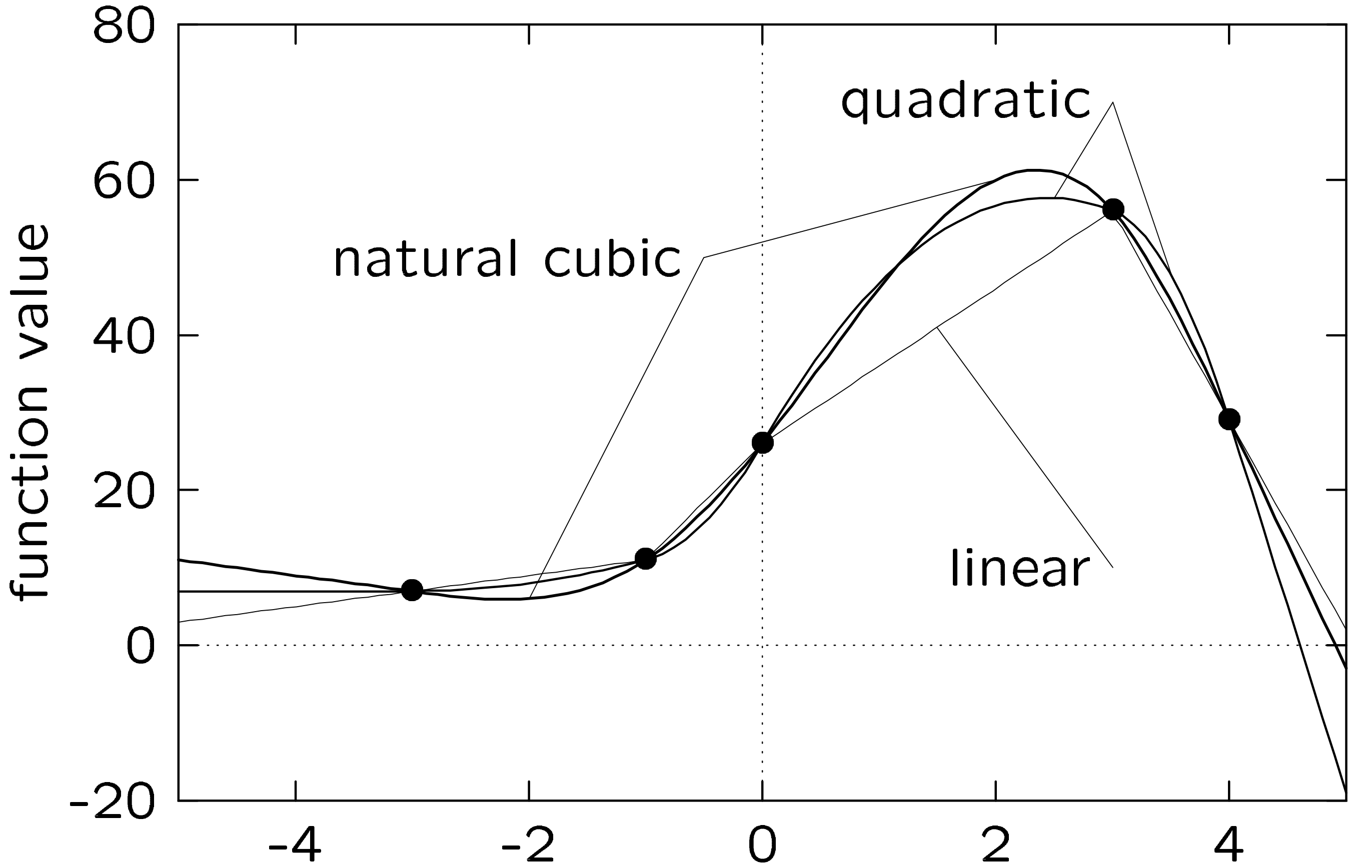
$\Rightarrow$ Summary

Interpolation vs. Splines—Serpentine Curve



Vs. oscillatory interpolator—even linear spline is better.

Three Splines



Increased smoothness with increase of degree.

Ordinary Differential Equations

$\Rightarrow$ Introduction
- Euler Method
- Higher Order Taylor Methods
- Runge-Kutta Methods
- Summary

Ordinary Differential Equation—Definition

- ODE = an equation
  - ∗ involving one or more derivatives of $x(t)$
  - ∗ $x(t)$ is unknown and the desired target
  - ∗ somewhat opposite of numerical differentiation

- E.g.: $(x''')^{\frac{3}{7}}(t) + 37\, t\, e^{x^2(t)} \sin \sqrt[4]{x'(t)} - \log \frac{1}{t} = 42 \quad \Rightarrow$
  Which $x(t)$'s fulfill this behavior?

- "Ordinary" (vs. "partial") = *one* independent variable t

- "Order" = highest (composition of) derivative(s) involved

- "Linear" = derivatives, including zeroth, appear in linear form

- "Homogeneous" = all terms involve some derivative (including zeroth)

Analytical Approach

- Good luck with previous equation, but others . . .
- Shorthand: $x = x(t)$, $x' = \frac{d(x(t))}{dt}$, $x'' = \frac{d^2(x(t))}{dt^2}$, . . .
- Analytically solvable
  * $x' - x = e^t \quad \Rightarrow \quad x(t) = t\,e^t + c\,e^t$
  * $x'' + 9x = 0 \quad \Rightarrow \quad x(t) = c_1 \sin 3t + c_2 \cos 3t$
  * $x' + \frac{1}{2x} = 0 \quad \Rightarrow \quad x(t) = \sqrt{c - t}$
- c, c_1 and c_2 are arbitrary constants
- Need more conditions/information to pin down constants
  * Initial value problems (IVP)
  * Boundary value problems (BVP)

Here: IVP for first-order ODE.

First-Order IVP

- General form:

$$x' = f(t, x), \;\; x(a) \text{ given}$$

- Note: non-linear, non-homogeneous

- Examples
  * $x' = x + 1, \;\; x(0) = 0 \quad \Rightarrow \quad x(t) = e^t - 1$
  * $x' = 6t - 1, \;\; x(1) = 6 \quad \Rightarrow \quad x(t) = 3t^2 - t + 4$
  * $x' = \frac{t}{x+1}, \quad x(0) = 0 \quad \Rightarrow \quad x(t) = \sqrt{t^2 + 1} - 1$

- Physically: e.g., t is time, x is distance and $f = x'$ is speed/velocity

Another optimistic scenario . . .

RHS Independence of x

- $f = f(t)$ but $f \neq f(x)$

- E.g.

$$\begin{cases} x' = 3t^2 - 4t^{-1} + \left(1 + t^2\right)^{-1} \\ x(5) = 17 \end{cases}$$

- Perform indefinite integral

$$x(t) = \int \frac{d(x(t))}{dt}\, dt = \int f(t)\, dt$$

- Obtain

$$\begin{cases} x(t) = t^3 - 4 \ln t + \arctan t + C \\ C = 17 - 5^3 + 4 \ln 5 - \arctan 5 \end{cases}$$

And now for the bad news . . .

Numerical Techniques

- Source of need
  - ∗ Usually analytical solution is not known
  - ∗ Even if known, perhaps very complicated, expensive to compute

- Numerical techniques
  - ∗ Generate a table of values for $x(t)$
  - ∗ Usually equispaced in t, stepsize $= h$
  - ∗ ⚠ with small h, and far from initial value
    roundoff error can accumulate and kill

Ordinary Differential Equations

- Introduction

⇒ Euler Method

- Higher Order Taylor Methods
- Runge-Kutta Methods
- Summary

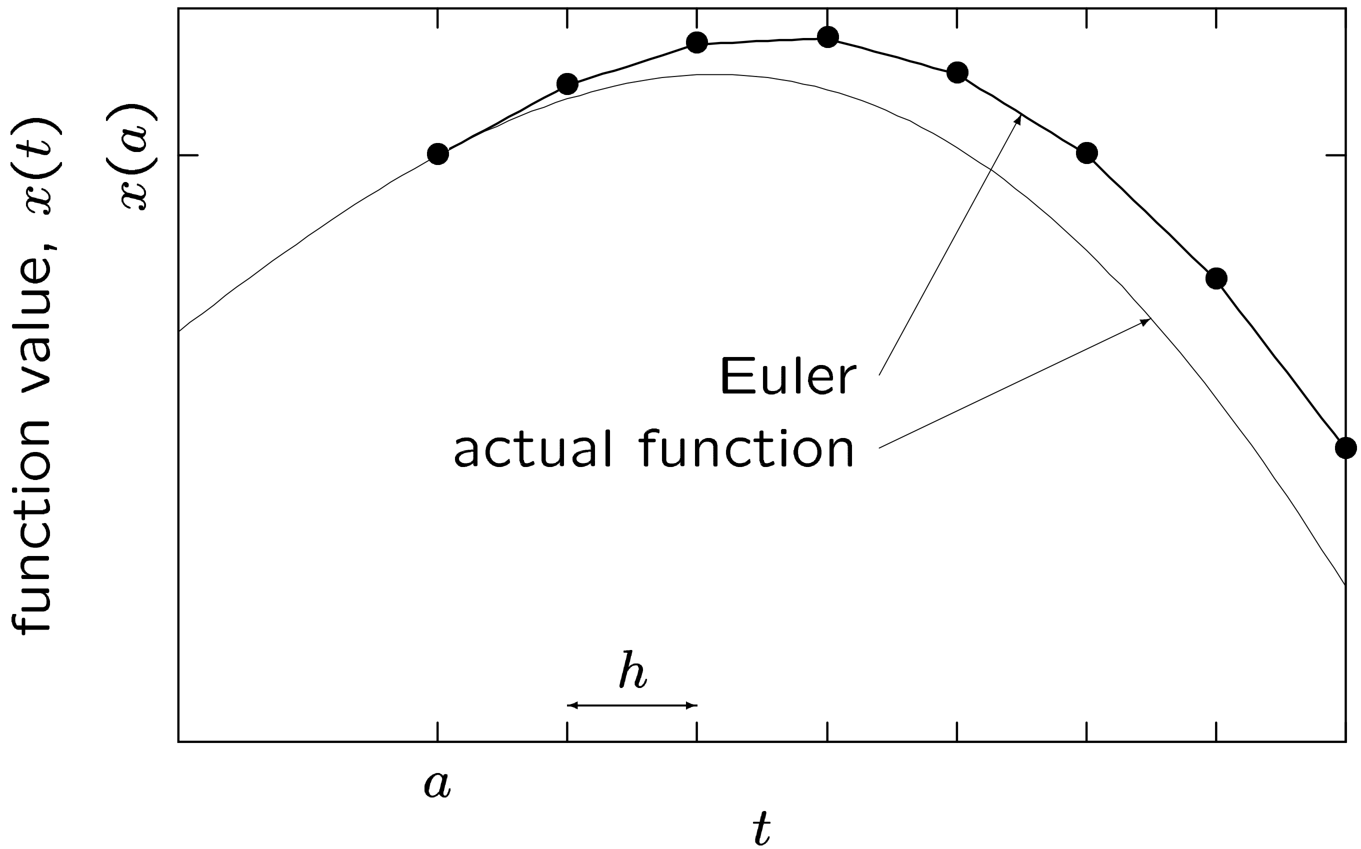
Euler Method

- First-order IVP: given $x' = f(t, x)$, $x(a)$, want $x(b)$
- Use first 2 terms of Taylor series (i.e., $n = 1$) to get from $x(a)$ to $x(a+h)$

$$x(a+h) = x(a) + h\,\underbrace{x'(a)}_{\text{use } f(a,x(a))} + \overbrace{O\big(h^2\big)}^{\text{truncation error}}$$

- Repeat to get from $x(a+h)$ to $x(a+2h)$, ...
- Total $n = \frac{b-a}{h}$ steps until $x(b)$
- Note: units of time/distance/speed are consistent

Euler Method—Example



- When will the slopes match up at the points?

Okay, but not great. What is the accuracy?

Euler Method—Pros and Cons

- Note: straight lines connecting points
  - $*$ from Euler construction (linear in h)
  - $*$ can be used for subsequent linear interpolation

- Advantages
  - $*$ Accurate early on: $O\left(h^2\right)$ for first step
  - $*$ Only need to calculate given function $f(t, x(t))$
  - $*$ Only one evaluation of $f(t, x(t))$ needed

- Disadvantages
  - $*$ Pretty inaccurate at b
  - $*$ Cumulative truncation error: $n \times O\left(h^2\right) = O(h)$
  - $*$ This is aside from (accumulative) roundoff error

How about more terms of the Taylor series?

Ordinary Differential Equations

- Introduction
- Euler Method

$\Rightarrow$ Higher Order Taylor Methods

- Runge-Kutta Methods
- Summary

Taylor Method of Order 4

- First-order IVP: given $x' = f(t,x)$, $x(a)$, want $x(b)$

- Use first 5 terms of Taylor series (i.e., $n = 4$) to get from $x(a)$ to $x(a+h)$

$$x(a+h) = x(a) + h\underbrace{x'(a)}_{\text{use } f(a,x(a))} + \frac{h^2}{2!}x''(a) + \frac{h^3}{3!}x'''(a) + \frac{h^4}{4!}x^{(iv)}(a) + O\left(h^5\right)$$

- Use f', f'' and f''' for x'', x''' and $x^{(iv)}$, respectively

- Repeat to get from $x(a+h)$ to $x(a+2h)$, . . .

- Note: units of time/distance/speed still are consistent

Order 4 is a standard order used.

Taylor Method—Numerical Example

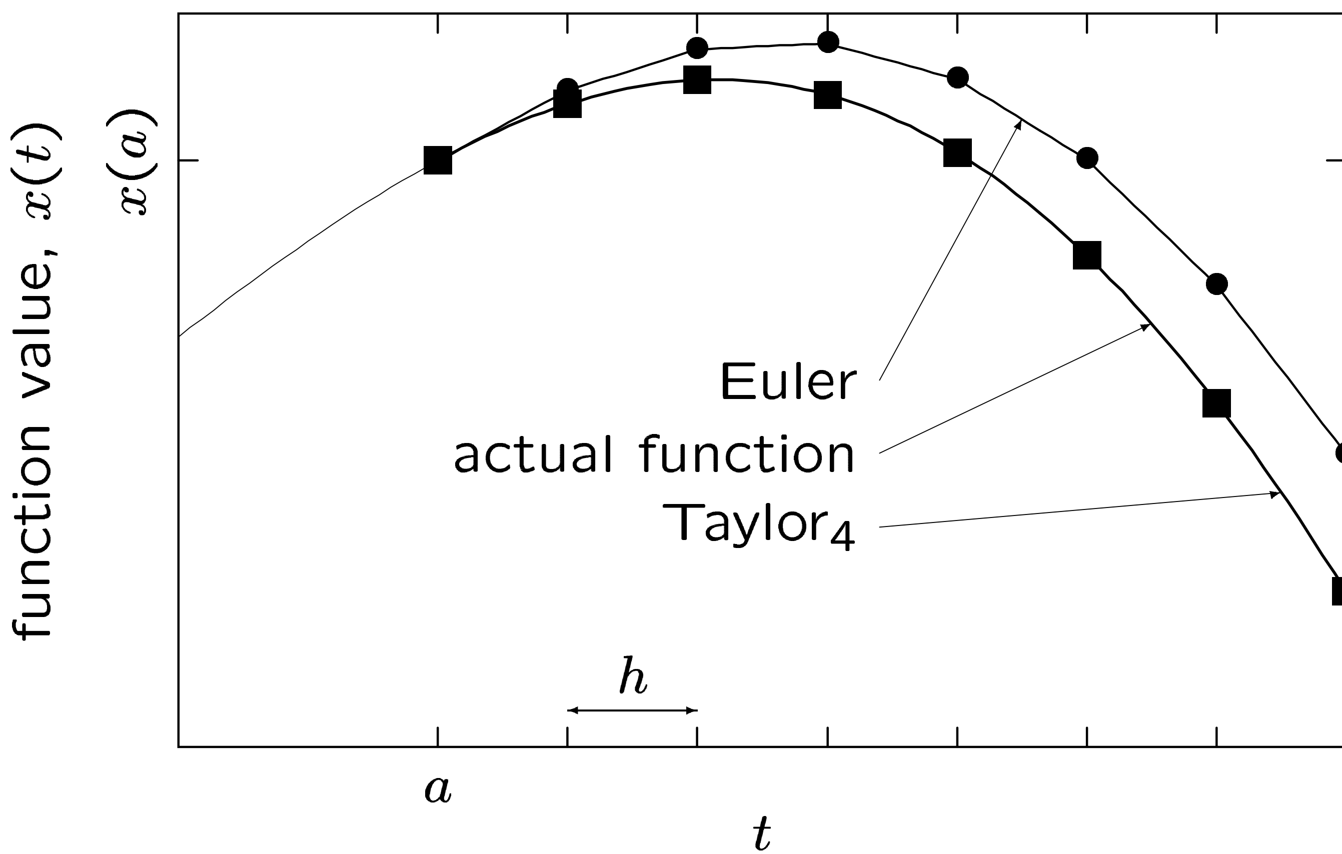
- First-order IVP: $x' = 1 + x^2 + t^3$, $x(1) = -4$, want $x(2)$
- Derivatives of $f(t, x)$

$$\begin{aligned} x'' &= 2x\,x' + 3t^2 \\ x''' &= 2x\,x'' + 2\left(x'\right)^2 + 6t \\ x^{(iv)} &= 2x\,x''' + 6x'\,x'' + 6 \end{aligned}$$

- Solution values of $x(2)$, $n = 100$
  - actual: 4.3712 (5 significant digits)
  - Euler: 4.2358541
  - Taylor$_4$: 4.3712096

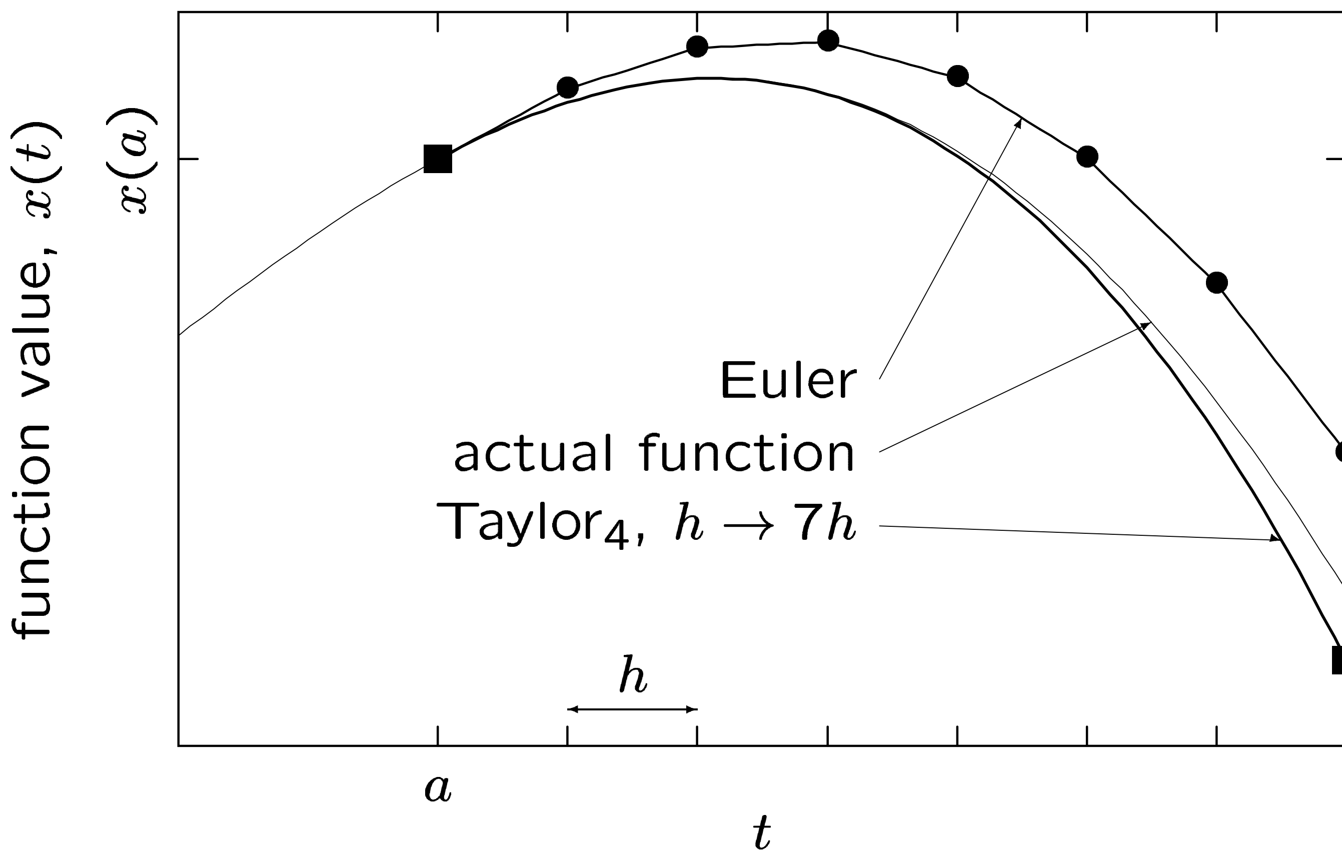
How about the earlier graphed example?

Taylor Method of Order 4—Example



Single step truncation error of $O\left(h^5\right) \Rightarrow$ excellent match.

Taylor Method of Order 4—Larger Step



Even single Taylor step beats Euler.

Taylor Method—Pros and Cons

- Note: graphs connecting points: from construction (P_4 in h)
- Advantages
  - ∗ Very accurate
  - ∗ Cumulative truncation error: $n \times O\big(h^5\big) = O\big(h^4\big)$
- Disadvantages
  - ∗ Need derivatives of $f(t, x(t))$ which might be
    - ⋆ analytically: difficult
    - ⋆ numerically: expensive—computationally and/or accuracy-wise
    - ⋆ just plain impossible
  - ∗ Four new evaluations each step (Euler was just one)

How to avoid the extra derivatives?

Ordinary Differential Equations

- Introduction
- Euler Method
- Higher Order Taylor Methods
- $\Rightarrow$ Runge-Kutta Methods
- Summary

Motivation

- We want to avoid calculating derivatives of $f(t, x(t))$
- Similar to Newton→secant motivation
- Also, recall different approaches for higher accuracy
  - ∗ Taylor series: more derivatives at one point
  - ∗ Numerical differentiation: more function evaluations, at various points
- Runge-Kutta (RK) of order m: for each step of size h
  - ∗ evaluate $f(t, x(t))$ at m interim stages
  - ∗ arrive at accuracy order similar to Taylor method of order m

Runge-Kutta Methods: RK2 and RK4

- Each $f(t, x(t))$ evaluation builds on previous
- Weighted average of evaluations produces $x(t+h)$
- Error for order m is $O\left(h^{m+1}\right)$ for each step of size h
- Note: units of time/distance/speed—okay

<u>RK2:</u>

$$x(t+h) = x(t) + \tfrac{1}{2}(F_1 + F_2)$$

$$\begin{cases} F_1 &= hf(t, x) \\ F_2 &= hf(t+h, x+F_1) \end{cases}$$

<u>RK4:</u>

$$x(t+h) = x(t) + \tfrac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4)$$

$$\begin{cases} F_1 &= hf(t, x) \\ F_2 &= hf\left(t + \tfrac{1}{2}h, x + \tfrac{1}{2}F_1\right) \\ F_3 &= hf\left(t + \tfrac{1}{2}h, x + \tfrac{1}{2}F_2\right) \\ F_4 &= hf(t+h, x+F_3) \end{cases}$$

Ordinary Differential Equations

- Introduction
- Euler Method
- Higher Order Taylor Methods
- Runge-Kutta Methods

$\Rightarrow$ Summary

Summary—First-Order IVP Solvers

- Complex and complicated IVPs require numerical methods
- Usually generate table of values, at constant stepsize h
- Euler: simple, but not too accurate
- High-order Taylor: very accurate, but requires derivatives of $f(t, x(t))$
- Runge-Kutta: same order of accuracy as Taylor, without derivative evaluations
- Error sources
  * Local truncation (of Taylor series approximation)
  * Local roundoff (due to finite precision)
  * Accumulations and combinations of previous two

Least Squares Method

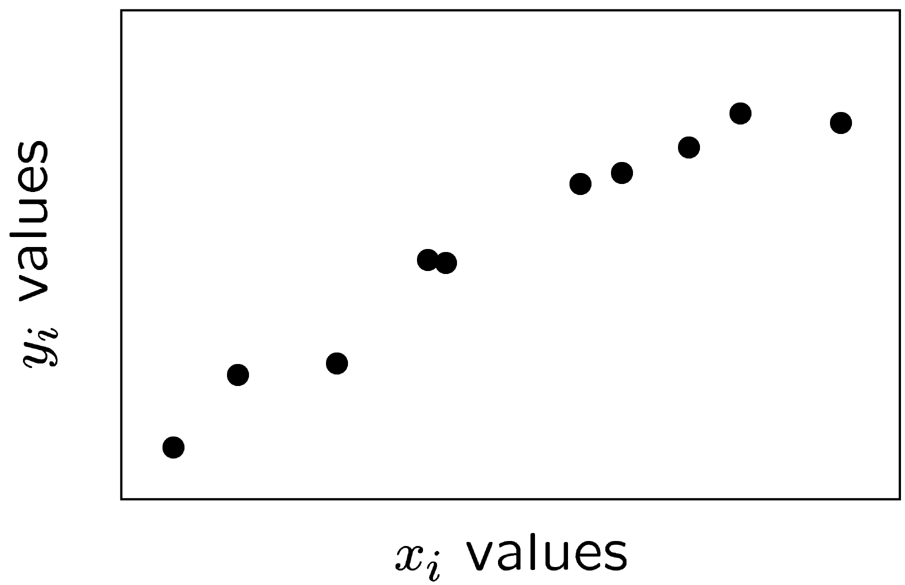
$\Rightarrow$ Motivation and Approach

- Linearly Dependent Data
- General Basis Functions
- Polynomial Regression
- Function Approximation

Source of Data

- Have the following tabulated data:

x	x_0	x_1	$\cdots$	x_m
y	y_0	y_1	$\cdots$	y_m



- E.g., data from experiment

- Assume known dependence, e.g. linear, i.e., $y = ax + b$

What a and b do we choose to represent the data?

Most Probable Line

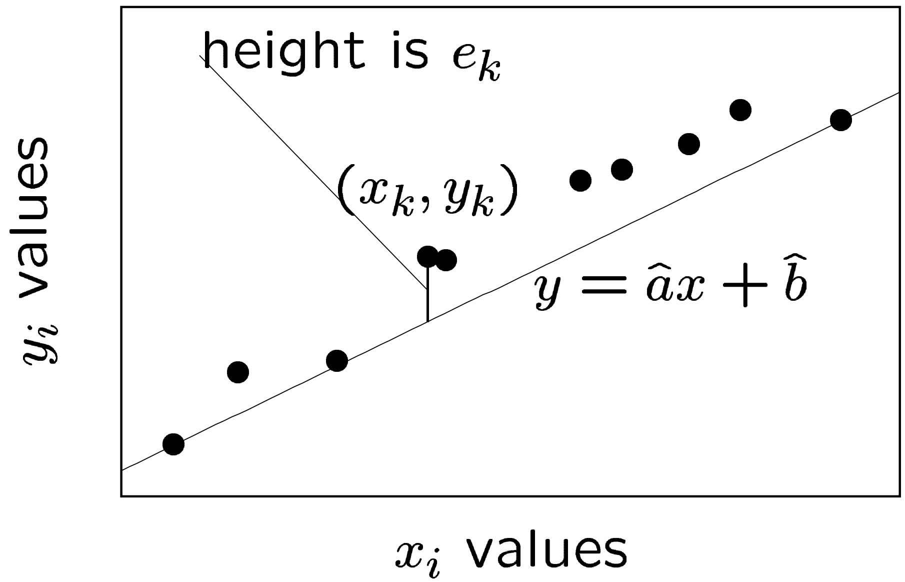
- For each point, consider the equation $y_i = ax_i + b$ with the two unknowns a and b
- One point $\Rightarrow \infty$ solutions
- Two points (different x_i) $\Rightarrow$ one unique solution
- $>$ two points $\Rightarrow$ in general no solution

$>$ two points $\Rightarrow$ What is most probable line?

Estimate Error

- Assume estimates $\hat{a}$ and $\hat{b} \Rightarrow$

  error at (x_k, y_k): $e_k = \hat{a}x_k + \hat{b} - y_k$



- Note:
  - $*$ vertical error, *not* distance to line (a harder problem)
  - $*$ $|e_k| \Rightarrow$ no preference to error direction

How do we minimize all of the $|e_k|$?

Vector Minimizations

Minimize:

- largest component: $\min_{a,b} \max_{0 \le k \le m} |e_k|$, "min-max"

- component sum: $\min_{a,b} \sum_{k=0}^{m} |e_k|$, linear programming

  Note: $|\cdot|$ won't allow errors to cancel

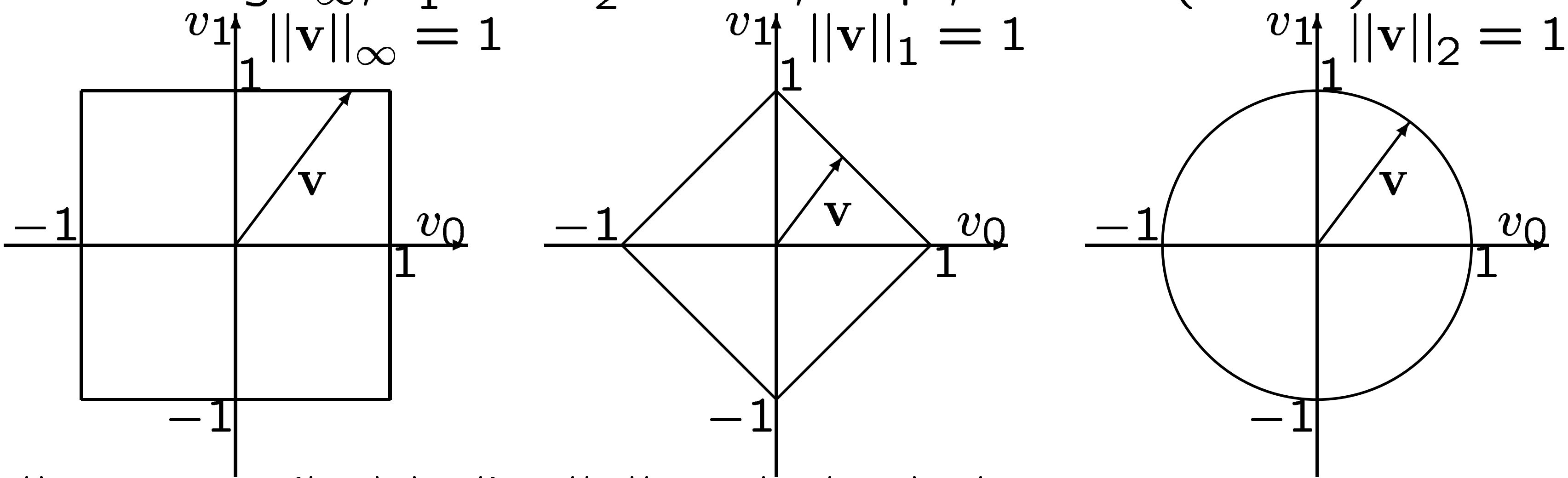
- component squared sum: $\min_{a,b} \underbrace{\sum_{k=0}^{m} e_k^2}_{\equiv \phi(a,b)}$, least squares

Why use least squares?

ℓ_p Norms

- Definition: $||\mathbf{v}||_p \equiv \left(\sum_{k=0}^{m} |v_k|^p \right)^{\frac{1}{p}}$

- Minimizing ℓ_∞, ℓ_1 and ℓ_2 norms, resp., in 2–D ($m = 1$):



$||\mathbf{v}||_\infty = \max(|v_0|, |v_1|)$ $\quad ||\mathbf{v}||_1 = |v_0| + |v_1|$ $\quad ||\mathbf{v}||_2 = \sqrt{v_0^2 + v_1^2}$

- Why use ℓ_2?
  - ∗ Can use calculus (see below)
  - ∗ If error is normally distributed $\Rightarrow$

    get maximum likelihood estimator

$\phi(a, b)$ Minimization

- How do we minimize $\phi(a,b) = \sum_{k=0}^{m} e_k^2$ wrt a and b?
- Standard calculus: $\dfrac{\partial \phi}{\partial a} \overset{\text{set}}{=} 0$ and $\dfrac{\partial \phi}{\partial b} \overset{\text{set}}{=} 0 \Rightarrow$
  two equations with two unknowns
- If dependence of y on a and b is linear (and consequently, dependence of $\phi(a,b)$ is quadratic) $\Rightarrow$
  minimization leads to linear system for a and b
  (*linear* least squares)
- Example also had linearly dependent data, i.e., y linear in x

Minimization of our example, . . .

Least Squares Method

- Motivation and Approach
- ⇒ Linearly Dependent Data
- General Basis Functions
- Polynomial Regression
- Function Approximation

LLS for Linearly Dependent Data—Method

Function to minimize:

$$\phi(a,b) = \sum_{k=0}^{m} e_k^2 = \sum_{k=0}^{m} (ax_k + b - y_k)^2$$

lead to two differentiations:

$$2\sum_{k=0}^{m} (ax_k + b - y_k)x_k = 0, \text{ and } 2\sum_{k=0}^{m} (ax_k + b - y_k) = 0$$

or as a system of linear equations in a and b:

$$\begin{aligned}
\left(\sum_{k=0}^{m} x_k^2\right)a + \left(\sum_{k=0}^{m} x_k\right)b &= \left(\sum_{k=0}^{m} x_k\, y_k\right) \\
\left(\sum_{k=0}^{m} x_k\right)a + \quad (m+1)b &= \left(\sum_{k=0}^{m} y_k\right)
\end{aligned}$$

Coefficient matrix = cross-products of a and b coefficients.

LLS for Linearly Dependent Data—Solution

We obtain:

$$a = \frac{1}{d}\left[(m+1)\sum_{k=0}^{m} x_k\, y_k - \sum_{k=0}^{m} x_k \sum_{k=0}^{m} y_k\right]$$

and

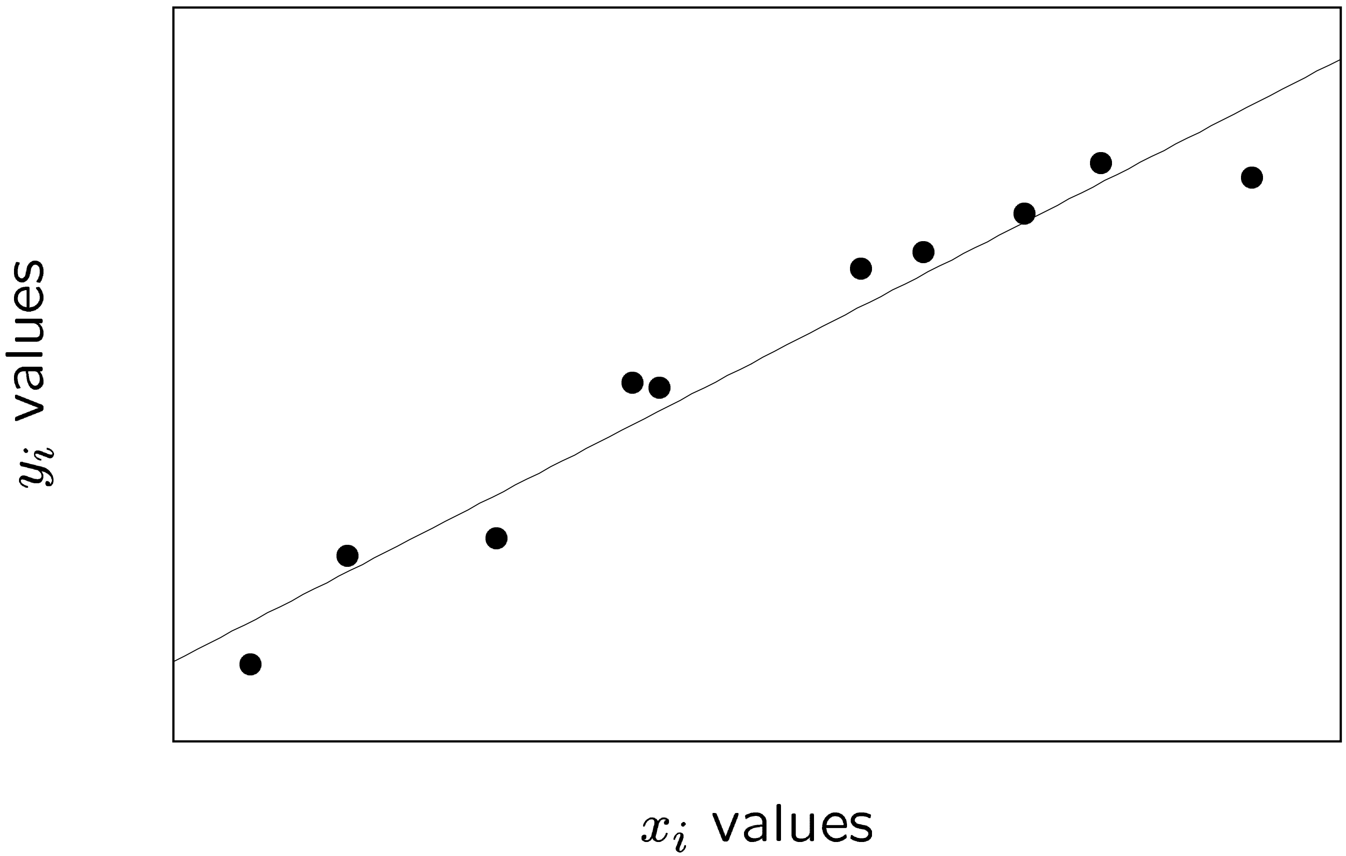
$$b = \frac{1}{d}\left(\sum_{k=0}^{m} x_k^2 \sum_{k=0}^{m} y_k - \sum_{k=0}^{m} x_k \sum_{k=0}^{m} x_k\, y_k\right)$$

where d is the determinant:

$$d = (m+1)\sum_{k=0}^{m} x_k^2 - \left(\sum_{k=0}^{m} x_k\right)^2$$

What does this look like?

LLS Solution for Sample Data



What about non-linearly dependent data?

Least Squares Method

- Motivation and Approach
- Linearly Dependent Data

$\Rightarrow$ General Basis Functions

- Polynomial Regression
- Function Approximation

Non-Linearly Dependent Data

- Linear least squares—for *linear* combination of any functions, e.g.:

$$y = a \ln x + b \cos x + c e^x$$

- Minimization of ϕ: three differentiations:

$$\frac{\partial \phi}{\partial a} \stackrel{\text{set}}{=} 0, \; \frac{\partial \phi}{\partial b} \stackrel{\text{set}}{=} 0 \text{ and } \frac{\partial \phi}{\partial c} \stackrel{\text{set}}{=} 0$$

- Elements of matrix: sums of cross-products of functions:

$$\sum_{k=0}^{m} \ln x_k \, e^{x_k}, \sum_{k=0}^{m} (\cos x_k)^2, \ldots$$

A more general form, . . .

Linear Combinations of General Functions

- $m+1$ points $\{(x_0,y_0),(x_1,y_1),\ldots,(x_m,y_m)\}$
- $n+1$ "basis" functions $g_0,g_1,\ldots,g_n$, such that

$$g(x)=\sum_{j=0}^{n} c_j\, g_j(x)$$

- Error function ϕ

$$\phi(c_0,c_1,\ldots,c_n)=\sum_{k=0}^{m}\left(\sum_{j=0}^{n} c_j\, g_j(x_k)-y_k\right)^2$$

- Minimization:

$$\frac{\partial\phi}{\partial c_i}=2\sum_{k=0}^{m}\left(\sum_{j=0}^{n} c_j\, g_j(x_k)-y_k\right)g_i(x_k)\overset{\text{set}}{=}0,\quad i=0,\ldots,n$$

Pulling it together, . . .

Normal Equations

- "Normal equations" :

$$\sum_{j=0}^{n} \left(\sum_{k=0}^{m} g_i(x_k)\, g_j(x_k) \right) c_j = \sum_{k=0}^{m} y_k\, g_i(x_k), \quad i = 0, \ldots, n$$

- Note: $n+1$ equations (i.e., rows) and $n+1$ columns

- (Coefficient matrix)$_{i\,j} = \sum_{k=0}^{m} g_i(x_k)\, g_j(x_k)$

- Possible solution method: Gaussian elimination

- Require of $g_j(x)$ for any solution method
  - ∗ linear independence (lest there be no solution)
  - ∗ appropriateness (e.g., not sin's for linear data)
  - ∗ well-conditioned matrix (opposite of ill-conditioned)

Choice of Basis Functions

- What if basis functions are unknown?
- Choose them for numerically "good" coefficient matrix (at least not ill-conditioned)
- Orthogonality $\Rightarrow$ diagonal matrix, would be nice
- Orthonormality $\Rightarrow$ identity matrix, would be best, i.e., $\sum_{k=0}^{m} g_i(x_k)\, g_j(x_k) = \delta_{i\,j}$ and compute coefficients directly

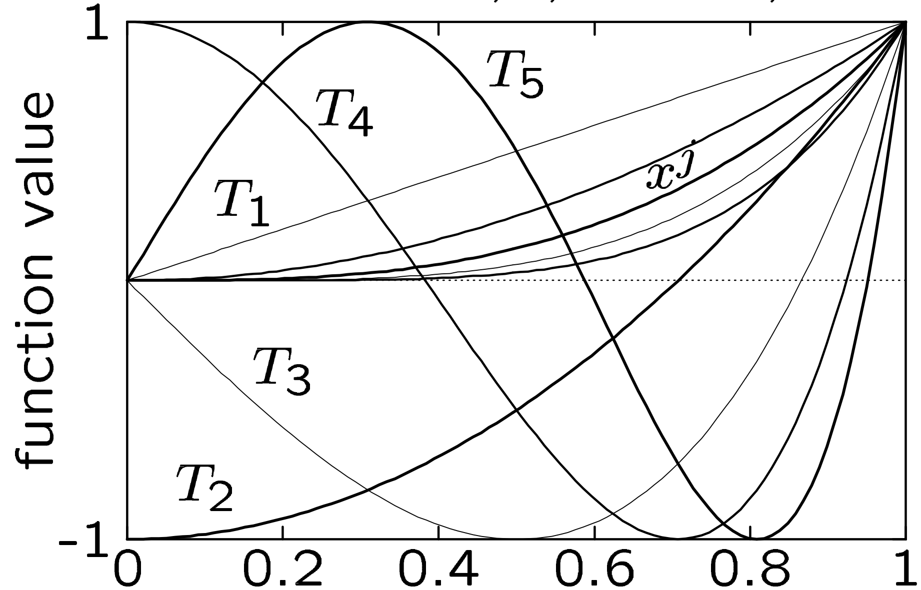
$$c_i = \sum_{k=0}^{m} y_k\, g_i(x_k), \quad i = 0, \ldots, n$$

- Can be done with Grahm-Schmidt process

Another method for choosing basis functions, . . .

Chebyshev Polynomials

- Assume that the basis functions are $\in P_n$, $x_i \in [-1, 1]$
- $1, x, x^2, x^3, \ldots$ are too alike to describe varying behavior
- Use Chebyshev polynomials: $1, x, 2x^2 - 1, 4x^3 - 3x, \ldots$



. . . with Gaussian elimination produces accurate results.

Least Squares Method

- Motivation and Approach
- Linearly Dependent Data
- General Basis Functions
- ⇒ Polynomial Regression
- Function Approximation

Motivation and Definition

- Want to smooth out data to a polynomial $p_N(x)$
- Problem: what degree N polynomial?
- For $m+1$ points, certainly $N < m$, as $N = m$ is interpolation
- Define variance σ_n^2

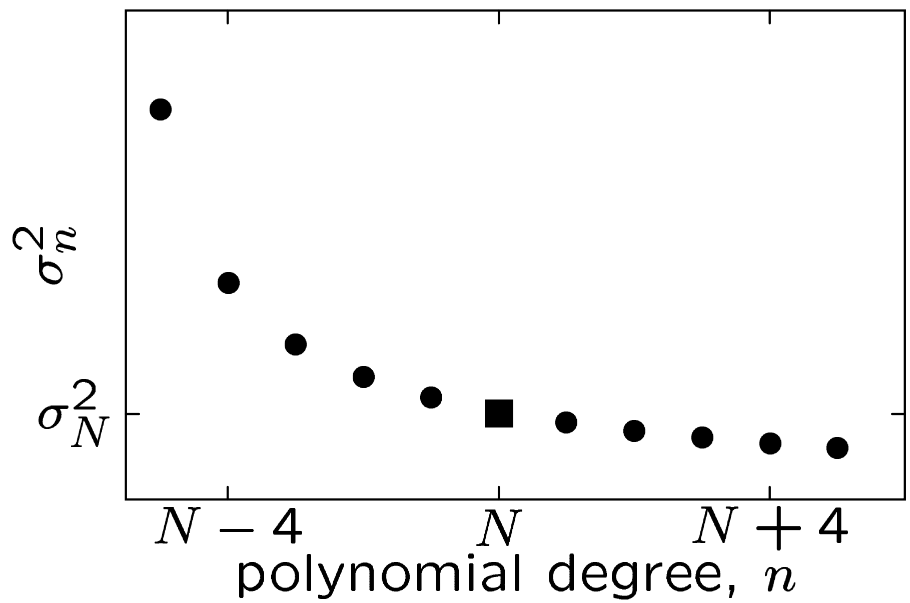
$$\sigma_n^2 = \frac{1}{m-n} \sum_{k=0}^{m} [y_k - p_n(x_k)]^2 \quad (m > n)$$

Regression Theory

- Statistical theory: if data (sans noise) is really of $p_N(x)$, then:

$$\sigma_0^2 > \sigma_1^2 > \sigma_2^2 > \cdots > \sigma_N^2 = \sigma_{N+1}^2 = \sigma_{N+2}^2 = \cdots = \sigma_{m-1}^2$$

- With noisy data stop when $\sigma_N^2 \approx \sigma_{N+1}^2 \approx \sigma_{N+2}^2 \approx \cdots$



Least Squares Method

- Motivation and Approach
- Linearly Dependent Data
- General Basis Functions
- Polynomial Regression

⇒ Function Approximation

Continuous Data

- Given $f(x)$ on $[a, b]$, perhaps from experiment
- Replace complicated or numerically expensive $f(x)$ with

$$g(x) = \sum_{j=0}^{n} c_j \, g_j(x)$$

- Continuous analog of error function

$$\phi(c_0, c_1, \ldots, c_n) = \int_a^b [g(x) - f(x)]^2 dx$$

- Can also weight parts of the interval differently

$$\phi(c_0, c_1, \ldots, c_n) = \int_a^b [g(x) - f(x)]^2 \, w(x) \, dx$$

Normal Equations and Basis Functions

- Differentiating, we get the normal equations

$$\sum_{j=0}^{n}\left[\int_a^b g_i(x)\,g_j(x)\,w(x)\,dx\right]c_j = \int_a^b f(x)\,g_i(x)\,w(x)\,dx, \quad i = 0,\ldots,n$$

- Want orthogonality of (coefficient matrix)$_{i\,j}$

$$\int_a^b g_i(x)\,g_j(x)\,w(x)\,dx = 0, \quad i \neq j$$

- For weighting interval ends, use Chebyshev polynomials since

$$\int_{-1}^{1} T_i(x)\,T_j(x)\,\frac{1}{\sqrt{1-x^2}}\,dx = \begin{cases} 0, & i \neq j, \\ \frac{\pi}{2}, & i = j > 0 \\ \pi, & i = j = 0 \end{cases}$$

Simulation

$\Rightarrow$ Random Numbers
- Monte Carlo Integration
- Problems and Games

Motivation

- Typical problem: traffic lights (sans clover leaf)
  * given traffic flow parameters ...
  * how to determine the optimal period
  * how to distribute the time per period
  * note: these are all inter-dependent

- Analytically very hard (or impossible)

- Empirical simulation can approach the problem

- Need to implement randomization for modeling various conditions

Less mathematical, but not less important.

Random Numbers—Usage

- With simulation ⇒ assist understanding of
  - standard/steady state conditions
  - various perturbations
- *Monte Carlo*: running a process many times with randomization
  - help draw statistics

Random Numbers—Requirements

- Not ordered, e.g., monotonic or other patterns
- Equal distribution
- Often RNG produce $x \in [0, 1)$
- Desired (demanded!): $P(a, a + h) = h$; *independent* of a
- Low or no periodicity
- No easy generating function from one number to the next
  - can be deceivingly random-looking
  - e.g.: digits of π

Random Number Generators

- Computers are deterministic $\Rightarrow$ not an easy problem

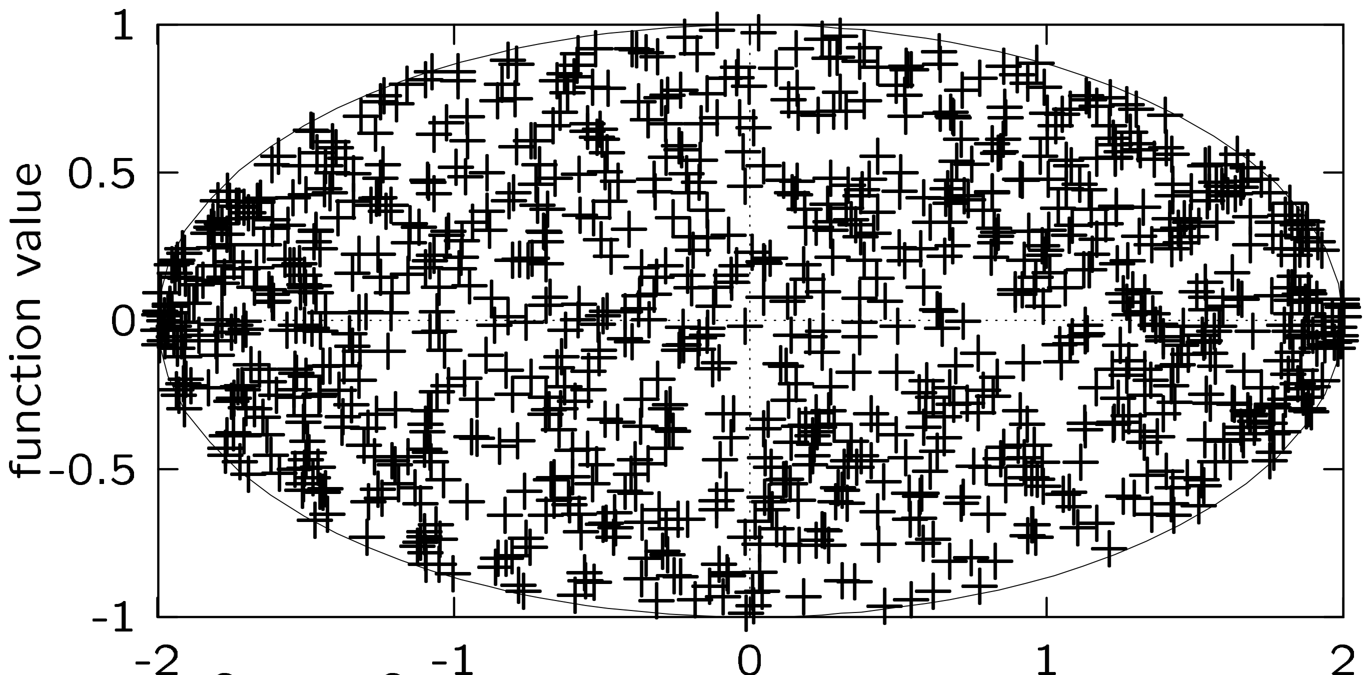
- Current computer $\frac{1}{100}$ of seconds—not good
  * for requests every $< \frac{1}{100}$ second
  * for any requests with periodicity of $\frac{1}{100}$ second

- Often based on Mersenne primes (so far, 40 of them)
  * definition: $2^k - 1$, for some k
  * e.g.: $k = 31 \Rightarrow$ 2,147,483,647
  * largest (as of 17 November 2003): $k = 20,996,011 \Rightarrow$ 6,320,430 decimal digits!
  * other usages: cryptology

Testing and Using a RNG

- "Not all RNG were created equal!"
- One can (and should) histogram a RNG
- Not obvious (nor necessarily known)
  - number of trials necessary for testing a RNG
  - number of trials necessary when using a RNG
- For ranges other than $[0, 1)$: apply obvious mapping

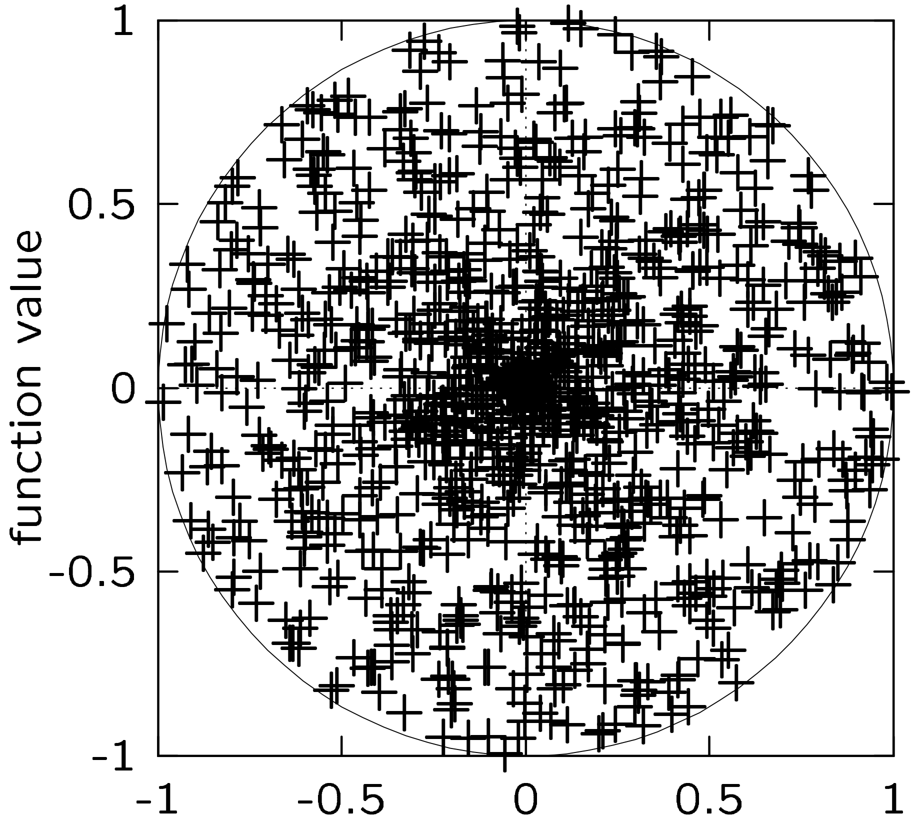
Incorrect Usage—In an Ellipse



- Equation: $x^2 + 4y^2 = 4$
- Generation algorithm:
  - ∗ $x_i \in$ rng$(-2, 2)$, $y_i \in$ rng$(-1, 1)$
  - ∗ y_i correction: $y_i \leftarrow (y_i/2)\sqrt{4 - x_i^2}$

Points bunch up at ends ⇒ non-uniformity.

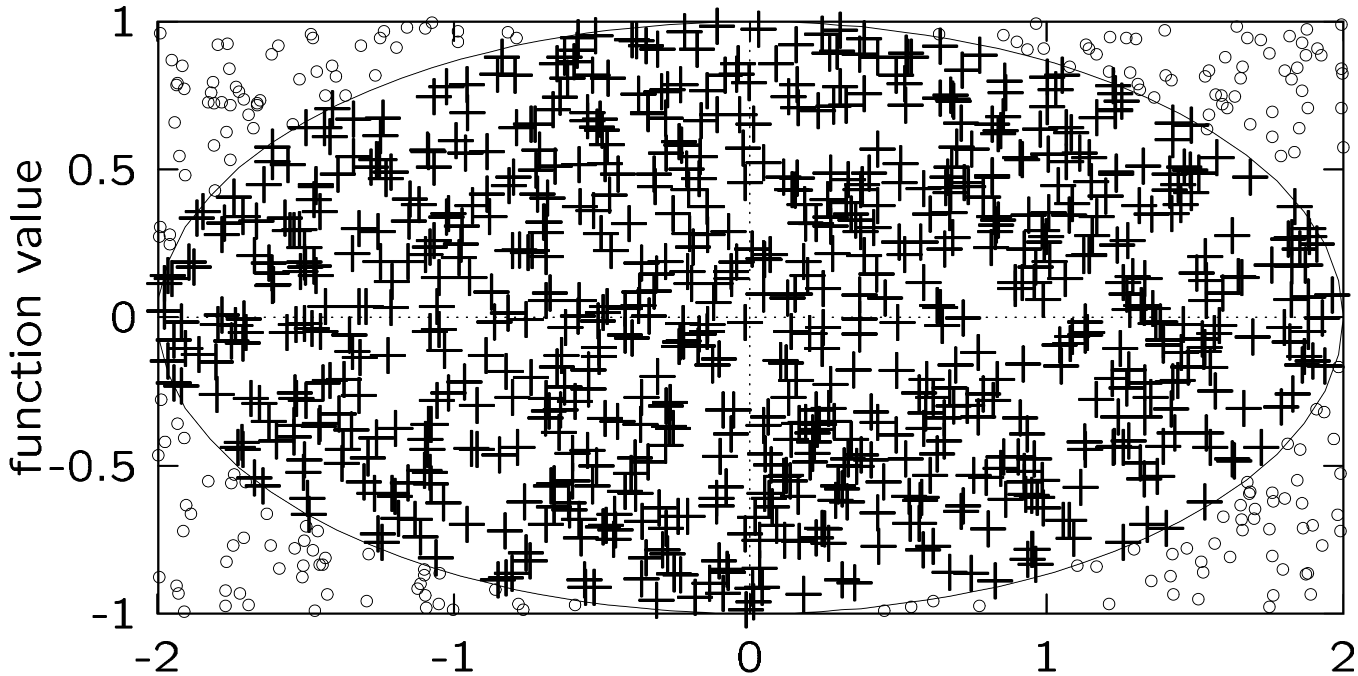
Incorrect Usage—In a Circle



- Generation algorithm: $\theta_i \in \text{rng}(0, 2\pi)$, $r_i \in \text{rng}(0, 1)$

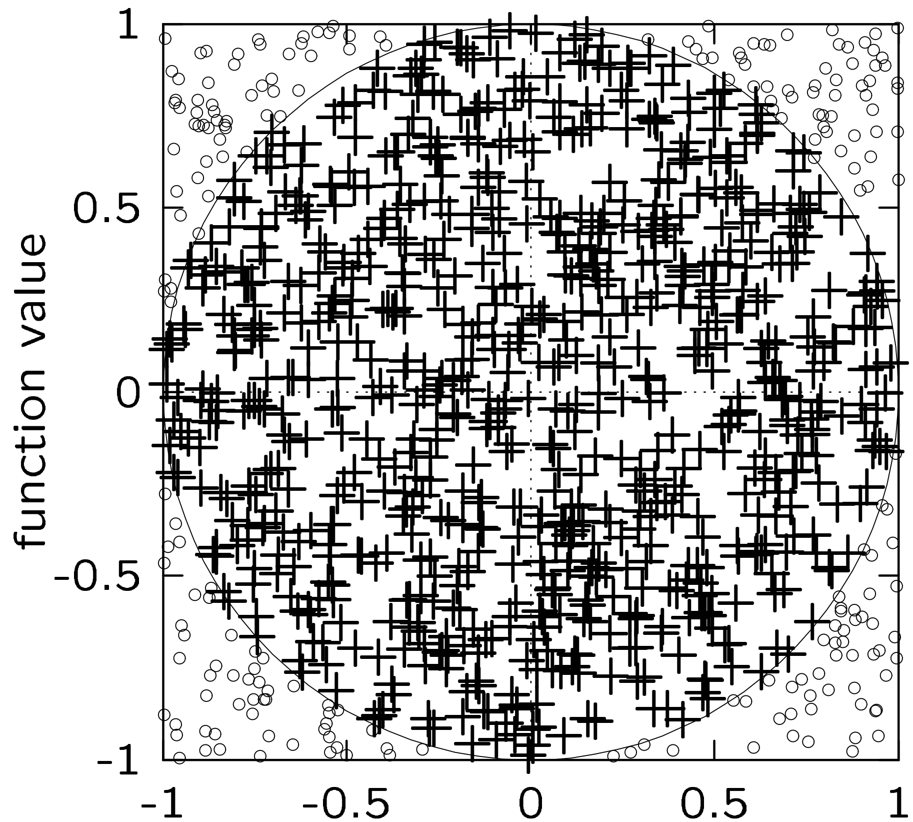
Points bunch in the middle $\Rightarrow$ non-uniformity.

correct Usage—In an Ellipse



- Generate extra points, discarding exterior ones

Correct Usage—In a Circle



- Generate extra *Cartesian* points, discarding exterior ones

Simulation

- Random Numbers

⇒ Monte Carlo Integration

- Problems and Games

Numerical Integration

- Motivation: to solve $\int_0^1 f(x)dx$
- Possible solutions
  - ∗ Composite Trapeziod Rule
  - ∗ Composite Simpson's Rule
  - ∗ Romberg Algorithm
  - ∗ Guassian Quadrature
- Problem: sometimes things are more difficult, particularly in higher dimensions
- Monte Carlo solution: for $x_i \in \text{rng}(0, 1)$

$$\int_0^1 f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

- Error (from statistical analysis): $O(1/\sqrt{n})$

Higher Dimensions and Non-Unity Domains

- In 3–D: for $(x_i, y_i, z_i) \in \text{rng}(0,1)$

$$\int_0^1 \int_0^1 \int_0^1 f(x,y,z)dx\,dy\,dz \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i, y_i, z_i)$$

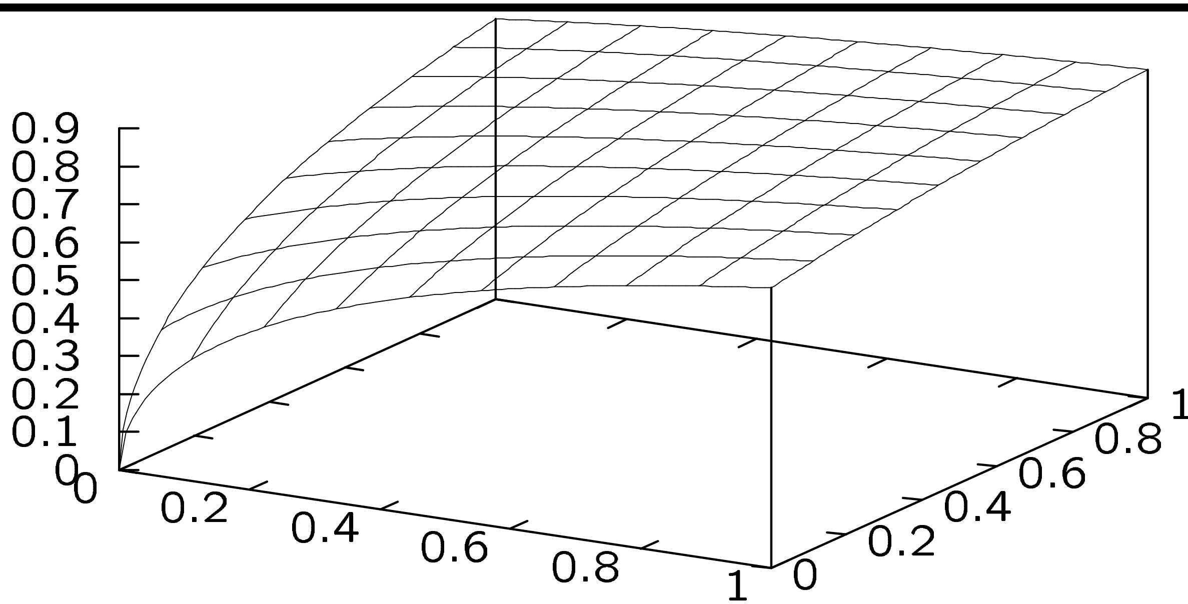
- Non-unity domain: for $x_i \in \text{rng}(a,b)$

$$\int_a^b f(x)dx \approx (b-a)\frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

- In general:

$$\int_A f \approx (\text{size of } A) \times (\text{average of } f \text{ for } n \text{ random points in } A)$$

Sample Integration Problem



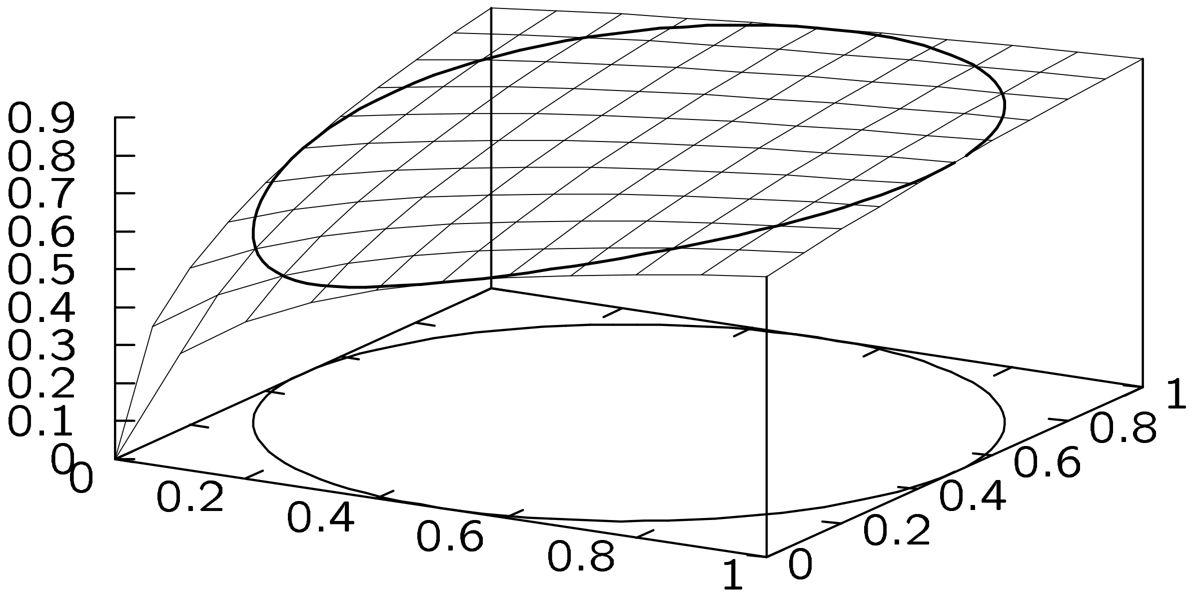
- Integral:

$$\iint_{\Omega} \sin \sqrt{\ln (x+y+1)}\, dx\, dy$$

- Domain:

$$\Omega = \left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2 \leq \frac{1}{4}$$

Sample Integration Solution



- Solution: $\dfrac{\pi}{4n}\sum_{i=1}^{n} f(p_i)$, p_i chosen properly (how?)

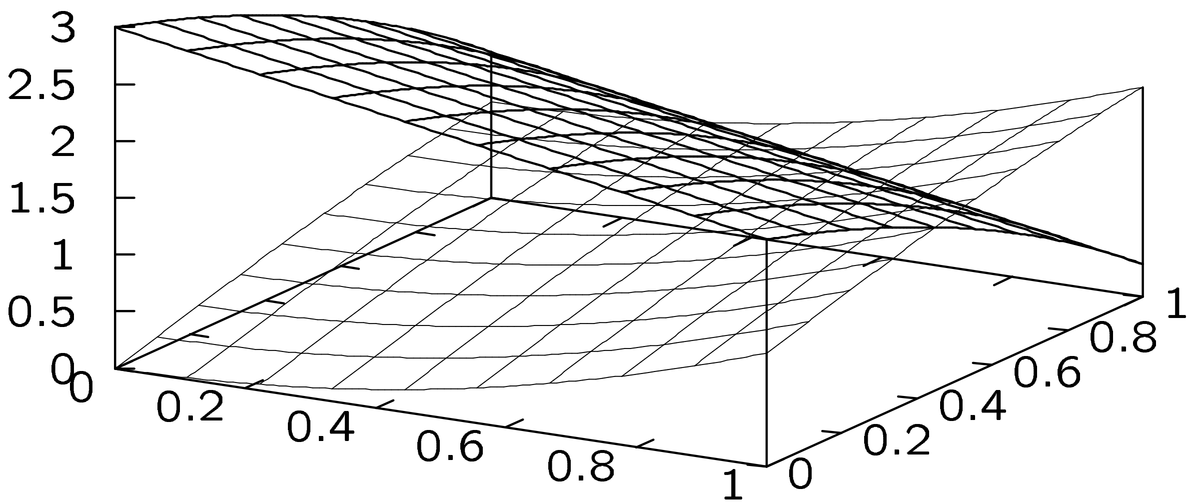
Computing Volumes

- Problem: determine the volume of the region which satisfies:

$$\begin{cases} 0 \leq x \leq 1 \quad 0 \leq y \leq 1 \quad 0 \leq z \leq 3 \\ x^2 + \sin y \leq z \\ x + e^y + z \geq 4 \end{cases}$$

- Solution
  * generate random points in $(0, 0, 0) \ldots (1, 1, 3)$
  * determine percentage which satisfies constraints

Geometric Interpretation



- Desired volume is on the left hand side, between the graphs

Simulation

- Random Numbers
- Monte Carlo Integration

⇒ Problems and Games

Probability/Chance of Dice and Cards

- Dice
  - 12, for 2 die, 24 throws
  - 19, for many die
  - loaded die
- Cards
  - shuffling in general
  - straight flush
  - royal flush
  - 4 of a kind

Can be calculated exactly, or approximated by simulation.

Miscellaneous Problems

- How many people for probable coinciding birthdays?
- Buffon's Needle
  - ∗ lined paper
  - ∗ needle of inter-line length
  - ∗ probability of dropped needle crossing a line?
- Monty Hall problem
- Neutron shielding ("random walk")
- n tennis players $\Rightarrow$ how many matches?
- 100 light switches, all off
  - ∗ person i switches multiples of i, $i = 1, \ldots, 100$
  - ∗ which remain on?

Problems with somewhat difficult analytic solutions.