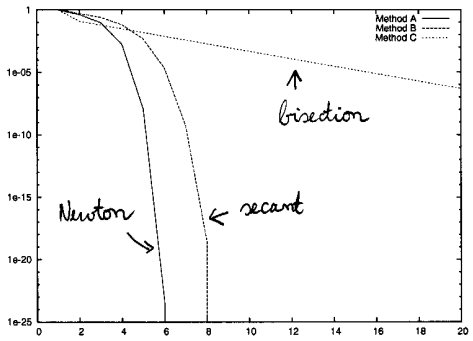
1. The following graphs compare the performance of the bisection, Newton, and secant root finding method for two different functions.

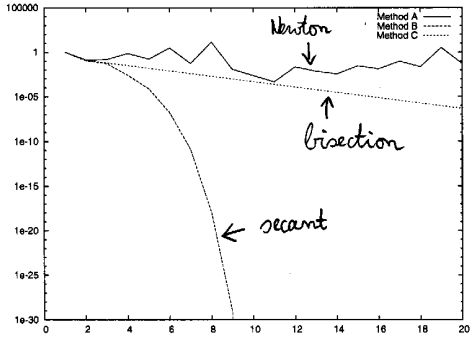
   The first test is for finding the root at $x = 0$ of the function

$$f(x) = x + x^2 \sin x .$$



   The second test is for finding the root at $x = 0$ of the function

$$g(x) = x + x^2 \sin \frac{1}{x} .$$



   (a) Identify methods A, B, and C. (The labeling in the two graphs is the same.) Comment briefly on each match.

   (b) **Extra credit.** Analyze why one of the methods fails so badly for finding the root of g while the others do just fine.

(10+*10*)

(a) f is a smooth function with a simple root at $x=0$.
Thus, all methods have their expected order:

Bisection : linear convergence

Secant : superlinear, ≈ 1.6

Newton : quadratic (faster than secant)

Function g, when continuously extended by setting $g(0)=0$, also has a root at $x=0$, but is not smooth (to be specific, one can show that g is differentiable, but not continuously differentiable, at $x=0$). So we expect

Bisection : linear convergence, method depends only on continuity!

Newton : expect trouble near the root as f' not continuous

Secant : Does not involve explicit derivatives, but analysis as done in class involves at least bounded second derivatives.
This is surprising because the graph shows that the secant method works fine with comparable speed of convergence as before.

(b) [illegible] of the answer [illegible] see why the secant method does well even though our analysis from class is not applicable.

Use FTC:

$$g(x_k) - g(x_{k-1}) = \int_{x_{k-1}}^{x_k} g'(x)\,dx = \int_{x_{k-1}}^{x_k} \left(1 + 2x \sin\frac{1}{x} - \cos\frac{1}{x}\right) dx$$

Suppose $x_{k-1} > x_k > 0$. (monotonic convergence, the analysis will show that this ordering will be preserved under iteration provided x_1 and x_0 are sufficiently small) (*

Then

$$\left| \frac{g(x_k) - g(x_{k-1})}{x_k - x_{k-1}} - 1 \right| \leq 2x_{k-1} + \frac{1}{x_{k-1} - x_k} \left| \int_{x_k}^{x_{k-1}} \cos\frac{1}{x}\,dx \right|$$

We need to analyze

$$\left| \int_{x_k}^{x_{k-1}} \cos\frac{1}{x}\,dx \right| = \left| \int_{\frac{1}{x_{k-1}}}^{\frac{1}{x_k}} \frac{\cos y}{y^2}\,dy \right|$$

$$\leq \underbrace{\left| \int_{\frac{1}{x_{k-1}}}^{\frac{1}{x_{k-1}}+\pi} \frac{\cos y}{y^2}\,dy \right|}_{\leq \frac{\pi}{\left(\frac{1}{x_{k-1}}\right)^2} = \pi x_{k-1}^2} + \sum_{j=0}^{\infty} \left| \int_{\frac{1}{x_{k-1}}+\pi+2\pi j}^{\frac{1}{x_{k-1}}+\pi+2\pi(j+1)} \frac{\cos y}{y^2}\,dy \right|$$

Set $x = \frac{1}{x_{k-1}} + \pi + 2\pi j$

$z = y - x$

$$\Rightarrow \int_{\frac{1}{x_{k-1}}+\pi+2\pi j}^{\frac{1}{x_{k-1}}+\pi+2\pi(j+1)} \frac{\cos y}{y^2}\,dy = \int_0^{2\pi} \frac{\cos(x+z)}{(x+z)^2}\,dz$$

$$= \int_0^{2\pi} \cos(x+z) \left[\frac{1}{x^2} - \underbrace{\frac{(2x+z)z}{x^2(x+z)^2}}_{\leq \frac{4xz}{x^4} \leq \frac{8\pi}{x^3}} \right] dz$$

The integral over the first term vanishes, so that we can estimate the entire sum as

$$\sum_{j=0}^{\infty} \left| \int_{\frac{1}{x_{k-1}}+\pi+2\pi j}^{\frac{1}{x_{k-1}}+\pi+2\pi(j+1)} \frac{\cos y}{y^2}\,dy \right| \leq \sum_{j=0}^{\infty} \frac{8\pi}{\left(\frac{1}{x_{k-1}} + \pi + 2\pi j\right)^3}$$

$$\leq 8\pi \,\frac{1}{\pi} \sum_{j=0}^{\infty} \frac{1}{\left(\frac{1}{\pi x_{k-1}} + 1 + j\right)^3}$$

$$\leq 8 \int_0^{\infty} \frac{1}{\left(\frac{1}{\pi x_{k-1}} + t\right)^3}\,dt = 4 \frac{1}{\left(\frac{1}{\pi x_{k-1}}\right)^2} = 4\pi^2 x_{k-1}^2$$

We conclude that

$$\frac{g(x_k) - g(x_{k-1})}{x_k - x_{k-1}} = 1 + O(x_{k-1})$$

provided that $|x_k| \leq \alpha |x_{k-1}|$ with $\alpha \ll 1$. (**)

I.e., we get the same error expression as for the secant method applied to a smooth next order perturbation of $g(x) = x$. Thus, the standard error analysis for the secant method with smooth function applies, (*) and (**) are consistently satisfied under iteration, and the order of convergence is as usual.

Remark: This detailed hard analysis was NOT required for a full score. A good answer would have included, in one way or another, the (geometric) observation that the secant method picks up an average slope of g near the root.

2. The secant method, which can be written as

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

involves subtraction of almost equal numbers as the sequence approaches a limit. Yet, the method proves very robust in practice. We shall thus analyze the propagation of rounding errors for the secant method.

(a) Assume that all operations are exact *except* for the evaluation of f. Why can we expect that this simplified analysis still captures the essential error behavior of the secant method?

(b) Show that one step of the secant method in floating point is approximately

$$x_{k+1} \approx x_k - (x_k - \xi) \frac{1}{1 \mp \delta \frac{x_{k-1} - \xi}{x_k - x_{k-1}}},$$

where ξ denotes a simple root of f, for some $\delta \ll 1$.
Hint: It is easiest to write

$$\mathrm{fl}\left(\frac{f(x_{k-1})}{f(x_k)}\right) = \frac{f(x_{k-1})}{f(x_k)}(1+\delta), \qquad (*)$$

then use truncated Taylor series to estimate the quotient.

(c) How does (*) imply that the secant method is robust to rounding errors?

(5+10+5)

(a)
- The main problem with computing differences of almost equal numbers is the amplification of relative errors in the input data. The introduction of new errors is either negligible, or even zero.
- In each step, we can treat x_k and x_{k-1} as exact, and study the decrease in error with respect to the true root ξ. This is the only error that ultimately matters.

(b) Since $f(x_k) = f(\xi + (x_k - \xi)) = \underbrace{f(\xi)}_{=0} + f'(\xi)(x_k - \xi) + \text{h.o.t.}$,

$$fl\left(\frac{f(x_{k-1})}{f(x_k)}\right) = \frac{f'(\xi)\,(x_{k-1}-\xi)}{f'(\xi)\,(x_k-\xi)}\,(1+\delta) = \frac{x_{k-1}-\xi}{x_k-\xi}\,(1+\delta)$$

where δ incorporates errors introduced through rounding but also through truncating the Taylor series above:

$$|\delta| \approx \max\{\varepsilon_{machine},\ |x_{k-1}-\xi|,\ |x_k-\xi|\}$$

$$\Rightarrow x_{k+1} = x_k - \frac{x_k - x_{k-1}}{1 - \frac{x_{k-1}-\xi}{x_k-\xi}(1+\delta)} = x_k - (x_k-\xi)\,\frac{x_k-x_{k-1}}{x_k-\xi-(x_{k-1}-\xi)(1+\delta)}$$

$$= x_k - (x_k-\xi)\,\frac{x_k-x_{k-1}}{(x_k-x_{k-1})-\delta(x_{k-1}-\xi)} = x_k - (x_k-\xi)\,\frac{1}{1-\delta\frac{x_{k-1}-\xi}{x_k-x_{k-1}}}$$

(c) Provided $\delta\,\frac{x_{k-1}-\xi}{x_k-x_{k-1}} \ll 1$,

$$x_{k+1} \approx x_k - (x_k-\xi)\left(1+\delta\,\frac{x_{k-1}-\xi}{x_k-x_{k-1}}\right) = \xi - \delta\,\frac{(x_k-\xi)(x_{k-1}-\xi)}{x_k-x_{k-1}}$$

Thus, so long as $|\delta| \approx |x_{k-1}-\xi|$, we are back in the standard convergence analysis. Thus, rounding errors will only become significant if $\varepsilon_{machine} \approx |x_{k-1}-\xi|$.

3. Compute the LU factorization without pivoting of the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 14 \end{pmatrix}.$$

(15)

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 14 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\underbrace{\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 3 & 8 & 14 \end{pmatrix}}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}\underbrace{\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 2 & 5 \end{pmatrix}}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}\underbrace{\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}}_{=:U}$$

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} \quad (= U^T)$$

4. (a) Prove that if $A \in M(n \times n)$ is positive definite, then all diagonal entries must be positive.

   (b) Let $A \in M(n \times n)$ be symmetric and positive definite, with $A = L + D + R$ denoting the splitting of A into its lower left triangular, diagonal, and upper right triangular parts (as in the derivation of the Jacobi and Gauss–Seidel algorithms).

   Show that the so-called *symmetric Gauss–Seidel preconditioner*

$$B_{SGS} = (D + R)\, D^{-1}\, (D + L)$$

   is indeed symmetric and positive definite.

(10+10)

(a) A pos. def. $\Rightarrow v^T A v > 0 \quad \forall v \neq 0$

Take $v = e_i$: $\quad 0 < e_i^T A e_i = a_{ii}$ $\qquad A = (a_{ij})$

(b) A symmetric : $L = R^T$

$$\Rightarrow B_{SGS}^T = \left((D+R)\, D^{-1}\, (D+R^T) \right)^T$$

$$= \underbrace{(D+R^T)^T}_{= D+R} \underbrace{(D^{-1})^T}_{= D^{-1}} \underbrace{(D+R)^T}_{= D+R^T}$$

$$= B_{SGS}$$

To show that B_{SGS} is pos. def., take $v = (D+R^T)^{-1} w$

(note that A pos. def. $\Rightarrow A$ nonsingular $\Rightarrow D+L$ nonsingular !)

$$\Rightarrow v^T B_{SGS} v = w^T (D+R)^{-1} (D+R)\, D^{-1} (D+R^T) (D+R^T)^{-1} w$$

$$= w^T D^{-1} w$$

From (a): D pos. def. $\Rightarrow D^{-1}$ pos. def. $\Rightarrow w^T D^{-1} w > 0 \quad \forall w \in \mathbb{R}^n$ □

5. Let

$$A = \begin{pmatrix} \varepsilon & 1 \\ 1 & \varepsilon \end{pmatrix}$$

with $|\varepsilon| < 1$.

   (a) Show that the Jacobi method for solving $Ax = b$ does not converge for any choice of right hand side b and starting vector x_0.

   (b) Suggest a "Jacobi method with pivoting" so that the modified Jacobi iteration converges for the given matrix A.

(10+10)

(a) Jacobi iteration: $A = D + (A - D)$

$$x_{k+1} = D^{-1} b - \underbrace{D^{-1}(A-D)}_{=: B} x_k$$

where $B = \begin{pmatrix} \frac{1}{\varepsilon} & 0 \\ 0 & \frac{1}{\varepsilon} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\varepsilon} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

Eigenvalues are $\pm \frac{1}{\varepsilon}$, hence $\rho(B) = \frac{1}{|\varepsilon|} > 1$.

(b) Let's permute the matrix (e.g. recursively by row index) such that the largest element below or to the right of a diagonal element is moved onto the diagonal. Here, we'll solve

$$PAx = Pb \qquad \text{with } P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$PA = D + (PA - D) \qquad \text{with } D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Rightarrow x_{k+1} = D^{-1} Pb - D^{-1} \underbrace{(PA - D)}_{=: B} x_k$$

$$\Rightarrow B = \varepsilon \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \Rightarrow \rho(B) = |\varepsilon| < 1, \text{ hence convergence.}$$

6. (a) State the definition of the norm of a matrix induced by a vector norm; state the definition of the condition number of a matrix.

   (b) Let $A \in M(n \times n)$ be invertible and $\mathbf{b} \in \mathbb{R}^n$. Let $\mathbf{x}^*$ denote the exact solution to the linear system $A\mathbf{x}^* = \mathbf{b}$ and let $\mathbf{x}$ denote an approximation to $\mathbf{x}^*$. Show that

$$\frac{\|\mathbf{x}^* - \mathbf{x}\|}{\|\mathbf{x}^*\|} \leq \kappa(A) \frac{\|\mathbf{b} - A\mathbf{x}\|}{\|\mathbf{b}\|}$$

   where the vector norm $\|\cdot\|$ is arbitrary, and the condition number is defined with respect to the induced matrix norm.

(5+10)

(a) Let $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a vector norm. Then for $A : \mathbb{R}^n \to \mathbb{R}^n$

$$\|A\| = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|}$$

$$\text{cond}(A) = \|A\| \, \|A^{-1}\|$$

(b) $Ax^* = b \Rightarrow A(x^* - x) = b - Ax$

$$\Rightarrow x^* - x = A^{-1}(b - Ax)$$

$$\Rightarrow \|x^* - x\| \leq \|A^{-1}\| \, \|b - Ax\|$$

$$\|A\| \, \|x^*\| \geq \|b\|$$

$$\frac{\|x^* - x\|}{\|A\| \, \|x^*\|} \leq \|A^{-1}\| \frac{\|b - Ax\|}{\|b\|}$$

the claim follows through multiplication by $\|A\|$.