MTH 373/573:
Scientific Computing

Practice Final

IIIT-Delhi, Monsoon 2022
Kaushik Kalyanaraman

Multiple Choice Questions

Problem 1: Floating point numbers

Consider the following Python output that has been obtained after appropriately setting up variables and instantiating them:

```
>>> type(x)
float
>>> x > 0
True
>>> 1.0 + x
1.0
```

What can we conclude about x?

(a) x is equal to zero.

(b) x is a subnormal number.

(c) x is less than underflow (UFL).

(d) x is less than machine epsilon (ε_{M})

D

By definition, the machine epsilon ε_{M} is that floating point number which, under a prescribed rounding rule choice, when added to the floating point number 1.0 gives us the next floating point number. Thus, if $1.0 + x$ is unchanged as in the code snippet, $x < \varepsilon_{\mathrm{M}}$.

Problem 2: Floating point arithmetic

Given n positive floating point numbers x_i, in which order should the summation $\sum\limits_{i=1}^{n} x_i$ be computed in order to minimize the rounding error?

(a) Smallest to largest.

(b) Largest to smallest.

(c) Rounding error is independent of order.

(d) None of these.

A

Recall that for an arithmetic operation (such as summation here) for two numbers x and y in a floating point number system, we have that the rounding error is given by:

$$\mathrm{fl}(x + y) = (x + y)(1 + \delta),$$

where $|\delta| \leq \varepsilon_{\mathrm{M}}$. Thus, writing this out for a few terms in the summation (where all terms are positive) $S = x_1 + x_2 + \cdots + x_n$, we have that:

$$\begin{aligned}
S_1 &= \mathrm{fl}(x_1),\\
S_2 &= \mathrm{fl}(\mathrm{fl}(x_1) + \mathrm{fl}(x_2)) = (\mathrm{fl}(x_1) + \mathrm{fl}(x_2))(1 + \delta_1),\\
S_3 &= \mathrm{fl}(\mathrm{fl}(x_1) + \mathrm{fl}(x_2) + \mathrm{fl}(x_3)) = \mathrm{fl}(S_2 + \mathrm{fl}(x_3)) = (S_2 + \mathrm{fl}(x_3))(1 + \delta_2)\\
&= ((\mathrm{fl}(x_1) + \mathrm{fl}(x_2))(1 + \delta_1) + \mathrm{fl}(x_3))(1 + \delta_2)\\
&= \mathrm{fl}(x_1) + \mathrm{fl}(x_2) + \mathrm{fl}(x_3) + \mathrm{fl}(x_1)\delta_1\delta_2 + \mathrm{fl}(x_2)\delta_1\delta_2 + \mathrm{fl}(x_3)\delta_2\\
S_4 &= \mathrm{fl}(x_1) + \mathrm{fl}(x_2) + \mathrm{fl}(x_3) + \mathrm{fl}(x_4) + \mathrm{fl}(x_1)\delta_1\delta_2\delta_3 + \mathrm{fl}(x_2)\delta_1\delta_2\delta_3 + \mathrm{fl}(x_3)\delta_2\delta_3 + \mathrm{fl}(x_4)\delta_3,
\end{aligned}$$

and so on. Thus, we see that the multipliers $\delta_1, \delta_2, \ldots$ in the sum will maximum scaling effect, and consequently largest round off error, if (without loss of generality) $x_1 \geq x_2 \geq \cdots \geq 0$. Thus, to minimize round off error, we wish to sum from the smallest number to the largest.

Problem 3: Stable algorithm

Why is it desirable to use a *stable algorithm* in numerical computation?

(a) It will improve the conditioning of the underlying problem.

(b) It will always produce accurate results.

(c) It is relatively insensitive to changes in the input data.

(d) It is relatively insensitive to errors made during the computation.

D

Conditioning of a problem is a smooth problem's property whereas stability is a discrete or algorithm's property. Thus, a stable algorithm cannot improve conditioning though an unstable algorithm on a well conditioned problem will give incorrect results.
A stable algorithm cannot also therefore guarantee accurate results unless the original problem is well conditioned.
Changes in input and its effect on the output or solution is again, by definition, related to the condition number and conditioning.
Thus, a stable algorithm can only guarantee that computational errors do not propagate, or are amplified by, a lot.

Problem 4: Conditioning of Function Evaluation

What is the (approximate) relative condition number of evaluating $f(x) = \log(1+x)$ at $x = 3$?

(a) $\frac{3}{4\log 4}$ (b) $\frac{1}{4\log 4}$ (c) $\log(4)$ (d) $\frac{1}{12\log 4}$

A

By definition, we have that the relative condition number for function evaluation is:

$$\kappa(x) = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x}{(1+x)\log(1+x)}\right|.$$

Thus, at $x = 3$, the relative condition number is $\frac{3}{4\log 4}$.

Problem 5: Linear algebra

Which of the following statements about an $n \times n$, real-valued matrix A is not logically equivalent to the others?

(a) There is an $n \times n$ matrix B such that $AB = BA = I$.

(b) $\det(A) \neq 0$

(c) The column rank of A and the row rank of A are identical.

(d) There is no vector $x \neq 0$ such that $Ax = 0$.

C

Statements (a), (b) and (d) are equivalent definitions of a nonsingular or full rank matrix A. However, row rank of A being equal to the column rank does not ensure that the matrix is full rank; for a rank deficient A too, this is true.

Problem 6: Partial pivoting

What is the initial pivot element of the matrix

$$\begin{bmatrix} 4 & -8 & 1 \\ 6 & 5 & 7 \\ 0 & -10 & -3 \end{bmatrix}$$

using partial pivoting?

(a) 0 (b) 4 (c) 6 (d) -10

C

For partial pivoting, the initial pivot is the largest magnitude entry in the first column, hence 6.

Problem 7: Linear system residual

When does a small residual, $r = b - Ax$, indicate an accurate solution to the linear system $Ax = b$ for $A \in \mathbb{R}^{n \times n}$?

(a) Always

(b) When A is nonsingular

(c) When A is well conditioned

(d) When b is in the span of the columns of A

C

The relationship between the relative residual and the error in the solution of a linear system is provided by:

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|A\| \, \|x\|}.$$

Thus, a small relative residual implies an accurate solution only if the condition number is small, that is, if the linear system matrix A (equivalently, the problem of finding a solution for the linear system) is well conditioned.

Problem 8: Symmetric positive definite systems

Which of the following is an advantage of solving symmetric positive definite linear systems (in contrast to general systems)?

(a) Requires about half as much work

(b) Requires about half as much storage

(c) Requires no pivoting for numerical stability

(d) All of the above

D

All options (a), (b) and (c) are true for solving a linear system with a symmetric positive definite matrix as compared with a general linear system.

Problem 9: Linear least squares residual

Consider the following linear least squares problem:

$$\begin{bmatrix} 5 & 3 \\ 0 & c \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \simeq \begin{bmatrix} 1 \\ 2 \\ a \end{bmatrix}$$

Which of the following statements is *False*?

(a) For all $c \neq 0$ there is a unique least-squares solution.

(b) For all $c \neq 0$ the residual has a 2-norm of $\|r\|_2 = |a|$.

(c) For all $c \neq 0$ the least-squares solution only exists if $a = 0$.

(d) For $c = 0$ there is still a least-squares solution, but it's not unique.

C

For any $c \neq 0$, the system matrix is full rank. Thus, for any a, the linear least squares problem has a unique solution. Furthermore, in this case, the residual is the component of the right hand side vector not in the subspace spanned by $\begin{bmatrix} 5 & 0 & 0 \end{bmatrix}^T$ and $\begin{bmatrix} 3 & c & 0 \end{bmatrix}^T$. Clearly, given the zero row, thus for b, the component $\begin{bmatrix} 0 & 0 & a \end{bmatrix}^T$ will not be in this subspace and as a result orthogonal to it. Thus, the residual is $\|r\| = |a|$.
For $c = 0$, the matrix is rank deficient and hence there are infinitely many solutions.
Thus, the false statement is option (c).

Problem 10: Data fitting

If a first-degree polynomial $x_1 + x_2 t$ is fit to the three data points:

t	1	2	3
y	1	1	2

using linear least squares, what are the resulting values of the parameters x_1 and x_2?

(a) $x_1 = 1, x_2 = 0$

(b) $x_1 = 1/3, x_2 = 1/2$

(c) $x_1 = 1/2, x_2 = 1/2$

(d) $x_1 = -1, x_2 = 1$

B

Solve the following least squares system to obtain the coefficients:

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \simeq \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}.$$

Problem 11: Householder's method

Which of the following is a disadvantage of solving an overdetermined system by Householder as opposed to normal equations?

(a) Householder is generally less stable.

(b) Householder is generally less accurate.

(c) Householder is generally more expensive.

(d) Householder can only be applied to symmetric systems.

C

Quoting from Section 3.7 in [Heath, 2018]: "The normal equations method is easy to implement: it simply requires matrix multiplication and Cholesky factorization. Moreover, reducing the problem to an $n \times n$ system is very attractive when $m \gg n$. By taking advantage of its symmetry, forming the cross-product matrix $A^T A$ requires about $mn^2/2$ multiplications and a similar number of additions. Solving the resulting linear system by Cholesky factorization requires about $n^3/6$ multiplications and a similar number of additions. Unfortunately, the normal equations method produces a solution whose relative error is proportional to $\kappa(A)^2$, and the required Cholesky factorization can be expected to break down if $\kappa(A) \approx 1/\sqrt{\varepsilon_M}$ or worse."

"For solving dense linear least squares problems, the Householder method is generally the most efficient and accurate of the orthogonalization methods. It requires about $mn^2 - n^3/3$ multiplications and a similar number of additions. It can be shown that the Householder method produces a solution whose relative error is proportional to $\kappa(A) + \|r\|^2 \kappa(A)^2$, which is the best that can be expected since this is the inherent sensitivity of the solution to the least squares problem itself. Moreover, the Householder method can be expected to break down (in the back-substitution phase) only if $\kappa(A) \approx 1/\varepsilon_M$ or worse."

"For nearly square problems, $m \approx n$, the normal equations and Householder methods require about the same amount of work. But for highly overdetermined problems, $m >> n$, the Householder method requires about twice as much work as the normal equations method. On the other hand, the Householder method is more accurate and more broadly applicable than the normal equations method."

Problem 12: Householder matrix

Which of the following statements about Householder matrices is *False*?

(a) They are represented by orthonormal matrices.

(b) The first Householder matrix produce zeros in all but the first entry of the target vector.

(c) Using a finite number of Householder matrices, one can obtain an upper triangular matrix that is similar to the original matrix.

(d) Using QR factorization, one can use Householder matrices to solve a linear system in $O(n^3)$ time.

C

A Householder matrix is indeed orthogonal, and the first Householder matrix $H_1 \in \mathbb{R}^{n\times n}$ acting on a vector $a \in \mathbb{R}^n$ indeed zeros out all but the first entry of this vector but while preserving the norm. In applying a Householder transformation H to an arbitrary vector u, we note that:

$$Hu = \left(I - 2\frac{vv^T}{v^Tv}\right)u = u - \left(2\frac{v^u}{v^Tv}v\right).$$

Computing v^Tu and v^Tv each involves n multiplications and $n-1$ additions thus for a total cost of $\mathcal{O}(n)$. The addition of a vector u with a scaled version (by the factor $2\frac{v^u}{v^Tv}$) of another vector u also costs $\mathcal{O}(n)$ since each component of the resulting vector requires 1 addition and 1 multiplication. Thus, Hu costs $\mathcal{O}(n)$. In the first step, the computation requires $n\mathcal{O}(n) = \mathcal{O}(n^2)$ computations for each column of a matrix A whose QR decomposition we are computing. In the second step, since the first column of A would have been orthgonalized, we will have to perform $(n-1)\mathcal{O}(n) = \mathcal{O}(n^2)$ and so on and for the last step only $\mathcal{O}(n)$ computations. Summing it all together, the QR decomposition via Householder matrices is of total cost $\mathcal{O}(n^3)$. Solving the upper triangular linear system incurs only $\mathcal{O}(n^2)$ via back substitution. Thus, the cost of solving the linear system is indeed $\mathcal{O}(n^3)$.

In fact, the statement that we can use only a finite number of Householder matrices to obtain an upper triangular matrix that is similar to the original matrix is *False* since for the Schur factorization of A, we have that $A = QTQ^T$.

Problem 13: Sequence convergence rate

The convergence of the sequence:

$$\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \cdots \frac{1}{2^k}, \cdots\right\}$$

is of what of order?

(a) Linear

(b) Superlinear

(c) Quadratic

(d) Higher than quadratic

A

By using the notion of rate of convergence, we have that:

$$\lim_{k\to\infty} \frac{|x_{k+1}|}{|x_k|} = \lim_{k\to\infty} \frac{1/2^{k+1}}{1/2^k} = \lim_{k\to\infty} \frac{1}{2} = \frac{1}{2} < 1,$$

thus, the rate of convergence is linear.

Problem 14: Polynomial interpolation

Suppose that you have already computed a polynomial interpolant for n data points, and someone gives you a new data point. Which of the following choice of basis functions would allow you to determine the new higher-degree interpolating polynomial *without* restarting from scratch (that is, without changing either the previously computed coefficients or basis functions)?

(a) Monomial basis.

(b) Newton basis.

(c) Lagrange basis

(d) None of the above.

B

This is one of the advantages of using the Newton basis polynomials. For monomials and Lagrange basis, a computed interpolant polynomial cannot be extended to a new higher degree one.

Problem 15: Interpolating polynomial

Consider interpolating the following three points using a quadratic polynomial in the *Newton basis*:

$$(-1, 2), \quad (0, 3), \quad (1, -5).$$

What is the coefficient for the first basis function if the points are interpolated in the order specified above?

(a) 1 (b) 2 (c) 3 (d) 4

B

The polynomial interpolant using Newton basis in this case would be:

$$p(t) = x_1 + x_2(t - t_1) + x_3(t - t_1)(t - t_2),$$

where $t_1 = -1$, $t_2 = 0$ and $t_3 = -5$. Let the data values be denoted as $y_1 = 2$, $y_2 = 3$ and $y_3 = -5$. For computing the coefficients x_1, x_2 and x_3, we will thus solve the following linear system:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & t_2 - t_1 & 0 \\ 1 & t_3 - t_1 & (t_3 - t_1)(t_3 - t_2) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \implies \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -5 \end{bmatrix},$$

which yields $x_1 = 2$, $x_2 = 1$, and $x_3 = -9/2$. Thus, the first coefficient is 2.

Problem 16: Cubic spline

Suppose we are given n pieces of data (x_i, y_i). Is a cubic spline interpolating these point unique?

(a) No, there are 3 free parameters.

(b) No, there are 2 free parameters.

(c) No, there is 1 free parameter.

(d) Yes, it is unique.

B

Problem 17: Bases and nodes for interpolation

Which of the following combinations of basis sets and interpolation nodes lead to an increasingly ill-conditioned generalized Vandermonde matrix (also called *basis* matrix) as the number of interpolation nodes increases?

(a) Lagrange basis with equispaced points

(b) Monomial basis with equispaced points

(c) Newton basis with equispaced points

(d) Chebyshev polynomial basis with Chebyshev points

B

A monomial basis with equispaced points yields a progressively ill conditioned matrix with more number of interpolation nodes.

Problem 18: Finite differences

Which of the following finite difference formulas is a second-order accurate approximation of the first derivative at x?

Hint: You can use Taylor expansions to figure out which of the answers below is the right one.

(a) $\dfrac{f(x+h)-f(x)}{h}$

(b) $\dfrac{f(x)-f(x-h)}{h}$

(c) $\dfrac{-3f(x)+4f(x+h)-f(x+2h)}{2h}$

(d) $\dfrac{-2f(x-2h)+2f(x)-2f(x+2h)}{4h}$

C

The following Taylor series for f around x are useful:

$$\begin{aligned}
f(x+2h) &= f(x) + 2f'(x)h + 2f''(x)h^2 + \frac{4f'''(x)}{3}h^3 + \mathcal{O}(h^4),\\
f(x+h) &= f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f'''(x)}{6}h^3 + \mathcal{O}(h^4),\\
f(x-h) &= f(x) - f'(x)h + \frac{f''(x)}{2}h^2 - \frac{f'''(x)}{6}h^3 + \mathcal{O}(h^4),\\
f(x-2h) &= f(x) - 2f'(x)h + 2f''(x)h^2 - \frac{4f'''(x)}{3}h^3 + \mathcal{O}(h^4).
\end{aligned}$$

From these, and through some gentle algebraic manipulations, we can arrive at the errors in each of the derivative computation as follows:

$$\begin{aligned}
\frac{f(x+h)-f(x)}{h} &= f'(x) + \mathcal{O}(h),\\
\frac{f(x)-f(x-h)}{h} &= f'(x) + \mathcal{O}(h),\\
\frac{-3f(x)+4f(x+h)-f(x+2h)}{2h} &= f'(x) + \mathcal{O}(h^2),\\
\frac{-2f(x-2h)+2f(x)-2f(x+2h)}{4h} &= f''(x) + \mathcal{O}(h).
\end{aligned}$$

Thus, option (c) is not even an approximation of the first derivative but for the second derivative. Clearly, option (c) is the required second order accurate approximation of the derivative.

Free Form Questions

Problem 19: Elementary elimination matrices

Given a vector $a \in \mathbb{R}^n$, an *elementary elimination matrix* $M_k \in \mathbb{R}^{n \times n}$ can annihilate all entries of the vector below the kth position provided $a_k \neq 0$ by the following transformation:

$$M_k\, a = \begin{bmatrix} 1 & \cdots & 0 & 0 & 0 & \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -m_{k+1} & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -m_n & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_k \\ a_{k+1} \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_k \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where $m_i = a_i/a_k, i = k+1, \dots, n$. Equivalently, $M_k = \mathbb{I} - me_k^T$ where $\mathbb{I}$ is the $n \times n$ identity matrix, $m = \begin{bmatrix} 0 & \dots & 0 & m_{k+1} & \dots & m_n \end{bmatrix}^T$ and e_k is the kth column of $\mathbb{I}$.

(a) Show that $M_k^{-1} = \mathbb{I} + me_k^T$ (10 points).

> M_k can be written $M_k = \mathbb{I} - me_k^T$, where e_k is the column vector with 1 in position k and 0 elsewhere. The sparsity pattern of m implies that $e_k^T m = 0$, and therefore $\left(\mathbb{I} - me_k^T\right)\left(\mathbb{I} + me_k^T\right) = \mathbb{I} - m\left(e_k^T m\right)e_k^T = \mathbb{I}$. Thus, $M_k^{-1} = \mathbb{I} + me_k^T$.

(b) If $j < k$, $M_j = \mathbb{I} - me_j^T$, and $M_k = \mathbb{I} - me_k^T$, what is the product M_jM_k? Show your computations. (10 points)

> $$M_jM_k = \left(\mathbb{I} - m_je_j^T\right)\left(\mathbb{I} - m_ke_k^T\right) = \mathbb{I} - m_je_j^T - m_ke_k^T + m_je_j^Tm_ke_k^T.$$
>
> If $j < k$, then by sparsity pattern of e_j and m_k, $e_j^Tm_k = 0$ since the entries of e_j in positions $k+1, \dots, n$ is 0 while the jth entry of m_k is 0. Thus, $M_jM_k = \mathbb{I} - m_je_j^T - m_ke_k^T$.

(c) Consider the matrix:

$$A = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 3 & 0 & 0 \\ 0 & 9 & 4 & 9 \\ 5 & 0 & 8 & 10 \end{bmatrix}$$

Determine a ***unit lower triangular*** matrix M and an ***upper triangular*** matrix U such that $MA = U$. (5 points)

We have:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -5 & 0 & 0 & 1 \end{bmatrix} \implies A_1 = M_1 A = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 3 & 0 & 0 \\ 0 & 9 & 4 & 9 \\ 0 & 0 & 8 & 10 \end{bmatrix}$$

$$M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \implies A_2 = M_2 A_1 = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 4 & 9 \\ 0 & 0 & 8 & 0 \end{bmatrix}$$

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} \implies U = M_3 A_2 = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 4 & 9 \\ 0 & 0 & 0 & -18 \end{bmatrix}$$

Thus, $M = M_3 M_2 M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \\ -5 & 6 & -2 & 1 \end{bmatrix}$ and $U = M_3 A_2 = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 4 & 9 \\ 0 & 0 & 0 & -18 \end{bmatrix}$.

Note: The computation in part (b) does not apply here, and M has to be explicitly computed!

Problem 20: Secant method for root finding

Recall that Newton's method for finding a root of $f(x) = 0$ is:

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

(a) Show how you would use Secant method to find $\sqrt{5}$. (1)

We can compute $\sqrt{5}$ by solving for the root of $f(x) = x^2 - 5$. Then, Secant method can be applied to finding this root as:

$$x_{k+1} = x_k - (x_k^2 - 5) \frac{x_k - x_{k-1}}{x_k^2 - 5 - x_{k-1}^2 + 5} = x_k - (x_k^2 - 5) \frac{x_k - x_{k-1}}{(x_k - x_{k-1})(x_k + x_{k-1})} = x_k - \frac{(x_k^2 - 5)}{(x_k + x_{k-1})},$$

$$\implies x_{k+1} = \frac{x_k^2 + x_k x_{k-1} - x_k^2 + 5}{x_k + x_{k-1}} = \frac{x_k x_{k-1} + 5}{x_k + x_{k-1}},$$

$$\implies x_{k+1} = \frac{x_k x_{k-1} + 5}{x_k + x_{k-1}}.$$

(b) Starting with initial value $x_0 = 0$ and $x_1 = 1$, perform two steps of secant method to compute

x_2 and x_3 for part (a). You can show your solution to 2 digits of precision after the decimal place. (10 points)

$$\begin{aligned}
x_2 &= \frac{x_1 x_0 + 5}{x_1 + x_0} = \frac{1 \cdot 0 + 5}{1 + 0} = \frac{5}{1} = 5, \\
x_3 &= \frac{x_2 x_1 + 5}{x_2 + x_1} = \frac{5 \cdot 1 + 5}{5 + 1} = \frac{10}{6} = \frac{5}{3} \approx 1.67 \\
x_4 &= \frac{x_3 x_2 + 5}{x_3 + x_2} = \frac{5/3 \cdot 5 + 5}{5/3 + 5} = \frac{40}{20} = 2 \\
x_5 &= \frac{x_4 x_3 + 5}{x_4 + x_3} = \frac{2 \cdot 5/3 + 5}{2 + 5/3} = \frac{25}{11} \approx 2.27
\end{aligned}$$

Note: I have shown some additional iterations for the purpose of illustration.

Problem 21: Linear least squares

Let $A \in \mathbb{R}^{m \times n}$ $(m \geq n)$, and $b \in \mathbb{R}^m$. Prove that a solution to the linear least squares problem: $Ax \simeq b$ always exists. Further, prove that such a solution is unique if and only if $\text{rank}(A) = n$ (that is, A is full rank).

Let us explicitly denote $f(x) := \|b - Ax\|_2^2$. Then we have that:

$$\|r\|_2^2 = r^T r = (b - Ax)^T (b - Ax) = b^T b - b^T A x - x^T A^T b + x^T A^T A x.$$

Since we seek a minimizer to $\|r(x)\|_2$, this is equivalent to finding the minimizer for $\|r(x)\|_2^2$. Thus, from the first-order necessary condition for existence of a minimum, we need that:

$$\nabla f(x) = 0,$$

that is, the gradient of $f(x)$ is 0. This gives us:

$$\nabla f(x) = \nabla \left(b^T b - b^T A x - x^T A^T b + x^T A^T A x \right) = -2A^T b + 2A^T A x = 0.$$

Thus, for this necessary condition for existence of a minimizer, we have that for x:

$$A^T A x = A^T b.$$

To check that this is indeed the minimizer, we need to verify the second-order sufficiency condition by computing the Hessian matrix. For x to be minimizer, we require that $H_f(x)$ be positive definite. For, $f = \|b - Ax\|_2^2$, we get that:

$$H_f = 2A^T A.$$

Now, given $A \in \mathbb{R}^{m \times n}$, $m \geq n$, $A^T A$ is positive definite if and only if A is full rank.

Proof. Let us first show the claim that A is full rank if and only if $A^T A$ is positive definite.

($\Longrightarrow$) Since A full rank, for all $x \neq 0$, $Ax \neq 0$. Thus, $\|Ax\|_2^2 = (Ax)^T(Ax) = x^T A^T A x > 0$ by property of norms. Hence, A postive definite.

($\Longleftarrow$) Given that $A^T A$ is positive definite, this means that for any nonzero x, $x^T(A^T A)x > 0$. Equivalently, $x^T A^T A x > 0 \implies \|Ax\|_2^2 > 0$ hence, $Ax \neq 0$ and so A is full rank. □

This shows that the solution to the normal equations is unique if A is full rank; otherwise, $A^T A$ is at least positive semidefinite and there exists infinitely many solutions for the normal equations.

Problem 22: Finite difference approximation

Given a smooth function $f : \mathbb{R} \to \mathbb{R}$, we wish to approximate its first and second derivatives at a given point x. For a given step size h, consider the following Taylor series expansions:

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f'''(x)}{6}h^3 + \cdots,$$

and

$$f(x-h) = f(x) - f'(x)h + \frac{f''(x)}{2}h^2 - \frac{f'''(x)}{6}h^3 + \cdots.$$

Solving for $f'(x)$ in each of these two series yields, respectively, the forward and backward difference formulas.

(a) Derive the *forward difference* formula as stated above. What is the error in this approximation? Show all your error analysis steps for full credit. (10 points)

For obtaining the forward difference formula, simply move $f(x)$ to the left hand side in the Taylor expansion for $f(x+h)$:

$$f'(x) = \frac{f(x+h) - f(x)}{h} + \mathcal{O}(h).$$

Thus, the forward difference formula is first order accurate.

(b) Derive the *backward difference* formula as stated above. What is the error in this approximation? Show all your error analysis steps for full credit. (10 points)

For the backward difference formula, move $f(x-h)$ to the right hand side in the Taylor expansion for $f(x-h)$ and rearrange to get:

$$f'(x) = \frac{f(x) - f(x-h)}{h} + \mathcal{O}(h).$$

Thus, the backward difference formula is also first order accurate.

(c) What order accuracy results if we *average* the forward and backward difference approximations? Show all your error analysis steps for full credit.

Taking the average of the forward and backward difference formula is equivalent to:

$$\begin{aligned} f'(x) &\approx \frac{1}{2}\left(\frac{f(x+h)-f(x)}{h}+\frac{f(x)-f(x-h)}{h}\right)=\frac{f(x+h)-f(x-h)}{2h} \\ &= \frac{1}{2}\left(\frac{f''(x)}{2}h+\frac{f'''(x)}{6}h^2+\cdots-\frac{f''(x)}{2}h+\frac{f'''(x)}{6}h^2+\cdots\right) \\ &= 2\frac{f'''(x)}{6}h^2+\cdots=\mathcal{O}(h^2). \end{aligned}$$

Thus, the averaged forward and backward difference formula, also called centered difference formula, is second order accurate.