Numerical Analysis

Example Sheet

v20190519.2

Bogdan Roman, Daniel Bates, Mario Cekic

Questions can be used for supervisions. Some may be revised and improved. Corrections and contributions (questions/solutions) will be gratefully received.

1. What is the resulting absolute error when subtracting two inputs x^* and y^* that are both subjected to errors? Apply your result to $x = 1, \epsilon_x = 0.01$ and $y = 2, \epsilon_y = 0.02$.

NOTE: Here we regard $x^* = x \pm \epsilon_x$ to mean that x^* can take any value in the interval $[x - \epsilon_x, x + \epsilon_x]$, i.e. $\pm\epsilon_x$ is in fact the largest possible interval of error (for x) centred around 0, i.e. $\pm\epsilon_x = [-\epsilon_x, \epsilon_x]$, and is not a fixed quantity like ϵ_x is. Recall the lectures' warning regarding this severe abuse of notation used in practice.

2. Derive from first principles an exact expression for the resulting relative error when dividing two inputs x^* and y^* that are both subjected to errors. What would be an approximation if η_y is very small. Apply your result to $x = 5, \eta_x = 5\%$ and $y = 8, \eta_y = 25\%$.

NOTE: Here we regard $x^* = x(1 \pm \eta_x)$ similarly to the above question, i.e. x^* can take any value in the interval $[x - \epsilon_x, x + \epsilon_x] = [x - |x|\eta_x, x + |x|\eta_x]$, i.e. $\pm\eta_x$ is in fact regarded as the largest possible interval of relative error (for x) centred around 0, i.e. $\pm\eta_x = [-|x|\eta_x, |x|\eta_x]$, and is not a fixed quantity like η_x is. Recall the lectures' warning regarding this severe abuse of notation used in practice.

3. For multivariate functions $f(x, y)$, an approximation for the absolute error ϵ_f can be obtained when ϵ_x and ϵ_y are *small*, similarly to the unidimensional case from the lectures, i.e. $\epsilon_f = \left|\frac{\partial f}{\partial x}\right|\epsilon_x + \left|\frac{\partial f}{\partial y}\right|\epsilon_y$.

Knowing that the relative error η_x is defined as $\eta_x = \frac{\epsilon_x}{|x|}$, find an approximation using this expression for the *relative* error when multiplying two inputs x and y that are subjected to errors. How does it compare to the exact expression? Discuss a scenario when this approximation is inaccurate.

4. Answer the previous question for the case of division between two inputs x and y subjected to errors.

5. Let $x_k > 0$ for $k = 1, \ldots, n$, and $y = x_1 + x_2 + \cdots + x_n$. Show that

$$\min\{\eta_{x_1}, \eta_{x_2}, \ldots, \eta_{x_n}\} \leq \eta_y \leq \max\{\eta_{x_1}, \eta_{x_2}, \ldots, \eta_{x_n}\}.$$

6. During lectures we considered storing values with 3 significant digits, and considered $x = 9.99$ and $y = 9.98$ then looked at the relative error of $x - y$. Answer the same question for $x = 10.0$ and $y = 9.99$, which at first glance may not look any different.

7. The standard formula for the solutions to quadratic equations,

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

can sometimes suffer from loss of significance, leading to a high relative error in the output. Describe a situation where this is a problem, and derive an alternative formula which does not suffer in this case.

8. Calculate the approximate value of the following functions to 3 significant digits, and estimate the relative error, comparing it to the relative error of x:

(a) $y = x^3 \sin x$ for $x = \sqrt{2}$ using $\sqrt{2} \approx 1.414$;
(b) $y = x \sin x$ for $x = \pi$ using $\pi \approx 3.142$;
(c) $y = e^x \cos x$ for $x = \sqrt{3}$ using $\sqrt{3} \approx 1.732$.

9. Use the Lagrange error bound formula to find a reasonable bound for the absolute error in approximating the quantity $\frac{17}{\sqrt{3}}$ with a third-degree Taylor polynomial for the function $g(x) = 17/\sqrt{4-x}$ about $x^* = 0$. Compare your bound with the actual absolute error.

10. Find the minimum polynomial order of the truncated Taylor series of $f(x) = e^x$ about $x^* = 0$ such that the error is guaranteed to be at most 10^{-10} for all $x \in [-10, 10]$.

11. By considering the Lagrange error bound, show that the (infinite) Taylor series for $\cos x$ is equal to $\cos x$ for all real x.

12. Given the error approximation $\epsilon_f \approx |f'(x^*)|\epsilon_x$, find the maximum absolute error of the input such that the absolute error of the output is smaller than 10^{-6} for the following cases:

a) $y = x^3 \sin x$, $x = \sqrt{2}$;
b) $y = x \ln x$, $x = \pi$;
c) $y = e^x \cos x$, $x = \sqrt{3}$.

13. When estimating derivatives, we want to select a large h to minimise loss of significance when we subtract similar values, but we also want a small h to improve the accuracy of the Taylor expansion used to estimate the derivative. Assume that the total error in our approximation of the derivative is minimised when the rounding error and truncation errors are equal.

If the relative rounding error in every computation of f is $\eta_f = 10^{-6}$, derive expressions which can be used to choose sensible values for h when estimating the derivative of $\cos x$ using (a) $D^+_{f'}$, (b) $D^0_{f'}$, (c) $D^-_{f'}$.

14. One way to numerically compute the second derivative is by using the approximation for the first derivative twice. For example:

$$D^0_{f''} \approx \frac{\frac{f(x+h)-f(x)}{h} - \frac{f(x)-f(x-h)}{h}}{h} = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

Assuming f is infinitely differentiable on $[a, b]$, show that for all $x, x \pm h \in [a, b]$, show that the $D^0_{f''}$ has 2nd (quadratic) order approximation.

15. Compare the results of midpoint Riemann, trapezium and Simpson methods on $f(x) = x^x$, in the interval $[0, 2]$ with 4 stripes. (The true integral is 2.83388....)

16. For a function $f(x)$ and stripes of width h:

(a) Show that when using the midpoint rule, the error in one stripe's area is bounded by

$$-\frac{1}{24}h^3 M_2 + O(h^4),$$

where M_2 is the largest value of $|f''(x)|$ within the stripe's interval.
Hint: consider the antiderivative of f, $F(x)$ (i.e. $\int_a^b f(x)dx = F(b) - F(a)$ and $\frac{dF}{dx} = f$). Then represent the true integral as $\int_{x_i-\frac{h}{2}}^{x_i+\frac{h}{2}} f(x)dx$ and apply a Taylor expansion.

(b) Show that when using the trapezium rule, the error in one stripe's area is

$$\frac{1}{12}h^3 M_2 + O(h^4).$$

(c) Show that it is possible to take a weighted average of the estimates from the midpoint and trapezium rules which eliminates the third-order term, leaving us with a fourth-order method.

(d) Prove that this fourth-order method is Simpson's method.

17. Whenever we evaluate a function, there will be some relative error η in the result due to rounding.
(a) How does this error propagate through to the approximation of the integral?

(b) How does the order of a method influence our choice of stripe width?
Hint: assume that total error is minimised when errors due to rounding are equal to the approximation error of the quadrature method.

18. How many random samples must be taken for Monte Carlo integration to achieve a relative error of η?
Hint: we're dealing with random samples, so a good measure of the error is the standard deviation in our approximations of the integral.

19. Give an example of an iteration with zeroth-order convergence.

20. Compute the minimum number of iterations required by the bisection method to converge to the root with maximum absolute error ϵ.

21. Give a list of problems of the bisection method.

22. Use fixed-point iteration to approximate a root of $x^3 - 7x + 2 = 0$ for $x \in [0, 1]$. Discuss the two choices for $g(x)$.

23. Try the previous exercise for $x_0 = 2.5$.

24. Prove the fixed-point iteration convergence criterion: If $g : I \to I$ (maps I onto itself) and is differentiable on I such that $|g'(x)| < 1$ for all $x \in I$ then g has exactly one fixed point r in I and the sequence (x_n) with any $x_0 \in I$ will converge to r.

Hint: Use the mean value theorem, but realise first that if $g : I \to I$ then for any $x \in I$ we have $g(x) \in I$, and why this is important.

25. Prove the fixed-point iteration divergence criterion: If $|g'(r)| > 1$ for all fixed points r of g then all sequences (x_n) will diverge (unless $x_0 = r$).

26. Some simple computer processors do not provide a division operation, and instead use a software implementation of the Newton-Raphson method to compute a reciprocal $\frac{1}{b}$ which can then be multiplied by a numerator a to find the result of $\frac{a}{b}$. Show how it is possible to compute $\frac{1}{2.41}$ to 8 decimal places using the Newton Raphson method (*without using division*).

27. Approximate the golden ratio using the secant method, with an initial interval of $[1, 2]$. How many iterations are required to achieve a result correct to 8 decimal places?

28. Find a minimum point of $f(x) = x^4 - 3x^3 + 5x + 1$ using gradient descent. Discuss the importance of "sufficiently small" when choosing γ. (*Hint: Desmos.*)

29. Perform 2 iterations of gradient descent on $f(x, y) = x^2y^2 + x^2 + y^2 - 3x - 2y$, starting at $x = 0,\ y = 1$, using optimal step lengths. What is the gradient at the point you reach? (This is the same function as was in the slides, but with a different starting point.)

30. In the knapsack problem we have a number of items, each with a weight and a value, and a knapsack with a limited weight capacity. The goal is to choose which items should go into the knapsack to maximise their value, without exceeding the weight capacity. Suggest how this can be formulated as a simulated annealing problem.

31. Given any three (x, y) coordinates, devise a linear system to fit a quadratic function to those points. Is this a useful way of implementing Simpson's method?

32. Use Gaussian elimination (with pivoting) to solve the following:

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}$$

What happens if we don't pivot?

33. Use Gaussian elimination (without pivoting) to solve:

$$\begin{pmatrix} 0.0002 & 1.044 \\ 0.2302 & -1.624 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 1.046 \\ 0.678 \end{pmatrix}$$

Now, solve it again, but round values to 4 significant figures after every operation. Explain what you see.

34. What happens if Gaussian elimination is attempted for non-square matrices?

35. Why is it easy to compute the inverse of a triangular matrix? Demonstrate this by finding the inverse of the following matrix. What is the complexity of this method?

$$\begin{pmatrix} 1 & 3 & 5 & 4 \\ 0 & 3 & 1 & 2 \\ 0 & 0 & 6 & 7 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

Hint: the inverse of an upper-triangular matrix is also upper-triangular.

36. Use LU factorisation to solve

$$\begin{pmatrix} 2 & 4 & 3 \\ 6 & 8 & 7 \\ 4 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 11 \\ 19 \\ 14 \end{pmatrix}$$

37. Use Cholesky factorisation to solve

$$\begin{pmatrix} 4 & 6 & 2 \\ 6 & 10 & 1 \\ 2 & 1 & 14 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 4 \\ 7 \\ -9 \end{pmatrix}$$

38. How do we solve $A\mathbf{x} = \mathbf{b}$ when we have an LDL^T decomposition for A?

39. Prove that for any matrix M, MM^T is symmetric.

40. Show that all diagonal elements of a positive-definite matrix are positive.

41. Prove that if a matrix is orthogonal, so is its transpose. (*Hint:* $(AB)^T = B^T A^T$.)

42. Prove that orthogonal matrices preserve (squared) norms, i.e. that $\|M\mathbf{v}\|^2 = \|\mathbf{v}\|^2$, where M is orthogonal.

43. Find a QR factorisation of

$$\begin{pmatrix} 3 & 6 & -1 \\ -6 & -6 & 1 \\ 2 & 1 & -1 \end{pmatrix}.$$

44. Given an orthonormal set of column vectors of an incomplete Q, show how to find a new vector which can be added to the set such that the set remains orthonormal:

(a) Show that there must be a row of Q whose magnitude is less than 1. (i.e. The row vectors are not yet orthonormal.)

(b) Starting with a column vector which is zero everywhere except the row found previously, generate a column vector which can be added to the set without breaking orthonormality.

(c) When is this useful?

45. Define the condition number K of a function f to be:

$$K(x) = \left| \frac{xf'(x)}{f(x)} \right|$$

What can you say about the condition number of $\cos x$?

46. Using $\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n}\right)^T$ show that:

(a) $\nabla \mathbf{x}^T \mathbf{b} = \mathbf{b}$

(b) $\nabla \mathbf{b}^T \mathbf{x} = \mathbf{b}$

(c) $\nabla \mathbf{x}^T A\mathbf{x} = 2A\mathbf{x}$ if A is real symmetric

(d) $\nabla(\mathbf{y} - \mathbf{x})^T A(\mathbf{y} - \mathbf{x}) = 2A(\mathbf{x} - \mathbf{y})$ if A is real symmetric

(e) $\nabla \|\mathbf{y} - A\mathbf{x}\|^2 = 2A^T(A\mathbf{x} - \mathbf{y})$

47. Given an incomplete set of measurements $y = (5, 5, 3, -1)$ at times $t = (0, 1, 2, 4)$:

(a) Fit a linear function to the data using the linear least squares method. i.e. Compute $\hat{\beta}_1, \hat{\beta}_2$ such that $y_i \approx \hat{\beta}_1 t_i + \hat{\beta}_2$.

(b) Compute the sum of squared residuals caused by this linear fit model.

(c) Compute the model's predicted (interpolated) value of the missing measurement at time 3.

48. Repeat the above exercise, but fit a quadratic function to the data (i.e. $y_i \approx \hat{\beta}_1 t_i^2 + \hat{\beta}_2 t_i + \hat{\beta}_3$).

49. Consider a set of measurements which contains a single outlier:

$$\begin{aligned} \text{inputs } x &= (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) \\ \text{measurements } y &= (0, 1, 24, 3, 4, 5, 6, 7, 8, 9) \end{aligned}$$

(a) Fit a linear model $y = \beta_0 + \beta_1 x$ using the least squares method.

(b) Fit the same linear model to the same dataset without the outlier (that is after removing 2 from the time vector and 24 from the measurements vector).

(c) Compute the goodness of fit R^2 in both cases above.

(d) Comment on why the least squares method is sensitive to such outliers?

50. Consider that instead of the standard linear least squares problem we were to minimize $S(\beta) = \|\mathbf{y} - X\beta\|^2 + \lambda\|\beta\|^2$, where $\lambda > 0$ is some scalar, and $\mathbf{y} \in \mathbb{R}^n$ is the vector of measurements, and $\beta \in \mathbb{R}^p$ is the unknown. This is called a regularization problem (*). Derive a matrix form solution $\hat{\beta}$ to this problem using the same technique as shown in the lectures. *Hint:* $\lambda x = \lambda I x$ for any vector x.
(*) Trivia: The first term is called the 'fidelity term', as it forces $X\beta$ to approximate $\mathbf{y}$, and the 2nd term is called the 'regularization term', as it forces the norm of β to be small, i.e. to regularize it, i.e. dampen large oscillations.

51. Repeat the above for $S(\beta) = \|y - X\beta\|^2 + \lambda\|W\beta\|^2$, where W is some p-by-m matrix. This is called a weighted regularization problem (*).
(*) Trivia: Here the matrix W acts on β and the minimization process tries to regularize the result $W\beta$ instead of β.

52. Consider we want to use weighted regularization to remove noise from some measured noisy signal y, i.e. to smooth y. This means we want that (a) the solution $\hat{\beta}$ to the problem to be as close to y as possible and (b) the 2nd derivative (in the discrete sense) of this solution to have low values in magnitude.

(a) Why is a function with smaller 2nd derivative in magnitude smoother?

(b) The discrete 1st derivative of a vector $x = [x_i], i = 1, \ldots, n$ has the form $x' = [x_i - x_{i-1}], i = 2, \ldots, n$. Find the matrix D that implements this operation, i.e. find D such that $Dx = x'$.

(c) Using the above matrix D give the expression and contents of the matrix E that implements the discrete 2nd derivative of x.

(d) Formulate the weighted regularization problem that smooths y. That is, identify the matrices X and W (Hint: obviously, one of them should be E).

(e) Derive the solution to this problem in matrix form.

(f) How do small vs large values of λ influence the result?

53. Implement the above in either Matlab/Octave as follows. Fetch the y vector of measurements from https://www.cl.cam.ac.uk/teaching/1819/NumAnalys/matlab/y_noise.txt and copy the entire text to clipboard. Then go to https://octave-online.net and input the following commands one by one:

```
y = [<paste from clipboard>];
N = length(y);
plot(y)
E = spdiags(ones(N,1) * [1 -2 1], 0:2, N-2, N);
lambda = 50;
F = speye(N) + lambda * E' * E;
x = F \ y;
plot(x)
```

Repeat the last 4 commands for $\lambda \in \{0.5, 5, 50, 500, 5000\}$, observe what happens, and comment on the result.

Data and example courtesy of Ivan Selesnick, NYU.

54. Show that the eigenvalues of a triangular matrix are its diagonal entries.

55. Let A, B, P be n-by-n matrices, and P be invertible. Show that if $B = P^{-1}AP$ (we say A and B are 'similar') then B and A have the same eigenvalues.

56. *[Recall Google's PageRank?]* For a square matrix A where every column sums to the same value k, show that k is an eigenvalue of A.

57. Given the eigenvectors $(\mathbf{v}_1, ..., \mathbf{v}_n)$ and eigenvalues $(\lambda_1, ..., \lambda_n)$ of a square matrix A, and a constant k, what can we say about the eigenvectors and eigenvalues of:

(a) kA

(b) A^k

(c) A's inverse

(d) A's transpose

58. What are the possible eigenvalues of an *idempotent* matrix (i.e. $A^2 = A$)?

59. Find the eigenvalues and eigenvectors of:

(a) The reflection matrix, $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

(b) A 90° rotation matrix, $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$.

(c) The transpose transformation.

60. Given that most computer languages today support 32-bit signed integers and double-precision floating point, can it be argued that having the integers is silly since they are a subset of the doubles?

61. Sketch a proof that integer comparison predicates can be applied to the bit patterns of IEEE unsigned floating point and mention any exceptions. (Or do a structured proof if feeling ambitious).

62. What do the following single precision IEEE bit patterns represent, where the most significant bit of the first-listed byte is the sign bit?

a) 00 00 00 00

b) 80 00 00 00

c) BF 01 00 00

d) 3F C0 00 00

e) 04 04 04 00

63. How is it possible to determine the value of $\sin x$ for any x, while only ever computing the Taylor expansion of $\sin x$ for $x \in [0, \frac{\pi}{4}]$? Show how pre-processing and post-processing are performed.

Hint: you might need to compute $\cos x$ *in this range too.*

64. The McLaurin series for $\ln(1 + x)$ is:

$$x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots \quad \text{for } |x| < 1.$$

Describe a range reduction procedure which ensures the series always acts on small (floating point) values of x. Show all pre- and post-processing steps.