Exercises Numerical Algorithms

Chris Stolk

November 19, 2021

These exercise are to be done using pen and paper only, unless otherwise noted. The section numbers correspond to the chapter numbers in Heath. When the section number is a letter, it concerns additional material.

(Some of these exercises are copied or adapted from Heath and other sources, some are made by the author.)

1 Scientific computing

1.1 Consider the problem of evaluating the function $\sin(x)$, in particular, the propagated data errors, i.e., the error in the function value due to a perturbation h in the argument x.

  (a) Estimate the absolute error in evaluating $\sin(x)$.

  (b) Estimate the relative error in evaluating $\sin(x)$.

  (c) Estimate the condition number for this problem.

  (d) For what values of the argument x is this problem highly sensitive?

1.2 Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = x - y$.

  (a) Measuring the size of the input (x, y) by $|x| + |y|$, and assuming that $|x| + |y| \approx 1$ and $x - y \approx \epsilon$, show that $\text{cond}(f) \approx 1/\epsilon$. What can you conclude about the sensitivity of substraction?

  (b) Estimate the condition number of subtraction in case $x = 10.01$ and $y = 10.0$. (N.B. since you are supposed to do this without a calculator you may approximate any numerical calculations and be 10 % off.)

1.3 In this exercise we consider truncation errors and rounding errors for numerical differentiation (cf. example 1.3). We assume the derivative $f'(x)$ is approximated by

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}.$$

  (a) Let M be a bound on $|f'''(t)|$ for t near x. Show that the truncation error in this approximation of $f'(x)$ can be estimated by $C_1 M h^2$, and determine a value for the constant C_1.

(b) Estimate the combined effect of rounding errors and cancellation. The result estimate should be of the form $C_2 h^{p_1} \epsilon_{rmmach}^{p_2}$, where C_2, p_1 and p_2 are constants that you have to determine. Assume that f can be computed to machine precision.

(c) Estimate the total error by the sum of the truncation error obtained in part (a) and the error computed in part (b). Determine a formula for the choice of h where the total error is minimal. Your formula should be of the form $C_3 \epsilon \text{mach}^{\alpha}$. What is α?

1.4 The following formulas are mathematically equivalent

$$\begin{array}{lll} a_1 = (\sqrt{2}-1)^6 & a_2 = (3-2\sqrt{2})^3 & a_3 = 99-70\sqrt{2} \\ a_4 = (\sqrt{2}+1)^{-6} & a_5 = (3+2\sqrt{2})^{-3} & a_6 = (99+70\sqrt{2})^{-1} \end{array}$$

Because of its finite precision, the computer approximates $\sqrt{2}$ by $\sqrt{2}(1+\epsilon)$. Which of the six formulas above gives the least accurate approximation of $(\sqrt{2}-1)^6$? Which formula gives the best result? You may use the result of exercise 1.2(a).

A Review of linear algebra

A.1 Solve the linear system $Ax = b$ for the following values of A and b,

(a) $A = \begin{bmatrix} 2 & -1 \\ 1 & 3 \end{bmatrix}$ and $b = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$.

(b) $A = \begin{bmatrix} 1 & -1 & -1 \\ 3 & -3 & 2 \\ 2 & -1 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 2 \\ 16 \\ 9 \end{bmatrix}$.

A.2 Find all solutions to the linear system $Ax = b$ in case

(a) $A = \begin{bmatrix} 1 & -2 & 2 \\ 3 & 2 & -1 \\ 2 & 4 & -3 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}$

(b) $A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$.

A.3 For which a is the following matrix of rank 3? Of rank 2? Of rank 1?

$$\begin{bmatrix} 1 & 1 & 3 & 1 \\ 2 & -1 & 0 & 1+a \\ -3 & 2 & 1 & -2 \end{bmatrix}$$

A.4 Prove that if R is a matrix in echelon form, then a basis for the row space of R consists of the nonzero rows of R.

2 Systems of linear equations

2.1 (a) Let $A = \begin{bmatrix} 1 & -1 & \alpha \\ 2 & 2 & 1 \\ 0 & \alpha & -3/2 \end{bmatrix}$ where α is a number in $\mathbb{R}$. For which α is A singular?

(b) Consider the following linear system of equations:

$$\begin{aligned} 2x + y + z &= 3 \\ 2x - y + 3z &= 5 \\ -2x + \alpha y + 3z &= 1. \end{aligned}$$

For what values of α does this system have an infinite number of solutions?

(c) Denote the columns of an $n \times n$ matrix A as A_k for $k = 1, \ldots, n$. We define the function $||A||_* = \max_k ||A_k||_2$. Show that $||A||_*$ is a norm, in that it satisfies the first three properties of a matrix norm (cf. §2.3.2).

2.2 Give the LU decomposition (without pivoting) of the following matrices

(a) $A = \begin{bmatrix} 2 & 2 & -1 \\ 4 & 0 & 4 \\ 6 & 2 & 10 \end{bmatrix}$.

(b) $A = \begin{bmatrix} 1 & 0 & 1 \\ a & a & a \\ b & b & a \end{bmatrix}$. In this case, also determine for which a, b the decomposition exists.

2.3 (a) Let $A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & -2 \\ 3 & 1 & 3 \end{bmatrix}$. Determine the matrix norms $\|A\|_1$ and $\|A\|_\infty$.

(b) Determine the condition number (with respect to the 2-norm) of $A = \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix}$.

(c) Let $A = \begin{bmatrix} 1 & 1.01 \\ 0.99 & 1 \end{bmatrix}$. Show using a pen-and-paper calculation that $\|A\|_2 \geq 1$. Determine A^{-1} and show that $\|A^{-1}\|_2 \geq 10000$. What can be concluded about $\operatorname{cond}(A)$?

2.4 (Diagonally dominant matrices) A matrix is said to be *column diagonally dominant* if, for each column j, the absolute value of the diagonal entry is greater than the sum of the absolute values of the off-diagonal entries, i.e., if

$$|a_{jj}| > \sum_{i, i \neq j} |a_{ij}|.$$

Show that after one step of Gaussian elimination with no pivoting, the remaining $(n-1) \times (n-1)$ submatrix is also column diagonally dominant. Deduce that no row exchanges will occur throughout the elimination process, even when partial pivoting is used.

2.5 Let A be a symmetric positive definite $n \times n$ matrix.

(a) We write A in block form as

$$A = \begin{bmatrix} a_{1,1} & w^* \\ w & K \end{bmatrix}.$$

Show that $a_{1,1}$ is positive and K is positive definite.

(b) In the first step of a Choleksy factorization algorithm A is written as

$$\begin{aligned} A &= \begin{bmatrix} a_{1,1} & w^* \\ w & K \end{bmatrix} \\ &= \begin{bmatrix} \alpha & 0 \\ w/\alpha & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & X \end{bmatrix} \begin{bmatrix} \alpha & w^*/\alpha \\ 0 & I \end{bmatrix} \end{aligned}$$

Express α and X in terms of K, w, and a_{11}.

(c) Argue that X must be positive definite.

2.6 Suppose we have a $2n \times 2n$ matrix A of the form $A = \begin{bmatrix} B & O \\ O & C \end{bmatrix}$ where B and C are nonsingular matrices of size $n \times n$. Suppose we want to solve a system $Ax = b$.

(a) What is the cost of computing an LU decomposition of A in the usual way?

(b) What is the cost of computing LU decompositions of B and C?

(c) Decompose b as $b = \begin{bmatrix} c \\ d \end{bmatrix}$, where c and d are vectors of length n. Explain how the system $Ax = b$ can be solved using the LU decompositions of B and C directly, while not computing the LU decomposition of A. Estimate the factor by which the computational cost is reduced compared to the situation where the LU decomposition of A is computed in the usual way.

2.7 Suppose we write a $(p+q) \times (p+q)$ matrix M in block form where A, B, C, D are respectively $p \times p, p \times q, q \times p$ and $q \times q$ matrices

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

(a) Verify that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ CA^{-1} & I_q \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I_p & A^{-1}B \\ 0 & I_q \end{bmatrix}$$

(b) Describe how a system $Mx = b$, with x and b in $\mathbb{R}^{p+q}$, can be solved by applying matrix-vector products with C and B and solves with A and $(D - CA^{-1}B)$.

(c) Suppose $p = 2m$ and $q = m$. What is the cost, to highest order, of LU-factorizing A and computing and LU-factorizing $D - CA^{-1}B$? Show that this cost, to highest order, is the same as that of factorizing M directly.

Although in this case no savings were obtained, the decomposition above is very useful for solving linear systems with many zeros, in other words where M is a sparse matrix. After applying a permutation of the indices such a matrix is written in the above form, where q is as small as possible and A is blockdiagonal, i.e. $A = \begin{bmatrix} E & O \\ O & F \end{bmatrix}$. This block-diagonal form then causes big savings in computational cost. Moreover, the procedure can be applied recursively.

3 Linear least squares

3.1 (a) Solve the problem of fitting a straight line to the three data points $(-1, 1)$, $(1, 2)$, $(2, 3)$ using the least-squares approach.

(b) Consider the problem of fitting a second order polynomial to the five data points $(-1, 0)$, $(0, 2)$, $(1, 2)$, $(3, 3)$, $(4, 2)$. Setup the overdetermined linear system for the least-squares problem.

3.2 (a) Let v be a vector in $\mathbb{R}^m$. What is the matrix associated with a Householder reflection in the hyperplane orthogonal to v.

(b) Show that this matrix is orthogonal.

3.3 (a) Consider the plane in $\mathbb{R}^3$ given by the equation

$$x_1 + x_2 + x_3 = 0.$$

Construct a matrix P which projects a given point on this plane. Hint: consider first the orthogonal complement of the plane.

(b) Consider the plane in $\mathbb{R}^3$ given by the equation

$$2x_1 - 2x_3 = 0$$

Construct a matrix $P \in \mathbb{R}^{3\times 3}$ which reflects a given point at this plane (computes the mirror image).

3.4 Suppose you are computing the QR factorization of the matrix

$$A = \begin{bmatrix} 1 & 4 & 1 \\ 1 & 0 & 1 \\ 3 & 3 & -1 \\ -1 & 0 & 1 \end{bmatrix}$$

by Householder transformations.

(a) How many Householder transformations are required?

(b) Determine the first Householder transformation. It is sufficient to determine the Householder vector v.

(c) Consider a general linear least squares problem $Ax \cong b$, where A is an $m \times n$ matrix, $m > n$. How can this problem be solved, assuming that a QR factorization $A = QR$ is given?

3.5 Let $A = U\Sigma V^T$, where $U \in \mathbb{R}^{3\times 3}$ is orthogonal, $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix}$, $V = \frac{1}{2}\begin{bmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{bmatrix}$, and $b := U \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}$.

(a) Verify that V is orthogonal

(b) Find $x \in \mathbb{R}^2$ that minimizes $\|Ax - b\|_2$.

3.6 Let A be diagonal $n \times n$ matrix with diagonal entries λ_j, $j = 1, \dots, n$, and suppose that the λ_j are real and non-negative and satisfy $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$.

(a) Prove from the definition of the matrix norm that $\|A\|_2 = \lambda_1$.

(b) Let B be a general real $n \times n$ matrix, with singular value decomposition $B = U\Sigma V^T$, and let σ_j, $j = 1, \dots, n$ be the singular values. What is the value of $\|B\|_2$? Derive this from the definition of the matrix norm (where you may use part (a)).

(c) What is the value of $\mathrm{cond}(B)$ (condition number using the matrix 2-norm)? Derive this from the definition and the previous parts of this exercise.

3.7 Let A be an $m \times n$ real matrix, $b \in \mathbb{R}^m$, and let $\gamma > 0$ be a real constant. We consider the problem of finding $x \in \mathbb{R}^n$ that minimizes the function

$$\|Ax - b\|^2 + \|\gamma x\|^2.$$

(a) Show that this amounts to solving the $(m+n) \times n$ linear least squares problem

$$\begin{bmatrix} A \\ \gamma I \end{bmatrix} x \cong \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{*}$$

(b) Formulate the normal equations for this problem

Let $A = U\Sigma V^T$ be the singular value decomposition of A and let σ_j be the singular values.

(c) It turns out that the solution to (*) is of the form

$$V \begin{bmatrix} T & O \end{bmatrix} U^T b.$$

where T is an $n \times n$ diagonal matrix in which the (j, j) entry is given by a formula of the form $f(\sigma_j, \gamma)$, for some function $f : \mathbb{R}^2 \to \mathbb{R}$ and O is an $n \times (m - n)$ block of zeros. Prove this statement and determine the function $f(\sigma, \gamma)$.

N.B. This procedure is called Tikhonov regularization. It is useful in parameter estimation problem where the linear least squares problem is ill-posed, i.e. the associated matrix has a large condition number.

4 Eigenvalue problems

4.1 Let $A = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}$.

(a) Compute the eigenvalues and eigenvectors of A.

(b) Which eigenvalue of this matrix is estimated by the power method?

(c) Compute 2 iterations of the power method with starting vector $x_0 = [1, 0]^T$ and estimate an eigenvalue from your results.

(d) Also estimate this eigenvalue using the Rayleigh quotient.

4.2 Let A be a diagonalizable $n \times n$ matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ for which $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$. For some arbitrary $x^{(0)}$, let $x^{(i+1)} := Ax^{(i)}$. Explain why, in nearly all cases, $\lim_{i\to\infty} \frac{\|x^{(i+1)}\|}{\|x^{(i)}\|} = |\lambda_1|$.

4.3 Let $A \in \mathbb{R}^{n\times n}$, and assume that v is an eigenvector of A with eigenvalue λ. Let σ be a constant, such that σ is not equal to an eigenvalue of A.

(a) Show that v is also an eigenvector for $(A - \sigma I)^{-1}$.

(b) What is the corresponding eigenvalue?

4.4 Consider a $(p + q) \times (p + q)$ matrix M that can be written in block form as

$$M = \begin{bmatrix} A & B \\ O & C \end{bmatrix}$$

where A is of size $p \times p$ and C is of size $q \times q$. Suppose λ is an eigenvalue of C with eigenvector w and λ is not an eigenvalue of A. Show that λ is an eigenvalue of M and determine the eigenvector in terms of A, B, C and w. (Hint: Write the eigenvector in the form $\begin{bmatrix} u \\ w \end{bmatrix}$, where u and v are vectors of length p and q respectively.)

5 Nonlinear equations

5.1 Consider the problem of solving the following nonlinear equation

$$x^3 = 3. \tag{*}$$

In the following, the values 1 and 2 may be used as starting points (choose the number of starting points appropriate for the method).

(a) Perform two steps of the bisection method and estimate the solution to (*).

(b) Perform one step of Newton's method and estimate the solution to (*).

(c) Perform one step of the secant method and estimate the solution to (*).

(d) What do you know about the convergence rates of the above three methods?

(e) Give a particular advantage of the bisection method. Same question for Newton's method and the secant method.

5.2 Carry out one iteration of Newton's method applied to the system

$$\begin{aligned} x_1 - x_2^2 &= 1 \\ x_1 + x_2 &= 5 \end{aligned}$$

with starting value $x_0 = [0, 2]^T$.

5.3 For each of the functions g below, consider the fixed point iteration associated with it. Answer the following question (i) determine all the fixed points; (ii) for the largest (right-most) fixed point, determine whether the fixed point iteration converges to it when the starting point is close enough to it; (iii) if convergence occurs, determine the rate of convergence.

(a) $g(x) = x^2 - 6$

(b) $g(x) = \frac{x^2+6}{2x-1}$

5.4 (a) Let f be a function $\mathbb{R} \to \mathbb{R}$ that is twice continuously differentiable. Observe that Newton's method for f can be written in the form of a fixed-point iteration

$$x^{(k+1)} = g(x^{(k)}).$$

Give an expression for g in terms of f.

(b) Give the definition of quadratic convergence of an iteration.

(c) Let x^* be a root of f such that $f'(x^*) \neq 0$. Show that Newton's method converges quadratically to x^* if started close enough to x^*. You may use the convergence properties of fixed-point iteration.