**Solutions to Exam on October 27, Numerical analysis I, 2021**

(1) Let $A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$ and $b = \begin{pmatrix} 4 \\ 3 \\ 3 \\ 1 \end{pmatrix}$. In computation of the solution to the equation $Ax = b$ we know that $b$ is perturbed by a vector $\delta b$ with $\|\delta b\|_\infty \leq 0.01$.

(a) Give an upper bound for $\|\delta x\|_\infty$, where $\delta x$ is the associated perturbation in the computed solution.

(b) Compute the condition number $\kappa_\infty(A)$ and compare it with the quotient between $\|\delta x\|/\|x\|$ och $\|\delta b\|/\|b\|$. Is the upper bound obtained by perturbation analysis tight?

*Solution.* (a) A straightforward computation gives $\|b\|_\infty = 4$, $\|A\|_\infty = 33$, $\|A^{-1}\|_\infty = 136$. Since $A(x + \delta x) = b + \delta b$, which is $A\delta x = \delta b$, we have $\delta x = A^{-1}\delta b$. Then

$$\|\delta x\|_\infty \leq \|A^{-1}\|_\infty \|\delta b\|_\infty = 136 \cdot 0.01 = 1.36.$$

(b) The solution $x = (1, -1, 1, -1)^T$, and $\|x\|_\infty = 1$. The condition number $\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = 33 \cdot 136 = 4488$. So

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} = 1.36, \quad \frac{\|\delta b\|_\infty}{\|b\|_\infty} = \frac{0.01}{4} = 0.0025 \quad \Rightarrow \text{ ratio} = \frac{1.36}{0.0025} = 544.$$

The perturbation analysis shows that this ratio is less than the condition number which is much larger. So this upper bound is not tight.

(2) Assume that the function $f(x)$ is three times continuously differentiable and $\alpha$ is a zero of $f$ but not a zero of its derivative.

(a) Show that the iteration

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2[f'(x_n)]^2 - f(x_n)f''(x_n)}, n = 1, 2, ...$$

can be obtained by applying Newton's method to the function $g(x) = \frac{f(x)}{\sqrt{|f'[x])|}}$.

(b) Argue that when the second derivative is very close to zero, the iteration is almost the same as the Newton's method iteration

(c) Show that, if $\{x_n\}$, $n = 0, 1, 2, ...$, generated by the above iteration converges in a neighborhood of $\alpha$, then the convergence is cubic.

*Solution.* (a) Applying Newton's method to $g$ ives

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

with

$$g'(x) = \frac{2[f'(x)]^2 - f(x)f''(x)}{2f'(x)\sqrt{|f'(x)|}},$$

and the result follows.

(b) When the second derivative is very close to zero, then

$$x_{n+1} \approx x_n - \frac{2f(x_n)f'(x_n)}{2[f'(x_n)]^2} = x_n - \frac{f'(x_n)}{f_(x_n)}$$

which is Newton's method (c). The easiest way (but pretty tedious though) is to compute the derivatives of the function

$$\varphi(x) = x - \frac{2f(x)f'(x)}{2[f'(x)]^2 - f(x)f''(x)},$$

that is

$$\varphi'(x) = \frac{f(x)^2\left(3f''(x)^2 - 2f^{(3)}(x)f'(x)\right)}{\left(f(x)f''(x) - 2f'(x)^2\right)^2}$$

and

$$\varphi''(x) = \frac{h(x)}{\left(2f'(x)^2 - f(x)f''(x)\right)^3},$$

where

$$\begin{aligned} h(x) = 2f(x)\Big(f'(x)^3\left(6f''(x)^2 - 2f(x)f^{(4)}(x)\right) + 12f(x)f^{(3)}(x)f'(x)^2f''(x) + \\ + f(x)f'(x)\left(-12f''(x)^3 + f(x)f^{(4)}(x)f''(x) - 2f(x)f^{(3)}(x)^2\right) \\ -4f^{(3)}(x)f'(x)^4 + f(x)^2f^{(3)}(x)f''(x)^2\Big). \end{aligned}$$

Clearly $\varphi'(\alpha) = \varphi''(\alpha) = 0$. So the convergence is at least cubic.

(3) Assume that the function $f$ is sufficiently smooth. Let $x_i = x_0 + ih$ and $h > 0$
  (a) Show that the formula
$$f(x_{\frac{1}{2}}) \approx \frac{1}{2}f(x_0) + \frac{1}{2}f(x_1) + \frac{1}{8}hf'(x_0) - \frac{1}{8}hf'(x_1)$$
  is exact for all third degree polynomials.
  (b) Derive an asymptotical (approximation) error estimate.
  (c) Use the formula and error estimate to determine $f(x) = e^{1/2}$, $x_0 = 0, x_1 = 0.2$ using 6-decimals. (Note that $e^{0.2} = 1.221403$.)

*Solution* See the textbook on pages 265-266.

(4) (a) What is the characteristic polynomial of the matrix
$$F = \begin{pmatrix} 0 & 0 & \cdots & 0 & -\gamma_0 \\ 1 & 0 & \cdots & 0 & -\gamma_1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -\gamma_{n-1} \end{pmatrix}?$$
  (b) Let $p(\lambda) = a_n\lambda^n + \cdots + a_0$ och $\gamma_i = a_i/a_n$, $i = 0, ..., n-1$ and $a_n \neq 0$. Show how Gersgorin's Theorem can be applied to obtain the statement that all the zeros of $p(\lambda)$, $\lambda_1, ..., \lambda_n$ satisfy
  (i) $|\lambda_i| \le \max\left\{\left|\frac{a_0}{a_n}\right|, \max_{1\le k\le n-1}\left(1 + \left|\frac{a_k}{a_n}\right|\right)\right\}$
  (ii) $|\lambda_i| \le \max\left\{1, \sum_{k=0}^{n-1}\left|\frac{a_k}{a_n}\right|\right\}$.
  (c) Compare these two estimates for $p(\lambda) = \lambda^3 - 2\lambda^2 + \lambda - 1$.
  (d) How would you solve polynomial equations, especially the polynomial has multiple zeros?

*Solution.* (a) Using induction we can show that the characteristic polynomial of the matrix $F$ is

$$\chi_F(s) = s^n + \gamma_{n-1}s^{n-1} + \cdots + \gamma_1 s + \gamma_0$$

(b) The Gerschgoring's discs for $F$ is

$$D_0 = \{z : |z| \le |\gamma_0|\}$$
$$D_i = \{z : |z| \le 1 + |\gamma_i]\}, \ i = 1, ..., n-2,$$
$$D_{n-1} = \{z : |z + \gamma_{n-1}| \le 1\}$$

Together with revered triangle inequality (for the last disc), we have that all eigenvalues of $F$, $\lambda_1, ..., \lambda_n$, satisfy

$$|\lambda_i| \le \max\left\{|\gamma_0|, \max_{1\le k\le n-1}(1 + |\gamma_k|)\right\}, i = 1, ..., n$$

Now the polynomial $p$ has the same zeros as the eigenvalues of $F$ and $\gamma_i = a_i/a_n$ we get the estimate in (i).

Since the eigenvalues of $F$ and the eigenvalues of $F^\top$ are the same, applying the Gercshgoring Theorem on $F^\top$ yields the following discs

$$\{z : |z| \le 1\}\text{and } \{z : |z + \gamma_{n-1}| \le \sum_{k=0}^{n-2} |\gamma_k|\}.$$

Using the same argument as in (i) we have

$$|\lambda_i| \le \left\{1, \sum_{k=0}^{n-1} |\gamma_k|\right\}, i = 0, 1, ..., n.$$

(c) By (b) (i) $|\lambda_i| \le 3$ (ii) $|\lambda_i| \le 4$. (i) gives better estimate.

(d) We convert the problem of finding zeros of a polynomial to the eigenvalues of the companion matrix such as $F$ above using for example shifter QR-algorithm. There are more efficient QR-algorithms for the companion matrix $F$ since it is a rank one update of an orthogonal matrix.

(5) Consider the initial value problem $y'(x) = f(x, y(x))$, $y(x_0) = y_0$.

(a) Derive both implicit and explicit Euler's methods for solving of this problem. Name, for each of them, at least one advantage and disadvantage, respectively.

(b) Determine the region where the methods are absolutely stable for $f(x, y) = ay$, where $a$ is a (possibly complex) constant.

(c) When is implicit Euler's method preferable? Why?

*Solution.* See e.g. the textbook pages 338, 350 for derivation of the Euler's method.

Euler's explicit method: $y_{n+1} = (1 + ha)y_n$. So $y_n \to 0$ if and only if $|1 + h| < 1$. So the absolutely stable region is a unit disc centered at $(-1, 0)$ on the complex plane, which is a subset of the left half plane, where $z = ha$.

Euler's implict method: $y_{n-1} = (1 - ah)^{-1}$. So $y_n \to 0$ if and only if $|1 - ah| > 1$. The absolutely stable region is the complex plane excluding the unit disc (including the circle) centered at $(1, 0)$. This include the whole left half-plane.

The explicit method is cheaper but the step size is limited (which is not suitable for stiff problem). The implicit method is more expansive since it needs solve (in general) a nonlinear equation at each step. But it is more suitable for some problems such as stiff problems

**You have finished the exam if your homework point $p_h \ge 15$ (i.e. p=20). Do (6a) if $p_h \in [10, 15)$ (i.e. p=10); do (6a) and (6b) if $p_h \in [5, 10)$ (i.e. p=5). Note that all your $p_h$ will be added.**

(6) (a) Let $y := \varphi(p, q) = -p + \sqrt{p^2 + q}$.

(i) Given the relative input errors $\varepsilon_p$ $\varepsilon_q$, determine the relative output error of the result $y$.

(ii) Show that the problem is well conditioned for $p > 0$, $q > 0$.

(iii) Propose a numerically stable algorithm to compute $y$.

(b) Consider a symmetric $n \times n$ matrix $A$.

(i) Show that the eigenvalue problem is well-condtioned.

(ii) Assume further that $A$ is symmetric positive definite tridiagonal. Propose an $O(n)$ running time algorithm to compute the Cholesky factor.

(iii) The finite difference method applied to the two-point boundary value problem: $\frac{d^2y}{dx^2} = 12x^2, 0 \leq x \leq 1$ with $y(0) = y(1) = 0$, using $x_j = 0 + (j-1)h$, $(j = 1, ..., J+1)$, results in a linear system of equations with the coefficient matrix $A$ being symmetric tridiagonal. How do you solve this system of equations? Do you invert the matrix $A$? What types of linear solver is more suitable if $J$ is very large? Write down at least one such numerical algorithm and the conditions under which the algorithm works.

(c) We can apply Newton-Raphson's method to find the positive solution of the equation $x^2 - c = 0$ to approximate $\sqrt{c}$ for $c > 0$. Write down the iteration $x_n$. Show that for all $0 < x_0 < \infty$, the sequence $\{x_n\}$ quadratically converges to $\sqrt{c}$.

*Solution.* (a)

$$\frac{\partial \varphi}{\partial p} = -\frac{y}{\sqrt{p^2+q}}, \quad \frac{\partial \varphi}{\partial q} = \frac{1}{2\sqrt{p^2+q}}$$

Error propagation theorem (Theorem 2.2.3) yilds

$$\Delta y \approx \frac{\partial \varphi}{\partial p}\Delta x + \frac{\partial \varphi}{\partial q}\Delta q = -\frac{y}{\sqrt{p^2+q}}\Delta p + \frac{1}{2\sqrt{p^2+q}}\Delta q$$

which gives

$$r_y \approx -\frac{p}{\sqrt{p^2+q}} r_p + \frac{p+\sqrt{p^2+q}}{2\sqrt{p^2+q}} r_q$$

Note that

$$\left| \frac{p}{\sqrt{p^2+q}} \right| \leq 1, \left| \frac{p+\sqrt{p^2+q}}{2\sqrt{p^2+q}} \right| \leq 1 \text{ för } q > 0.$$

So $y$ is well-conditioned if $p > 0, q > 0$ ochill-conditioned if $q \approx -p^2$.

An numerically stable algorithm can be

$$\begin{cases} s := p^2 \\ t := s + q \\ u := \sqrt{t} \\ v := p + u \\ y := q/v \end{cases}$$

Note that we don't have subtraction in the computation and the mappnig

$$\psi : u \to p + u \to \frac{q}{p+u} = \psi(u)$$

has the relative error (for $y$ is

$$\frac{1}{y}\frac{\partial\psi}{\partial u}\cdot\Delta u = \frac{-q}{y(p+u)^2}\cdot\Delta u = \underbrace{-\frac{\sqrt{p^2+q}}{p+\sqrt{p^2+q}}}_{k}\varepsilon = k\varepsilon$$

By inspection the amplifier factor $k$ satisfies $|k| < 1$. So it is numerically stable.

(b) (i) See the textbook page 209.

(ii) Let $a_1, ...a_n$ be the diagonal elements of $A$ and $b_2, ..., b_n$ be the off-diagonal elements under the diagonal. Let now the matrix $G$ be the Cholesky factor, i.e. $A = GG^T$ with the diagonal element $g_1, ..., g_n$ and off-diagonal element under the diagonal $h_2, ..., h_n$. We can prove that $g_i = \sqrt{a_i - h_i^2}$ and $h_i = b_i/g_{i-1}$ for $i = i : n$. Set $h_0 = 0$. This is an $= (n)$ algorithm.

(iii) The matrix obtained from the finite difference matrix is a tridiagonal matrix, but not positive definite. When the number of grid point is large a direct method will break down due to computer capacity. In this case an iterative method is preferred. Such algorithms are in the form $x^{(k+1)} = Bx^{(k)} + c$ (to solve $Ax = b$). Some examples are Jacobi method ($B = -(L+U)$, Gauss-Seidel metod ($B = -(I+L)^{-1}U$), (see the text book page 191 for details, or Problems 4 and 6 in the textbook page 196 and page196 respectively. The condition for its convergence is that the spectral radius of $B$ is less than 1.

(c) The iteration is

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right).$$

By some algebraic manipulation we have

$$x_{n+1} - \sqrt{c} = \frac{1}{2x_n}(x_n - \sqrt{c})^2.$$

Now for any $x_0 > 0$, the sequence is positive. Hence $x_n \geq \sqrt{c}$ for all $n = 0, 1, ....$ which yields

$$x_n - x_{n+1} = x_n - \frac{1}{2}\left(x_n + \frac{c}{x_n}\right) = \frac{x_n^2 - c}{2x_n} \geq 0$$

i.e. $x_1 \geq x_2 \geq \cdots \geq \sqrt{c}$. This means the sequence is positive, bounded and decreasing. SO it has a unique limit $a$. Solving $a = \frac{1}{2}\left(a + \frac{c}{a}\right)$ gives $a = \sqrt{c}$.