Computational math

Rodion Zaytsev

September 2020

Problem 1. *Prove that*

$$\|xy^*\|_F = \|xy^*\|_2 = \|x\|_2\|y\|_2 \quad \forall x, y \in C^n$$

Proof. By definition, of Frobenius norm,

$$\|xy^*\|_F = \sqrt{\sum |x_i y_j|^2} = \sqrt{\sum |x_i|^2}\sqrt{\sum |y_i|^2} = \|x\|_2\|y\|_2$$

On the other hand,

$$\|xy^*\|_2 = \sup_{\|z\|_2=1} \|xy^*z\|_2 = \sup_{\|z\|_2=1} |(y,z)|\|x\|_2 \leq \|y\|_2\|x\|_2$$

The supremum is achieved when $z = y/\|y\|_2$ □

Problem 2.

$$A = I + \alpha uu^*,\ \alpha \in \mathbb{C},\ u \in \mathbb{C}^n,\ \|u\|_2 = 1$$

Find all α such that A is unitary

Solution.

$$\begin{aligned} &I = AA^* = (I + \alpha uu^*)(I + \overline{\alpha}uu^*) = \\ &I + (\alpha + \overline{\alpha})uu^* + |\alpha|^2 uu^*uu^* = \\ &I + (2\operatorname{Re} + |\alpha|^2\|u\|_2^2)uu^* \\ &\Rightarrow 2\operatorname{Re}\alpha + |\alpha|^2 = 0 \end{aligned}$$

because $\|u\|_2 = 1$.
Hence, if $\alpha = x + iy$, where $x = \operatorname{Re}\alpha, y = \operatorname{Im}\alpha$, then

$$2x + x^2 + y^2 = 0 \iff (x+1)^2 + y^2 = 1$$

That is α lies on a circle. □

Problem 3. *Prove that*

$$cond(AB) \leqslant cond(A)cond(B)$$

Proof.

$$\begin{aligned} cond(AB) = \|(AB)^{-1}\|\|AB\| = \|B^{-1}A^{-1}\|\|AB\| \leqslant \\ \leqslant \|B^{-1}\|\|A^{-1}\|\|A\|\|B\| = \|A^{-1}\|\|A\|\|B^{-1}\|\|B\| = \\ = cond(A)cond(B) \end{aligned}$$

□

Problem 4. *Express $cond(B^TB)$ in terms of $cond(B)$ for 2-norm if B is real non-degenerate matrix.*

Proof. Let $B = U\Sigma V^*$ be the SVD decomposition. For 2-norm we have

$$\|B\|_2 = \max_k |\Sigma_k|$$

□

Since B is real

$$A = B^TB = B^*B = V|\Sigma|^2V^*$$

hence

$$\|A\|_2 = \max_k |\Sigma_k|^2 = \|B\|_2^2$$

On the other hand

$$B^{-1} = V\Sigma^{-1}U^* \Rightarrow \|B^{-1}\|_2 = \max_k \frac{1}{|\Sigma_k|}$$

But

$$A^{-1} = (B^*B)^-1 = B^{-1}(B^{-1})^* = V|\Sigma|^{-2}V^* \Rightarrow \|A\|_2 = \max_k |\Sigma_k|^{-2} = \|B^{-1}\|_2^2$$

Hence

$$cond(A) = \|A^{-1}\|_2\|A\|_2 = \|B^{-1}\|_2^2\|B^{-1}\|_2^2 = cond(B)^2$$

Problem 5. *Suppose LU decomposition without the choice of the leading element is calculated for a row-dominant matrix A. Prove that $\rho = \frac{\max_{i,j}|u_{i,j}|}{\max|a_{i,j}|} \leqslant 2$*

Proof. We will prove that after a forward Gauss step the matrix remains row-dominant.

$$\begin{aligned} a'_{kj} = a_{kj} - \frac{a_{k1}}{a_{11}}a_{1j} \implies \\ \sum_{j\geqslant 2, j\neq k} |a'_{kj}| = \sum_{j\geqslant 2, j\neq k} |a_{kj} - \frac{a_{k1}}{a_{11}}a_{1j}| \leqslant \sum_{j\geqslant 2, j\neq k} |a_{kj}| + |\frac{a_{k1}}{a_{11}}a_{1j}| \end{aligned} \tag{1}$$

Since A is row-dominant

$$\begin{aligned}&\sum_{j\geqslant 2, j\neq k} |a_{kj}| + |\frac{a_{k1}}{a_{11}}a_{1j}| < |a_{kk}| - |a_{k1}| + |\frac{a_{k1}}{a_{11}}|(|a_{11}| - |a_{1k}|) = \\ &|a_{kk}| - |\frac{a_{k1}}{a_{11}}a_{1k}| \leqslant |a_{kk} - \frac{a_{k1}}{a_{11}}a_{1k}| = |a'_{kk}|\end{aligned} \tag{2}$$

Let D be the diagonal of U, and let $DU' = U$. Then all entries of U' are not greater than 1, because it has 1s on the diagonal, and it is also row-dominant. Since the formula for inversion of triangular matrices with 1s on the diagonal is $(I+T)^{-1} = I - T$, $|[(u')^{-1}]_{ij}| \leqslant 1$

$$\begin{aligned}\max|u_{ij}| = \max|d_{ii}| = \max|\sum_j a_{ij}[(u')^{-1}]_{ji}| \leqslant \\ \leqslant \max\sum_j |a_{ij}| < 2\max|a_{ii}| = 2\max|a_{ij}|\end{aligned} \tag{3}$$

□

Problem 6. *Find the number of steps for Jacobi method to converge with precision 10^{-6} for a three-diagonal matrix, whose diagonal entries are 4, and the other do not exceed 1 in absolute value. The initial error $\|e\|_\infty < 10$*

Solution.

$$\begin{aligned}&\|D^{-1}(L+U)\|_\infty \leqslant 2 * \frac{1}{4} = \frac{1}{2} \implies \\ &\frac{10}{2^n} < 10^{-6} \implies n = \lceil 7\log(10)\rceil = 24\end{aligned} \tag{4}$$

□

Answer: 24 steps

Problem 7. *Prove that for second order linear systems Jacobi and Gauss-Seidel methods converge simultaneously.*

Proof. The criterion for convergence of Seidel's method is that the roots of

$$\begin{vmatrix} \lambda a & b \\ \lambda c & \lambda d \end{vmatrix} = 0$$

are less than 1. We can easily solve this equation, and express the criterion directly

$$ad\lambda^2 - bc\lambda = \lambda(ad\lambda - bc)$$

so the criterion is

$$\left|\frac{bc}{ad}\right| < 1$$

Similarly, the criterion for Jacobi's method

$$\begin{vmatrix} \lambda a & b \\ c & \lambda d \end{vmatrix} = ad\lambda^2 - bc = 0$$

can be reformulated as

$$\sqrt{\left|\frac{bc}{ad}\right|} < 1$$

Squaring this inequality, we get the equivalence of the two criterions. □

Problem 8. *Find a third order linear system for which the Gauss-Seidel method converges, but the Jacobi method diverges.*

Solution. Consider the matrix

$$A = \begin{pmatrix} 1 & 1 & 0 \\ a & 1 & 1 \\ -b & 0 & 1 \end{pmatrix}$$

It is easily calculated, that the criterion for Jacobi to converge is

$$t^3 - at - b = 0 \implies |t| < 1$$

and for Seidel

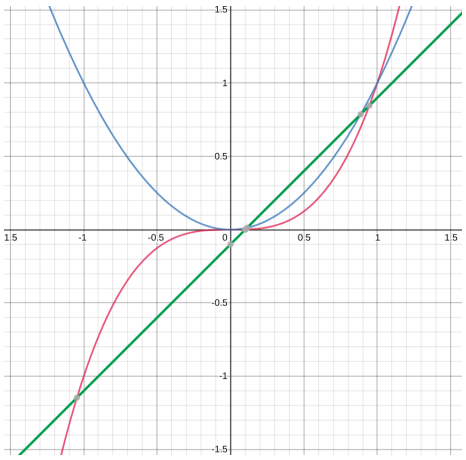
$$t^2 - at - b = 0 \implies |t| < 1$$

To provide a counterexample, we would like to choose such a, b that the first equation has roots outside $[-1, 1]$, and the second doesn't.

If we rewrite the equations as

$$t^3 = at + b, t^2 = at + b$$

The solution is given below graphically (clearly both roots for $t^2 = at + b$ lie within $[-1, 1]$, whereas for $t^3 = at + b$ there is one root below -1).



□

Problem 9. *Prove that the error of approximation of* e^x *on* $[0,1]$ *by a third degree polynomial with Chebyshev nodes does not exceed* 10^{-3}

Proof. The upper bound for the error is fiven by

$$\sup_{x\in[0,1]}\left|\frac{e^{\xi(x)}}{4!}\omega(x)\right| \leqslant \frac{e}{4!}\sup_{x\in[0,1]}|\omega(x)|$$

Consider a linear function which maps $[-1,1]$ to $[0,1]$ (because a linear map will preserve the degree of the polynomial). Such a map is given by

$$x \mapsto \frac{x+1}{2}$$

If we choose the transformed Chebyshev polynomial, that is the polynomial with transformed roots, we obtain the upper bound

$$\begin{aligned}\frac{e}{4!}\sup_{x\in[0,1]}\left|\tilde{T}_4(x)\right| = \frac{e}{4!}\sup_{x\in[-1,1]}\left|\tilde{T}_4\left(\frac{x+1}{2}\right)\right| = \\ = \frac{e}{4!}\sup_{x\in[-1,1]}\left|\prod_{i=1}^{4}\left(\frac{x+1}{2}-\frac{\xi_i+1}{2}\right)\right| = \frac{e}{2^4 4!}\sup_{x\in[-1,1]}|T_4(x)| = \\ = \frac{e}{2^7 4!} < \frac{3}{2^7\times 24} = \frac{1}{2^{10}} < \frac{1}{1000}\end{aligned}$$

Here, ξ_i are the roots of the Chebyshev polynomial of 4th degree. The Chebyshev polynomial is normalized so that the leading coefficient is 1.

□

Problem 10. *Evaluate the error of approximation of* e^x *at points* $0.05, 0.15$. *Interpolation method with nodes at* $x_0 = 0, x_1 = 0.1, x_2 = 0.2$ *is used.*

Solution. The upper bound is given by

$$\frac{e^{0.2}}{3!}|x(x-0.1)(x-0.2)|$$

Substituting $x = 0.05, 0.15$, we get $\omega(0.05) = \omega(0.15) = 375\times 10^{-6}$
Answer: the error for both points doesn't exceed

$$\frac{e^{0.2}\times 375\times 10^{-6}}{6} < 0.77\times 10^{-4}$$

□

x	0	$\pi/6$	$\pi/4$	$\pi/3$
$\sin(x)$	0	0.5	0.71	0.87

Problem 11. *With what accuracy can* $\sin(\pi/5)$ *be determined by polynomial interpolation from the table below? The absolute error at the nodes does not exceed* 10^{-2}.

Solution. The interpolating polynomial error is given by

$$\delta L(x) = \sum \delta f_j |l_j(x)| \leqslant 10^{-2} \sum |l_j(x)|$$

The sum is easily calculated by a program.

$$\delta L(\pi/5) \leqslant 1.2 \times 10^{-2}$$

As for the method error, the upper bound is

$$\sup_{x\in[0,\pi/3]} \left| \frac{\sin(x)}{4!} \omega(\pi/5) \right| = \frac{\sqrt{3}\pi^4}{2 \times 4! \times 22500} < 1.6 \times 10^{-4}$$

Clearly this error is negligible compared to the previous one.
Answer: the absolute error doesn't exceed 1.2×10^{-2} □

Problem 12. *Prove that the Lebesgue constant depends not on the size of the interval of interpolation, but on the relative distance between the points*

Proof. In other words we need to prove that the Lebesgue constant is invariant under affine transforms. We will first prove that the Lebesgue function is invariant under affine transforms. Indeed, let $\alpha(x) = ax + b$ be an affine map. Then all functions of Legendre basis are preserved (where we consider a new new Legendre basis, $\tilde{l}_j(x)$, for the set of transformed interpolation nodes)

$$\tilde{l}_j(\alpha(x)) = \prod_{i\neq j} \frac{\alpha(x) - \alpha(x_i)}{\alpha(x_j) - \alpha(x_i)} = \prod_{i\neq j} \frac{a(x - x_i)}{a(x_j - x_i)} = \prod_{i\neq j} \frac{x - x_i}{x_j - x_i} = l_j(x)$$

Hence

$$\tilde{\Lambda}(\alpha(x)) = \sum_j \left| \tilde{l}_j(\alpha(x)) \right| = \sum_j |l_j(x)| = \Lambda(x)$$

Since the interval of interpolation is also transformed by the map, we get

$$\tilde{\Lambda} = \max_{[\alpha(t_1),\alpha(t_2)]} \tilde{\Lambda}(x) = \max_{[t_1,t_2]} \tilde{\Lambda}(\alpha(x)) = \max_{[t_1,t_2]} \Lambda(x) = \Lambda$$

□

Problem 13. *Find the best approximation of* $\sin(x)$ *in* $L_2[0,1]$ *norm among polynomials of at most 2nd degree*

Solution. The Gram's matrix for basis $1, x, x^2$ is

$$\begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

Using a recursive formula,

$$\int_0^1 x^n \sin(x)dx = n\sin(1) - \cos(1) - n(n-1)\int_0^1 x^{n-2}\sin(x)dx$$

which is obtain by induction, it is easy to calculate the scalar products with $\sin(x)$

$$\langle \sin(x), 1\rangle = 1 - \cos(1) \tag{5}$$
$$\langle \sin(x), x\rangle = \sin(1) - \cos(1) \tag{6}$$
$$\left\langle \sin(x), x^2\right\rangle = 2\sin(1) + \cos(1) - 2 \tag{7}$$

We therefore need to solve a linear system

$$\begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 - \cos(1) \\ \sin(1) - \cos(1) \\ 2\sin(1) + \cos(1) - 2 \end{pmatrix}$$

It is easily solved by wolfram. The orthogonal projection (which is the best approximation) is

$$(-51 + 24\sin(1) + 57\cos(1)) - 12(-27 + 14\sin(1) + 28\cos(1))x + 30(-11 + 6\sin(1) + 11\cos(1))x^2$$

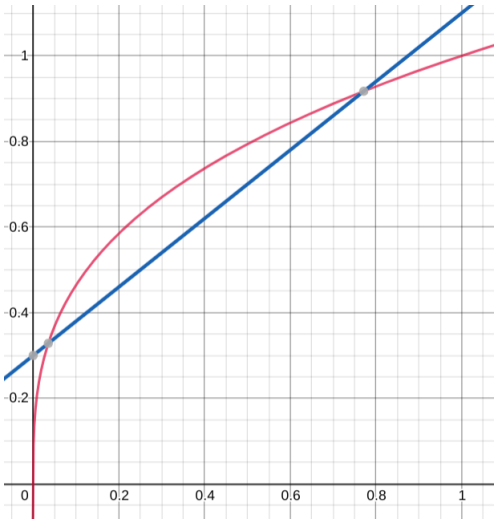
□

Problem 14. *Find the best approximation of* $x^{1/3}$ *in* $C[0,1]$ *norm among linear functions.*

Solution. Let $ax + b$ be the best approximation, and $E(x) = x^{1/3} - ax - b$ - the error. By Chebyshev's theorem, there must be $n + 2 = 3$ alternance points and therefore 2 roots of the error. From geometrical observations, the graphs of the functions $x^{1/3}, ax + b$ will be of the form

There are thus two alternance points at 0 and 1, so we have

$$E(0) = E(1) \implies b = a + b - 1 \implies a = 1 \tag{8}$$



The third alternance point is an extremum, so we find it by differentiation

$$E'(\tau) = 0 \implies \frac{1}{3}\tau^{-2/3} = a \implies \tau = (3a)^{-3/2}$$

Since $|E(\tau)| = |E(0)| = b$

$$(3a)^{-1/2} - a(3a)^{-3/2} - b = b \implies \frac{2}{3\sqrt{3a}} = 2b \implies b = \frac{1}{3\sqrt{3a}} = \frac{1}{3\sqrt{3}}$$

Hence, the best approximation is

$$x + \frac{1}{3\sqrt{3}}$$

□

Problem 15. *Calculate the order of approximation and optimal step size for numerical differentiation* $f'(x_2) \approx \frac{f_0 - 6f_1 + 3f_2 + 2f_3}{6h}$

The sum of the coefficients is zero, and expanding the first order coefficients yields $\frac{-2+6+2}{6} = 1$ as required. Expanding the second order, we get $2^2 - 6 + 2 = 0$, and the third order $(-2)^3 + 6 + 2 = 0$, so we have

$$f'(x_2) - \frac{f_0 - 6f_1 + 3f_2 + 2f_3}{6h} = \frac{\left((-2)^4 - 6 + 2\right) h^4 + o\left(h^4\right)}{6h} = O\left(h^3\right)$$

So the approximation is of the 3rd order. Approximating the error only with the lowest order terms, we obtain

$$E = \left(2^4 + 6 + 2\right) M_4 \frac{h^3}{6 \times 4!} + (1 + 6 + 3 + 2) M_0 \frac{\varepsilon}{6h} = \frac{1}{6} M_4 h^3 + 2M_0 \frac{\varepsilon}{h}$$

Differentiating w.r.t. h, we get

$$\frac{1}{2}M_4h^2 = 2M_0\frac{\varepsilon}{h^2} \implies h = \left(\frac{4M_0\varepsilon}{M_4}\right)^{\frac{1}{4}} \sim 2\times 10^{-4}\left(\frac{M_0}{M_4}\right)^{\frac{1}{4}}$$

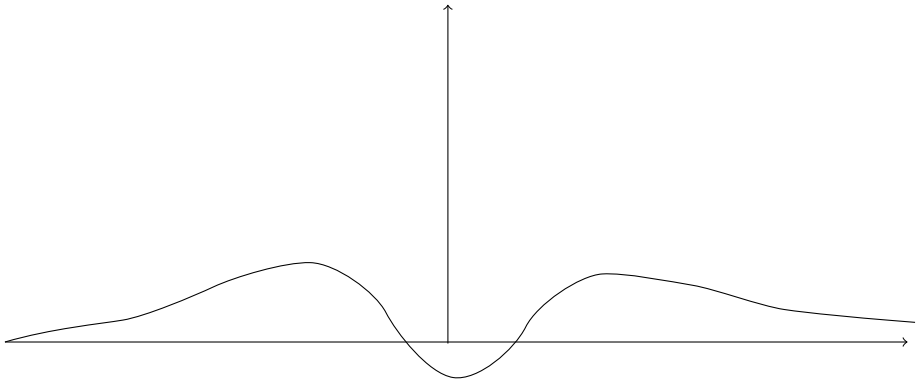
Thus h is in fact small, which justifies that we ignored higher terms. Substituting into expression for E, we get

$$E_{opt} = \frac{5M_0\varepsilon}{3}\sqrt[4]{\frac{M_4}{4M_0\varepsilon}} \sim 8\times 10^{-13}M_0\sqrt[4]{\frac{M_4}{M_0}}$$

Problem 16. *Evaluate the minimum number N of subintervals required to achieve $\varepsilon = 10^{-4}$ accuracy of a) $\int_0^1 e^{-x^2}dx$ and b) $\int_0^1 \sin x^2 dx$ using the trapezoidal rule*

We will use the formula $E \leqslant \frac{1}{12}M_2h^3$ for the trapezoidal rule.

a) $\frac{d^2}{dx^2}e^{-x^2} = 4x^2e^{-x^2} - 2e^{-x^2}$, differentiating again yields $-8x^3 + 12x = 0$. It is readily seen, that the rough sketch is



Comparing the extremum points, we find that $M_2 = 2$ is achieved at zero. Let $c_1 = 0, c_2 \dots c_n, c_n + 1 = 1$ be the subinterval boundaries. The upper bound for the error is therefore

$$E < 2\sum_{i=1}^{n}\frac{(c_{i+1} - c_i)^3}{12}$$

To minimize the sum, the points must be equidistributed. This is true for any such polynomial sum, because minimizing the contribution from two adjacent subintervals, we have

$$\frac{d}{dm}\left((m-a)^n + (b-m)^n\right) = n\left((m-a)^{n-1} - (b-m)^{n-1}\right) = 0 \implies m-a = b-m \implies m = \frac{a+b}{2}$$

Hence

$$E < \frac{2N}{12N^3} = \frac{1}{6N^2} = \varepsilon \implies N = \sqrt{\frac{1}{6\varepsilon}} \leqslant 41$$

b) We repeat the same procedure. The second derivative is $2\cos x^2 - 4x^2 \sin x^2$ and the third $-4x\left(2x^2 \cos x^2 + 3\sin x^2\right)$. This time there is only one extremum on $[0,1]$ (at 0 - it is 2). At 1, the second derivative is $2\cos 1 - 4\sin 1 \approx 2.3 > 2$. So $M_2 = 2.3$. Using the same technique as above we obtain $N \approx \sqrt{\frac{2.3}{12\times\varepsilon}} \approx 44$.

Problem 17. *Suggest how* $\int_0^1 \cos\frac{\pi}{x}dx$ *can be calculated with accuracy* $\varepsilon = 5 \times 10^{-5}$

Split the integral into two parts

$$\int_0^1 \cos\frac{\pi}{x}dx = \int_0^\delta \cos\frac{\pi}{x}dx + \int_\delta^1 \cos\frac{\pi}{x}dx$$

For the first integral, we have

$$|\int_0^\delta \cos\frac{\pi}{x}dx| \leqslant \delta$$

Estimate the upper bound for the second derivative

$$|\cos''\left(\frac{\pi}{x}\right)| = \frac{\pi^2}{x^4}\sin\frac{\pi}{x} + \frac{2\pi}{x^3}\cos\frac{\pi}{x} \leqslant \frac{\pi^2}{\delta^4} + \frac{2\pi}{\delta^3} < \frac{10}{\delta^4}$$

where the last inequality holds, because δ will be chosen small enough (it will be chosen smaller than ε). So if we are to calculate the integral from δ using the Newton-Cotes method with N subintervals, the totall error will not exceed

$$E \leqslant \delta + \frac{NM_2}{12}\left(\frac{1-\delta}{N}\right)^3 < \delta + \frac{10}{12N^2\delta^4} < \frac{1}{N^2\delta^4}$$

Hence we can choose N so that

$$\frac{1}{N^2\delta^4} \leqslant \varepsilon \implies N = \frac{1}{\delta^2\sqrt{\varepsilon}}$$

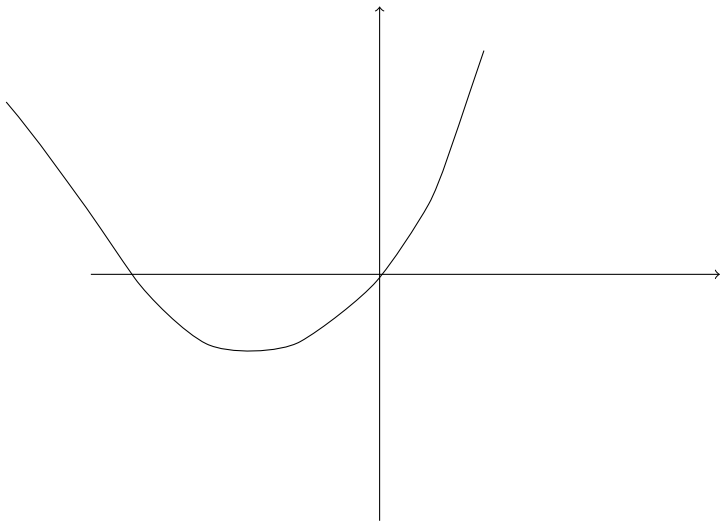
For example, we can choose $\delta = \frac{\varepsilon}{10} \implies N \sim 3 \times 10^8$. Actually the inequality $\delta + \frac{10}{12N^2\delta^4} < \frac{1}{N^2\delta^4}$ will not hold if we choose this approximate value of N, but it will if we set $N = \frac{1}{\delta^2}$ exactly, and clearly if we increase N, the error will not increase (this inequality is redundant - it was only used to estimate the values of N and δ, if we actually substitute the values of δ and N, the inequality $E < \varepsilon$ will hold).

Problem 18. *Suggest a simple iteration method to find the root of* $x = e^{2x} - 1$

Let $F(x) = e^{2x} - 1 - x$. We have $F'(x) = 0 \iff 2e^{2x} = 1$, in particular there is only one extremum, somewhere below 0, and it is a minimum. Since

$$F(-1) = e^{-2} > 0, F\left(-\frac{1}{2}\right) = e^{-1} - \frac{1}{2} < 0, F(0) = 0$$

the root we are looking for is located in $\left(-1, -\frac{1}{2}\right)$.



Let $f(x) = e^{2x} - 1$, notice that f is strictly increasing, hence

$$x \in \left(-1, -\frac{1}{2}\right) \implies -1 < e^{-2} - 1 \leqslant f(x) \leqslant e^{-1} - 1 < \frac{1}{2} - 1 = -\frac{1}{2}$$

Also

$$f'(x) = 2e^{2x} \leqslant 2e^{-2} < 1, x \in \left(-1, -\frac{1}{2}\right)$$

which means that f is a contraction map, so we can use the simple iteration method if we choose the initial approximation in $\left(-1, -\frac{1}{2}\right)$.

Problem 19. *Write the Newton's method formula for finding the root of a nonlinear equation. Estimate the number of iterations for accuracy* $\varepsilon = 10^{-5}$

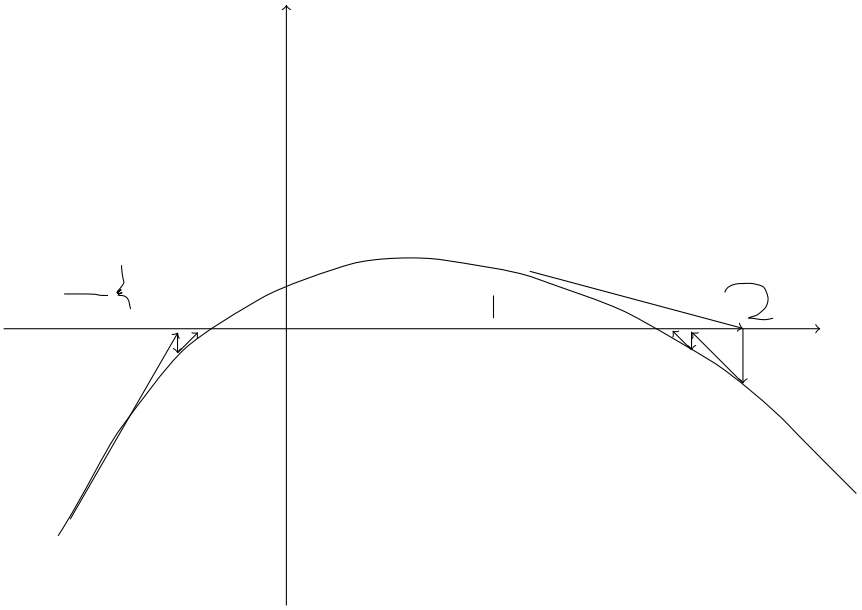
a) $F(x) = ln(x+2) - x^2 = 0$ First we will make a rough sketch of the graph. First find the extrema

$$F'(x) = \frac{1}{x+2} - 2x = 0 \iff x^2 + 2x = \frac{1}{2} \iff x = \pm\sqrt{\frac{3}{2}} - 1$$

Clearly the extremum with the minus sign is below -2, so it should be discarded. Also, we have

$$F(-1) = -1 < 0, F(0) = ln(2) > 0, F(1) = ln(3) - 1 > 0, F(2) = ln(4) - 4 < 0$$

so there are two roots - one in (-1,0) and the other in (1,2). Geometrically it is clear, that if we choose the initial approximation to the right of the maximum, the sequence will converge to the root on the right, and likewise for the left root.



Let's extimate the speed of convergence. We will only consider the root on the right, the calculations for the left one are similar. Since we have localized the root to be in (1,2), and from the sketch above it is clear, that if we choose 2 as initial approximation the sequence will stay there, it suffices find bounds of derivatives in the interval. The first derivative is negative and decreasing, so

$$|F'(x)| \geqslant |F'(1)| = 2 - \frac{1}{3} = \frac{5}{3}$$

For the second derivative we have

$$|F''(x)| = |-2 - \frac{1}{(x+2)^2}| \leqslant 2 + \frac{1}{9} = \frac{19}{9}$$

Hence

$$\gamma = \frac{||F''(\xi)||}{2||F'(\zeta)||} \leqslant \frac{19 \times 3}{18 \times 5} = \frac{57}{90}$$

Since the initial aproximation is closer to the root than $\frac{1}{\gamma} > 1$, we can use the formula

$$e_k \leqslant \gamma^{-1} (\gamma e_0)^{2^k} \leqslant \gamma^{2^k - 1}$$

Hence we will need

$$log_2 (1 + log_\gamma \varepsilon) \leqslant 5$$

iterations.

The formula for Newton's method in this case is

$$x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)} = x_k - \frac{ln(x_k + 2) - x_k^2}{\frac{1}{x_k + 2} - 2x_k} = x_k - \frac{(x_k + 2)\left(ln(x_k + 2) - x_k^2\right)}{1 - 4x_k - 2x_k^2}$$

b) $F(x) = e^x - 2x - 2 = 0$

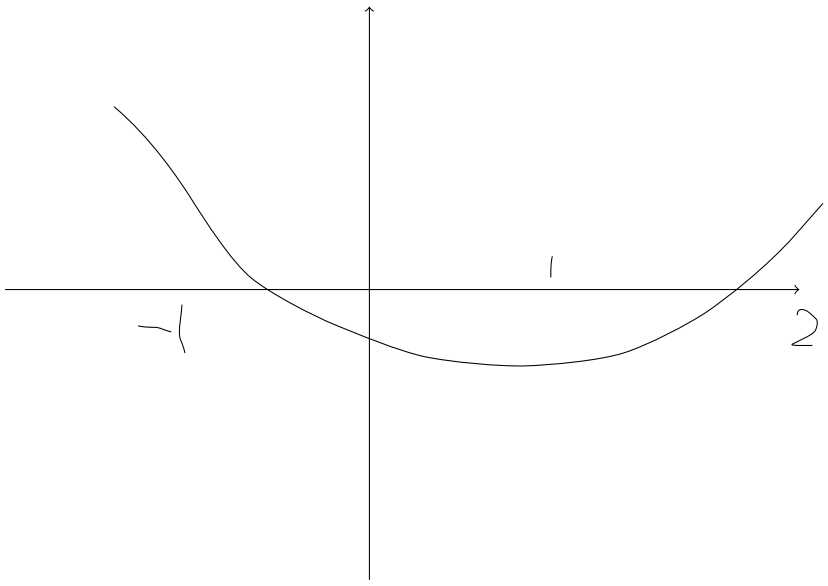
Similarly, we have

$$F'(x) = e^x - 2 = 0 \iff x = ln(2), F(2) = -2ln(2) < 0$$

Since

$$F(-1) = e^{-1} > 0, F(0) = -1 < 0, F(1) = e - 4 < 0, F(2) = e^2 - 6 > 0$$

we can localize the roots in (-1, 0) and in (1,2)



For the left root we have

$$|F'(x)| \geqslant |F'(0)| = 1, |F''(x)| = e^x \leqslant 1 \implies \gamma \leqslant \frac{1}{2}$$

If we take the initial approximation anywhere in (-1,0), the number of iterations needed will be

$$log_2\left(1+log_\gamma \varepsilon\right) \leqslant 5$$

For the right root, the localisation needs to be more accurate for the approximation to work. The iterative formula is given by

$$x_{k+1} = x_k - \frac{e^{x_k} - 2x_k - 2}{e^{x_k} - 2} = \frac{\left(x_k - 1\right) e^{x_k} + 2}{e^{x_k} - 2}$$

Problem 20. *The finite difference scheme*

$$U_{i+1} - 2U_i + U_{i-1} = \frac{h^2}{12}\left(f_{i+1} + 10f_i + f_{i-1}\right)$$

is used to solve an ODE

$$u''\left(x\right) = f\left(x\right), u\left(0\right) = a, u\left(1\right) = b$$

Find the approximation order of the scheme

We need to calculate the order of the residual vector. We will use the $||\cdot||_\infty$ norm. Expanding into Taylor series, we obtain

$$u\left(x_{i+1}\right) - 2u\left(x_i\right) + u\left(x_{i-1}\right) - \frac{h^2}{12}\left(f_{i+1} + 10f_i + f_{i-1}\right) =$$

$$= u^{(2)}\left(x_i\right)h^2 + u^{(4)}\left(x_i\right)\frac{h^4}{12} + u^{(6)}\left(x_i\right)\frac{2h^6}{6!} + o\left(h^6\right) - \frac{h^2}{12}\left(12f\left(x_i\right) + f^{(2)}\left(x_i\right)h^2 + f^{(4)}\left(x_i\right)\frac{h^4}{12} + o\left(h^4\right)\right) =$$

$$= \left(u^{(2)}\left(x_i\right) - f\left(x_i\right)\right)h^2 + \left(u^{(4)}\left(x_i\right) - f^{(2)}\left(x_i\right)\right)\frac{h^4}{12} + \left(\frac{2u^{(6)}\left(x_i\right)}{6!} - \frac{f^{(4)}\left(x_i\right)}{12^2}\right)h^6 + o\left(h^6\right) =$$

$$= O\left(h^6\right)$$

because the first two terms vanish, since

$$u''\left(x\right) = f\left(x\right) \implies u^{(4)}\left(x\right) = f''\left(x\right)$$

Hence the approximation is of the sixth order, which is two orders higher than if we just used $f\left(x\right)$ on the right hand side.