S.-T. Yau College Student Mathematics Contests 2022

Computational and Applied Mathematics

Solve every problem.

Problem 1. Consider $\{p_i(x)\}_{i=0}^{\infty}$, a family of orthogonal polynomials associated with the inner product

$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)w(x)\,dx, \quad w(x) > 0 \quad \text{for } x \in (-1, 1),$$

where $p_i(x)$ is a polynomial of degree i. Let $x_0, x_1, \dots, x_n$ be the roots of $p_{n+1}(x)$. Construct an orthonormal basis in the subspace of the polynomials of degree no more than n such that, for any polynomial in this subspace, the coefficients of its expansion into the basis are equal to the scaled values of this polynomial at the nodes $x_0, x_1, \dots, x_n$.

Solution: Start by considering $l_i(x)$, $i = 0, \dots, n$, the Lagrange interpolating polynomials of degree n for the nodes $x_0, x_1, \dots, x_n$. Let us compute the inner product of two such polynomials using the Gaussian quadrature, exact for the polynomials of degree less or equal to $2n + 1$. We have

$$\langle l_i, l_j\rangle = \int_{-1}^{1} l_i(x)l_j(x)w(x)\,dx = \sum_{k=0}^{n} l_i(x_k)l_j(x_k)w_k = \delta_{ij}w_i,$$

where $w_0, w_1, \dots, w_n$ are the positive weights of the quadrature.

We now normalize the Lagrange interpolating polynomials,

$$R_i(x) = \frac{1}{\sqrt{w_i}} l_i(x),$$

and obtain

$$\int_{-1}^{1} R_i(x)R_j(x)w(x)\,dx = \sum_{k=0}^{n} w_k R_i(x_k)R_j(x_k) = \sum_{k=0}^{n} w_k \frac{1}{\sqrt{w_i}}\delta_{ik}\frac{1}{\sqrt{w_j}}\delta_{jk} = \delta_{ij},$$

namely, these functions form an orthonormal basis. The coefficients of a function in this subspace are computed as projections on the basis,

$$f_i = \langle f, R_i\rangle = \int_{-1}^{1} f(x)R_i(x)w(x)\,dx = \sum_{k=0}^{n} w_k f(x_k)R_i(x_k) = \sum_{k=0}^{n} w_k f(x_k)\frac{1}{\sqrt{w_i}}\delta_{ik} = \sqrt{w_i} f(x_i).$$

Problem 2. Consider a 2D fixed point iteration of the form

$$x_{k+1} = f(x_k, y_k),\ y_{k+1} = g(x_k, y_k). \tag{1}$$

Assume that the vector-valued function $\vec{H}(x, y) = (f(x, y), g(x, y))^T$ is continuously-differentiable, and the infinity norm of the Jacobian matrix is less than 1 at a unique fixed point (x_∞, y_∞).

Now consider a new iteration:

$$x_{k+1} = f(x_k, y_k), \quad y_{k+1} = g(x_{k+1}, y_k). \tag{2}$$

Prove that iteration (2) is convergent, to the same fixed point as iteration (1), for the initial conditions sufficiently close to the fixed point.

Solution: First, we check that the new iteration has the same fixed point as the original iteration. For (1), $x_\infty = f(x_\infty, y_\infty)$, $y_\infty = g(x_\infty, y_\infty)$. Thus, we have for the new iteration,

$$\begin{aligned} f(x_\infty, y_\infty) &= x_\infty, \\ g(f(x_\infty, y_\infty), y_\infty) = g(x_\infty, y_\infty) &= y_\infty. \end{aligned}$$

Next, the Jacobian of the new iteration reads as

$$\mathbf{J}_2 = \begin{bmatrix} \partial_1 f(x,y) & \partial_2 f(x,y) \\ \partial_1 g(f(x,y),y)\partial_1 f(x,y) & \partial_1 g(f(x,y),y)\partial_2 f(x,y) + \partial_2 g(f(x,y),y) \end{bmatrix}. \tag{3}$$

The infinity norm of the Jacobian is the maximum absolute row sum. The first row has exactly the same absolute row sum as the Jacobian of the original iteration, thus we have

$$\left|\partial_1 f(x_\infty, y_\infty)\right| + \left|\partial_2 f(x_\infty, y_\infty)\right| < 1.$$

The absolute row sum for the second row is

$$\begin{aligned} &\left|\partial_1 g(f(x,y),y)\partial_1 f(x,y)\right| + \left|\partial_1 g(f(x,y),y)\partial_2 f(x,y) + \partial_2 g(f(x,y),y)\right| \\ &\leq \left|\partial_1 g(f(x,y),y)\right| \left(\left|\partial_1 f(x,y)\right| + \left|\partial_2 f(x,y)\right|\right) + \left|\partial_2 g(f(x,y),y)\right|. \end{aligned}$$

Evaluating at (x_∞, y_∞), we have

$$\begin{aligned} &\left|\partial_1 g(f(x_\infty, y_\infty), y_\infty)\right| \left(\left|\partial_1 f(x_\infty, y_\infty)\right| + \left|\partial_2 f(x_\infty, y_\infty)\right|\right) + \left|\partial_2 g(f(x_\infty, y_\infty), y_\infty)\right| \\ &\leq \left|\partial_1 g(f(x_\infty, y_\infty), y_\infty)\right| + \left|\partial_2 g(f(x_\infty, y_\infty), y_\infty)\right| \\ &= \left|\partial_1 g(x_\infty, y_\infty)\right| + \left|\partial_2 g(x_\infty, y_\infty)\right| \qquad < 1. \end{aligned}$$

Thus, the Jacobian of the new iteration has infinity norm less than 1 at the fixed point. Since the new iteration is continuously-differentiable, there must be a neighborhood of the fixed point such that iterations initialized in this neighborhood will converge.

Problem 3. Let $A \in \mathbf{R^{m \times m}}$ be a matrix with entries a_{ij} which satisfy

$$a_{ii} \geq \sum_{j \neq i} |a_{ij}| + 2, \quad a_{ii} \leq 7.$$

(a) Prove that A^{-1} exists.

(b) Prove that $\|A\|_\infty$ is the *max row sum* (of absolute values) of A.

(c) Find both a lower and upper bound for $\|A\|_\infty$.

(d) Now assume $A = A^T$. Find bounds for $\|A\|_2$ and $\|A^{-1}\|_2$.

Solution:

(a) Suppose A is singular and $v = [v_1\ v_2\ \cdots\ v_m]^T$ be an eigenvector corresponding to eigenvalue 0 with $\max_j |v_j| = |v_k|$. Consider the k-th row of the vector equation $Av = 0$,

$$a_{kk} v_k = -\sum_{j \neq k} a_{kj} v_j.$$

Therefore,

$$|a_{kk}| \leq \sum_{j \neq k} |a_{kj}| \left|\frac{v_j}{v_k}\right| \leq \sum_{j \neq k} |a_{kj}|,$$

which is a contradiction.

(b) Suppose the k-th row has the max absolute sum, *i.e.*,

$$\max_i \left\{ \sum_j |a_{ij}| \right\} = \sum_j |a_{kj}|.$$

For the vector v with $v_j = \text{sgn}(a_{kj})$, we have

$$\|Av\|_\infty \geq \sum_j |a_{kj}|,$$

since the right hand side is the k-th element of Av. Noting that $\|v\|_\infty = 1$, we have $\|A\|_\infty \geq \sum_j |a_{kj}|$. For the other inequality, we have that for any vector u,

$$\|Au\|_\infty = \max_i \left\{ \sum_j |a_{ij}| \|u_j\| \right\} \leq \max_i \left\{ \sum_j |a_{ij}| \right\} \|u\|_\infty.$$

(c) From given information,

$$2 \leq \sum_j |a_{ij}| = a_{ii} + \sum_{j \neq i} |a_{ij}| \leq a_{ii} + (a_{ii} - 2) \leq 12.$$

Therefore, $2 \leq \|A\|_\infty \leq 12$.

(d) $A = A^T$ means singular values are the absolute values of (real) eigenvalues. By the Gershgorin Theorem,

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|.$$

Therefore,

$$a_{ii} - \sum_{j \neq i} |a_{ij}| \leq \lambda \leq a_{ii} + \sum_{j \neq i} |a_{ij}|.$$

By given info.

$$2 \leq \lambda \leq 12.$$

Now since $A = A^T$ and A is invertible, the smallest and largest singular values of A^{-1} will be the reciprocal of the largest and smallest singular values of A respectively. Hence,

$$\tfrac{1}{12} \leq \|A^{-1}\|_2 \leq \tfrac{1}{2}.$$

Problem 4. Consider a system of ODE initial value problems of the form:

$$\frac{d}{dt} u = f(u), \quad u(0) = u_0.$$

Assume that $f(u)$ has the property that the forward Euler (FE) method:

$$U^{n+1} = U^n + k f(U^n),$$

satisfies

$$\|U^{n+1}\| \leq \|U^n\|$$

for some norm $\|\cdot\|$ and for all time-steps k, $0 < k \leq k_{FE}$. Now consider the 2-stage Runge-Kutta method:

$$\begin{aligned} U^{(1)} &= U^n + k\beta_{10} f(U^n), \\ U^{n+1} &= \{\alpha_{20} U^n + k\beta_{20} f(U^n)\} + \{\alpha_{21} U^{(1)} + k\beta_{21} f(U^{(1)})\} \end{aligned}$$

where

$$\beta_{10} \geq 0, \quad \beta_{20} \geq 0, \quad \beta_{21} \geq 0, \quad \alpha_{20} \geq 0, \quad \alpha_{21} \geq 0, \quad \alpha_{20} + \alpha_{21} = 1.$$

(a) Prove that the above 2-stage Runge-Kutta method also satisfies the inequality:

$$\|U^{n+1}\| \leq \|U^n\|$$

under some appropriate time-step restriction: $0 \leq k \leq k^*$, where you need to explicitly determine k^* in terms of k_{FE}.

(b) Explicitly determine the coefficients:

$$\beta_{10}, \quad \beta_{20}, \quad \beta_{21}, \quad \alpha_{20}, \quad \alpha_{21},$$

so that

(i) The method is second-order accurate; and

(ii) The maximum allowed time-step, k^*, is as large as possible.

Solution:

(a) The first stage is simply the forward Euler method with a time step $k\beta_{10}$. Therefore, as long as

$$k \cdot \max\{1, \beta_{10}\} \leq k_{FE},$$

we have that

$$\|U^{(1)}\| \leq \|U^n\|.$$

The second stage can be written as a linear combination of two forward Euler steps:

$$U^{n+1} = \alpha_{20}\left\{U^n + k\frac{\beta_{20}}{\alpha_{20}} f(U^n)\right\} + \alpha_{21}\left\{U^{(1)} + k\frac{\beta_{21}}{\alpha_{21}} f(U^{(1)})\right\}.$$

Requiring that

$$k \cdot \max\left\{1, \beta_{10}, \frac{\beta_{20}}{\alpha_{20}}, \frac{\beta_{21}}{\alpha_{21}}\right\} \leq k_{FE},$$

we get that

$$\|U^{n+1}\| \leq \alpha_{20}\|U^n\| + \alpha_{21}\|U^{(1)}\| \leq (\alpha_{20} + \alpha_{21})\|U^n\| = \|U^n\|.$$

Therefore, the 2-stage RK method satisfies $\|U^{(1)}\| \leq \|U^n\|$ under the following time-step constraint:

$$k \leq k^* = k_{FE} \cdot \min\left\{1, \frac{1}{\beta_{10}}, \frac{\alpha_{20}}{\beta_{20}}, \frac{\alpha_{21}}{\beta_{21}}\right\}.$$

(b) To see the local truncation error, apply the method to the function $f(u) = \lambda u$ (and let $z = k\lambda$):

$$U^{(1)} = (1 + z\beta_{10})\, U^n,$$
$$U^{n+1} = \alpha_{20}\left(1 + z\frac{\beta_{20}}{\alpha_{20}}\right) U^n + \alpha_{21}\left(1 + z\frac{\beta_{21}}{\alpha_{21}}\right) U^{(1)}.$$

Combining these two results (and using the fact that $\alpha_{20} + \alpha_{21} = 1$):

$$U^{n+1} = \left(1 + z\,(\beta_{20} + \beta_{21} + \alpha_{21}\beta_{10}) + \frac{z^2}{2}\,(2\beta_{10}\beta_{21})\right)U^n.$$

Therefore, for accuracy considerations we require that

$$\beta_{20} + \beta_{21} + \alpha_{21}\beta_{10} = 1 \quad \text{and} \quad \beta_{10}\beta_{21} = \tfrac{1}{2}.$$

For optimal stability we require that

$$0 \leq \beta_{10} \leq 1, \quad 0 \leq \beta_{20} \leq \alpha_{20} \leq 1, \quad 0 \leq \beta_{21} \leq \alpha_{21} \leq 1, \quad \alpha_{20} + \alpha_{21} = 1.$$

The optimal solution requires that

$$\beta_{10} = 1 \implies \beta_{21} = \tfrac{1}{2} \implies \beta_{20} + \alpha_{21} = \tfrac{1}{2} \implies \alpha_{21} = \tfrac{1}{2} \implies \beta_{20} = 0 \quad \text{and} \quad \alpha_{20} = \tfrac{1}{2}.$$

Putting these all together yields the following 2-stage RK method that is norm-preserving under the optimal time-step restriction $k \le k^*$:

$$\begin{aligned} U^{(1)} &= U^n + kf(U^n), \\ U^{n+1} &= \tfrac{1}{2}\left\{U^n + U^{(1)} + kf(U^{(1)})\right\}. \end{aligned}$$

Problem 5. Construct a third-order accurate Lax-Wendroff-type method for $u_t + au_x = 0$ ($a > 0$ is a constant) in the following way:

(a)
- Expand $u(t+k, x)$ in a Taylor series and keep the first four terms. Replace all time derivatives by spatial derivatives using the equation.
- Construct a cubic polynomial passing through the points $U^n_{j-2}, U^n_{j-1}, U^n_j, U^n_{j+1}$.
- Approximate the spatial derivatives in the Taylor series by the exact derivatives of the above constructed cubic polynomial.

(b) Verify that the truncation error is $O(k^3)$ if $h = O(k)$.

Solution:

(a) By Taylor's expansion in time, and using the equation, we have

$$\begin{aligned} u(t+k, x) &= u(t, x) + ku_t(t, x) + \frac{k^2}{2}u_{tt}(t, x) + \frac{k^3}{6}u_{ttt}(t, x) + O(k^4) \\ &= u(t, x) - aku_x(t, x) + \frac{(ak)^2}{2}u_{xx}(t, x) - \frac{(ak)^3}{6}u_{xxx}(t, x) + O(k^4). \end{aligned}$$

Use Lagrange interpolation to construct a cubic polynomial passing through $x_0 - 2h, x_0 - h, x_0, x_0 + h$. Since the interpolation is in the spatial variable, we omit the time variable for this part. The polynomial $p_3(x)$ satisfying $u = p_3 + O(h^4)$ is,

$$\begin{aligned} p_3(x) = &-\frac{(x - x_0 + h)(x - x_0)(x - x_0 - h)}{6h^3}u(x_0 - 2h) \\ &+ \frac{(x - x_0 + 2h)(x - x_0)(x - x_0 - h)}{2h^3}u(x_0 - h) \\ &- \frac{(x - x_0 + 2h)(x - x_0 + h)(x - x_0 - h)}{2h^3}u(x_0) \\ &+ \frac{(x - x_0 + 2h)(x - x_0 + h)(x - x_0)}{6h^3}u(x_0 + h). \end{aligned}$$

Also, the error is

$$\begin{aligned} u'(x_0) &= p_3'(x_0) + O(h^3), \\ u''(x_0) &= p_3''(x_0) + O(h^2), \\ u'''(x_0) &= p_3'''(x_0) + O(h). \end{aligned}$$

Now we need to substitute the expressions of $p_3'(x_0), p_3''(x_0), p_3'''(x_0)$ in place of u_x, u_{xx}, u_{xxx} respectively. By simple calculation,

$$\begin{aligned} p_3'(x_0) &= \frac{u(x_0 - 2h)}{6h} - \frac{u(x_0 - h)}{h} + \frac{u(x_0)}{2h} + \frac{u(x_0 + h)}{3h}, \\ p_3''(x_0) &= \frac{u(x_0 - h)}{h^2} - 2\frac{u(x_0)}{h^2} + \frac{u(x_0 + h)}{h^2}, \\ p_3'''(x_0) &= -\frac{u(x_0 - 2h)}{h^3} + 3\frac{u(x_0 - h)}{h^3} - 3\frac{u(x_0)}{h^3} + \frac{u(x_0 + h)}{h^3}. \end{aligned}$$

Plugging these expressions into the Taylor expression, we obtain

$$\begin{aligned} U_j^{n+1} = U_j^n &- \frac{ak}{6h}\left(U_{j-2}^n - 6U_{j-1}^n + 3U_j^n + 2U_{j+1}^n\right) \\ &+ \frac{(ak)^2}{2h^2}\left(U_{j-1}^n - 2U_j^n + U_{j+1}^n\right) \\ &- \frac{(ak)^3}{6h^3}\left(-U_{j-2}^n + 3U_{j-1}^n - 3U_j^n + U_{j+1}^n\right). \end{aligned}$$

(b) From the Taylor expression and the approximation errors,

$$\begin{aligned} \frac{u(t+k,x)-u(t,x)}{k} &= -au_x + \frac{a^2k}{2}u_{xx} - \frac{a^3k^2}{6h}u_{xxx} + O(k^3) \\ &= -\frac{a}{6h}\Big(u(t,x-2h) - 6u(t,x-h) + 3u(t,x) + 2u(t,x+h)\Big) + O(h^3) \\ &\quad + \frac{a^2k}{2h^2}\Big(u(t,x-h) - 2u(t,x) + u(t,x+h)\Big) + O(kh^2) \\ &\quad - \frac{a^3k^2}{6h^3}\Big(-u(t,x-2h) + 3u(t,x-h) - 3u(t,x) + u(t,x+h)\Big) + O(k^2h) \\ &\quad + O(k^3). \end{aligned}$$

Hence, if $h = O(k)$, then scheme is $O(k^3)$.

Problem 6. Suppose you have \$60K to invest and there are 3 investment options available. You must invest in multiples of \$10K. If d_i dollars are invested in investment i then you receive a net value (as the profit) of $r_i(d_i)$ dollars. For $d_i > 0$ we have

$$\begin{aligned} r_1(d_1) &= (7d_1 + 2) \times 10, \\ r_2(d_2) &= (3d_2 + 7) \times 10, \\ r_3(d_3) &= (4d_3 + 5) \times 10, \end{aligned}$$

and $d_1(0) = d_2(0) = d_3(0)$. All are measured in \$10K dollars. The objective is to maximize the net value of your investments. This can be formulated as a linear programming problem:

$$\begin{aligned} &\max_{d_1,d_2,d_3} r_1(d_1) + r_2(d_2) + r_3(d_3), \\ &\text{such that} \quad d_1 + d_2 + d_3 \le 6, \\ &d_i \ge 0 \quad i = 1,2,3 \quad \text{are integers.} \end{aligned}$$

Solution: We solve this problem by dynamical programming. The key elements are

1. The stage i is just investment i.
2. The decisions at stage i are d_i which is how much to be invested in i. The immediate return is $r_i(d_i)$ and it is obvious that $d_i \in \{0, 1, \dots, 6\}$.
3. The states at stage i is x_i = the total amount available to be invested to investment i.

Define $f_i(x_i)$ = maximum net value for stages i given that we have x_i dollars available. Then the recursion equation is

$$f_i(x_i) = \max_{d_i}\left\{r_i(d_i) + f_{i+1}(x_i - d_i)\right\}.$$

The boundary condition is $f_4(x_4) = 0$. The answer is $f_1(6)$. This is a backward recursion and the computation goes as:

Stage 3: Note that

$$f_3(x_3) = \max_{0 \le d_3 \le 6} \{r_3(d_3)\}.$$

We have the following table:

Table 1: $r_3(d_3) = 4d_3 + 5$

x_3	$d_3 = 0$	$d_3 = 1$	$d_3 = 2$	$d_3 = 3$	$d_3 = 4$	$d_3 = 5$	$d_3 = 6$	$f_3(x_3)$	d_3^*
0	0	-	-	-	-	-	-	0	0
1	0	9	-	-	-	-	-	9	1
2	0	9	13	-	-	-	-	13	2
3	0	9	13	17	-	-	-	17	3
4	0	9	13	17	21	-	-	21	4
5	0	9	13	17	21	25	-	25	5
6	0	9	13	17	21	25	29	29	6

Stage 2: Note that

$$f_2(x_2) = \max_{0 \le d_2 \le 6} \{r_2(d_2) + f_3(x_2 - d_2)\}.$$

We have the following table

Table 2: $r_2(d_2) + f_3(x_2 - d_2) = 3d_2 + 7 + f_3(x_2 - d_2)$

x_2	$d_2 = 0$	$d_2 = 1$	$d_2 = 2$	$d_2 = 3$	$d_2 = 4$	$d_2 = 5$	$d_2 = 6$	$f_2(x_2)$	d_2^*
0	0	-	-	-	-	-	-	0	0
1	9	10	-	-	-	-	-	10	1
2	13	19	13	-	-	-	-	19	1
3	17	23	22	16	-	-	-	23	1
4	21	27	26	25	19	-	-	27	1
5	25	31	30	29	28	22	-	31	1
6	29	35	34	33	32	31	25	35	1

Stage 1: Note that

$$f_1(x_1) = \max_{0 \le d_1 \le 6} \{r_1(d_1) + f_2(6 - d_1)\}.$$

We have the following table

Thus the optimal net present value is \$49000 and the investment policy is $d_1 = 4, d_2 = 1, d_3 = 1$.

Table 3: $r_1(d_1) + f_2(x_1 - d_1) = 7d_1 + 2 + f_2(6 - d_1)$

x_1	$d_1 = 0$	$d_1 = 1$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$	$d_1 = 5$	$d_1 = 6$	$f_1(x_1)$	d_1^*
6	35	40	43	46	49	47	44	49	4