Math 567 **Homework 3** **Due Oct. 11, 2020**
Grady Wright

1. *Linear two-point boundary value problems: Finite differences*

   (a) Implement a general function in whatever language you like for numerically solving the general two-point boundary value problem (TPBVP)

   $$u'' = p(x)u' + q(x)u + r(x), \quad x \in [0,1], \quad u(0) = \alpha, \; u(1) = \beta$$

   using centered, second-order finite differences. Your routine should be called `fd2tpbvp`, and should be entirely general (i.e. it should take as input functions representing $p(x)$, $q(x)$, & $r(x)$, values for α & β, and the number of interior discretization points m). Send me an electronic copy of the code and include a listing of the code in your write-up.

   (b) Use your function from part (a) to numerically approximate the TPBVP

   $$u'' = 2\tan(x)u' + 2, \quad u(0) = u(1) = 0,$$

   which has the exact solution $u(x) = (x-1)\tan(x)$. Compare the approximate solution to the exact solution using $m = 10, 20, 40, 80, 160$ equally spaced interior points over $[0,1]$ (i.e. $x_j = jh$, $j = 1, \ldots, m$, with $h = 1/(m+1)$) and verify numerically that the approximate solution is second-order accurate.

2. *Fictitious point method for Robin boundary conditions.* Consider the TPBVP

   $$u'' = p(x)u' + q(x)u + r(x), \quad x \in [a,b],$$

   with mixed boundary conditions

   $$u(a) = \alpha \;\text{ and }\; \beta_1 u(b) + \beta_2 u'(b) = \beta_3.$$

   Suppose we discretize the equation with $m+1$ equally-spaced subintervals of spacing h. Using the fictitious point method described in in class, derive the following second-order accurate finite difference approximation for u_{m+1}:

   $$-2u_m + \left[2 + h^2 q_{m+1} + (2 - hp_{m+1})h\frac{\beta_1}{\beta_2}\right] u_{m+1} = -h^2 r_{m+1} + (2 - hp_{m+1})h\frac{\beta_3}{\beta_2},$$

   where $u(b) \approx u_{m+1}$, $u(b-h) \approx u_m$, $p_{m+1} = p(b)$, $q_{m+1} = q(b)$, and $r_{m+1} = r(b)$.

3. *Neumann-Neumman boundary conditions.* Consider the 1D Poisson equation with Neumann-Neumann boundary conditions

   $$u''(x) = f(x), \; a \le x \le b, \; u'(a) = \sigma_0, \; u'(b) = \sigma_1. \tag{1}$$

   As discussed in class this equation has an infinite number of solutions if $\int_a^b f(x)\,dx = \sigma_1 - \sigma_0$ (the so-called compatibility condition), otherwise there is no solution. If $\sigma_0 = \sigma_1 = 0$ (i.e. the zero-flux condition) then it can be shown that solutions to the Poisson equation are unique up to an additive constant. If we discretize this BVP at $m+2$ equally-spaced points across

$[a,b]$ and use second-order accurate FD formulas in the interior and the second-order accurate fictitious point method at the boundary, we arrive at the linear system

$$\underbrace{\frac{1}{h^2}\begin{bmatrix} -2 & 2 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & -2 & 1 & \ddots & & & \vdots \\ 0 & 1 & -2 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 1 & -2 & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & 2 & -2 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_m \\ u_{m+1} \end{bmatrix}}_{\mathbf{u}} = \underbrace{\begin{bmatrix} f_0 + \frac{2}{h}\sigma_0 \\ f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_m \\ f_{m+1} - \frac{2}{h}\sigma_1 \end{bmatrix}}_{\mathbf{b}}, \tag{2}$$

where $h = (b-a)/(m+1)$, $f_j = f(x_j)$, and $x_j = a + jh$, $j = 0, \ldots, m+1$.

As discussed in class (and by simple inspection), the matrix A in this equation is singular since every row sums to zero. This means that the vector of length $m+2$

$$\mathbf{e} = [1 \quad 1 \quad \ldots \quad 1]^T$$

is an eigenvector corresponding to a zero eigenvalue: $A\mathbf{e} = \mathbf{0}$. Furthermore, it is the only eigenvector (up to a non-zero multiplicative constant) with this property, i.e. it is the only vector in the null-space of A. Another property of A is that the vector of length $m+2$

$$\mathbf{w} = \left[\tfrac{1}{2} \quad 1 \quad \ldots \quad 1 \quad \tfrac{1}{2}\right]^T$$

is an eigenvector of A^T corresponding to a zero eigenvalue: $A^T\mathbf{w} = \mathbf{0}$ or $\mathbf{w}^T A = \mathbf{0}^T$. It is also the only eigenvector (up to a non-zero multiplicative constant) with this property, i.e. it is the only vector in the null-space of A^T.

A result from linear algebra (known as the *Fredholm Alternative*) is that the system $A\mathbf{u} = \mathbf{b}$ has an infinite number of solutions if $\mathbf{w}^T\mathbf{b} = 0$ (which guarantees $\mathbf{b}$ is in the column space of A). This condition can be viewed as a discrete version of the continuous compatibility condition $\int_a^b f(x)\,dx = \sigma_1 - \sigma_0$.

This is all beautiful theory, but in practice what we want to do is actually compute one of the solutions to the BVP. This homework problem and the next one discuss two approaches (with this problem being the most general of the two) for computing a solution with a "direct" method. It should be noted that effective "iterative" methods can also be used. The overall goal of any of these methods is to make the problem unique by adding on an additional constraint. In the continuous problem, especially for the zero-flux case, this is typically accomplished by specifying the "total amount" of u in the domain, which mathematically we write as

$$\frac{1}{b-a}\int_a^b u(x)\,dx = U, \tag{3}$$

For the zero-flux case, this fixes the arbitrary constant in the solution to be equal to U. This same approach is followed in the discrete problem as well, with the discrete version of this integral taking the form

$$\frac{h}{b-a}\left[\frac{1}{2}u_0 + \sum_{j=1}^{m} u_j + \frac{1}{2}u_{m+1}\right] = U \quad \Longleftrightarrow \quad \mathbf{w}^T\mathbf{u} = (m+1)U,$$

where we have used the fact that $h = (b-a)/(m+1)$. Note that this is just the Trapezoidal rule approximation to (3).

The discrete problem (2) can now be converted into the following equality-constrained quadratic programming problem:

$$\begin{aligned} \min \;\; & J(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T A\mathbf{u} - \mathbf{u}^T\mathbf{b} \\ \text{subject to} \;\; & \mathbf{w}^T\mathbf{u} = (m+1)U \end{aligned} \tag{4}$$

The method of *Lagrange multipliers* can be used to solve this problem. The idea is to form the *Lagrangian*

$$\mathcal{L}(\mathbf{u},\lambda) = J(\mathbf{u}) + \lambda\left(\mathbf{w}^T\mathbf{u} - (m+1)U\right) = \frac{1}{2}\mathbf{u}^T A\mathbf{u} - \mathbf{u}^T\mathbf{b} + \lambda\left(\mathbf{w}^T\mathbf{u} - (m+1)U\right),$$

and find where the gradient is zero (in general one looks for the *saddle points* of $\mathcal{L}$). Here the gradient is defined as

$$\nabla = \begin{bmatrix} \frac{\partial}{\partial u_0} & \frac{\partial}{\partial u_1} & \cdots & \frac{\partial}{\partial u_{m+1}} & \frac{\partial}{\partial \lambda} \end{bmatrix}^T.$$

Applying this to the Lagrangian gives

$$\nabla\mathcal{L}(\mathbf{u},\lambda) = A\mathbf{u} - \mathbf{b} + \lambda\mathbf{w} = 0,$$

which is a linear system of $m+2$ equations and $m+3$ unknowns, $\mathbf{u}$ and λ. Using the constraint $\mathbf{w}^T\mathbf{u} = (m+1)U$ gives one extra equation and leads to the $(m+3)$-by-$(m+3)$ linear system of equations

$$\begin{bmatrix} A & \mathbf{w} \\ \mathbf{w}^T & 0 \end{bmatrix}\begin{bmatrix} \mathbf{u} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ (m+1)U \end{bmatrix}. \tag{5}$$

This is an example of a more general type of linear system called a "saddle point system".

(a) Show that the linear system (5) has a unique solution regardless of $\mathbf{b}$.

Hint: One way to proceed is to show that the only solution to (i) $A\mathbf{u} + \lambda\mathbf{w} = \mathbf{0}$ and (ii) $\mathbf{w}^T\mathbf{u} = 0$ (i.e. the system (5) with the right hand side set to zero) is $\mathbf{u} = \mathbf{0}$ and $\lambda = 0$. One way to show this is to consider what happens when multiplying the first equation by $\mathbf{w}^T$. This should give you the result that $A\mathbf{u} = \mathbf{0}$, which means $\mathbf{u} = \alpha\mathbf{e}$ for some α. You can then use the equation $\mathbf{w}^T\mathbf{u} = 0$ to show that $\alpha = 0$.

(b) Show that if $\mathbf{w}^T\mathbf{b} = 0$ in (5) then $\lambda = 0$. This means that the solution $\mathbf{u}$ from (5) solves the original system (2) and thus gives an approximate solution to the BVP (6).

Note: If $\mathbf{w}^T\mathbf{b} \neq 0$ then the solution $\mathbf{u}$ from (5) does not solve (2), however, $\mathbf{u}$ may still approximately solve the BVP (6). This will be reflected in the size of λ, since $\|A\mathbf{u} - \mathbf{b}\| = |\lambda|\|\mathbf{w}\|$. If $|\lambda| = O(h^2)$ then $\mathbf{u}$ will solve the BVP up to $O(h^2)$ since this is the truncation error of the FD method.

(c) Solve the BVP (6) numerically with $a = 0$, $b = 2\pi$, $m = 99$, $f(x) = -4\cos(2x)$, and $\sigma_0 = \sigma_1 = 0$ using the augmented linear system (5). Set $U = 0$ and check your answer against the true solution $u(x) = \cos(2x)$. Plot the error in the approximate solution and report the relative two-norm of the error. Also report the value of λ. What does this value tell you about your solution in regards to solving the system (2).

(d) Repeat part (c) but now for the function $f(x) = x$, $\sigma_0 = -\pi^2$, and $\sigma_1 = \pi^2$. Plot the numerical solution you obtain and report the value of λ. Does it seem reasonable that this approximately solves (2)? Explain.

4. *Neumann-Neumman boundary conditions and the Discrete Cosine Transform.* Consider again the Neumann-Neumann Poisson equation, but now with strictly zero-flux boundary conditions

$$u''(x) = f(x),\ a \le x \le b,\ u'(a) = 0,\ u'(b) = 0. \tag{6}$$

In this problem, we will look at different way to solve the corresponding linear system (2) that results from discretizing this problem with a second-order accurate FD formula. Here we again let $h = (b-a)/(m+1)$, $x_j = a + jh$, $u_j \approx u(x_j)$, and $f_j = f(x_j)$ for $j = 0, \ldots, m+1$. In this approximation, the fictitious point idea has been applied, which used the assumption

$$\frac{u_1 - u_{-1}}{2h} = 0 \Longleftrightarrow u_1 = u_{-1} \quad \text{and} \quad \frac{u_{m+2} - u_m}{2h} = 0 \Longleftrightarrow u_m = u_{m+2}, \tag{7}$$

where $u_{-1} \approx u(x_0 - h)$ and $u_{m+2} \approx u(x_{m+1} + h)$. The method that we will use is based on the Discrete Cosine Transform (DCT).

(a) Consider the vectors $\mathbf{u}$ and $\mathbf{f} = \mathbf{b}$ in (2) of length $m+2$. Let $\hat{\mathbf{u}}$ and $\hat{\mathbf{f}}$ denote the DCT of these vectors, respectively, i.e.

$$\hat{u}_k = \frac{1}{m+1}\left[\frac{1}{2}u_0 + \sum_{j=1}^{m} u_j \cos\left(\frac{\pi jk}{m+1}\right) + \frac{1}{2}u_{m+1}\cos(k\pi)\right] := \frac{1}{m+1}\sum_{j=0}^{m+1}{}'' u_j \cos\left(\frac{\pi jk}{m+1}\right),$$

$$\hat{f}_k = \frac{1}{m+1}\left[\frac{1}{2}f_0 + \sum_{j=1}^{m} f_j \cos\left(\frac{\pi jk}{m+1}\right) + \frac{1}{2}f_{m+1}\cos(k\pi)\right] := \frac{1}{m+1}\sum_{j=0}^{m+1}{}'' f_j \cos\left(\frac{\pi jk}{m+1}\right),$$

for $k = 0, \ldots, m+1$. Now, we can express the entries of $\mathbf{u}$ and $\mathbf{b}$ as

$$u_j = \hat{u}_0 + 2\sum_{k=1}^{m} \hat{u}_k \cos\left(\frac{\pi jk}{m+1}\right) + \hat{u}_{m+1}\cos(\pi j) := 2\sum_{k=0}^{m+1}{}'' \hat{u}_k \cos\left(\frac{\pi jk}{m+1}\right),$$

$$f_j = \hat{f}_0 + 2\sum_{k=1}^{m} \hat{f}_k \cos\left(\frac{\pi jk}{m+1}\right) + \hat{f}_{m+1}\cos(\pi j) := 2\sum_{k=0}^{m+1}{}'' \hat{f}_k \cos\left(\frac{\pi jk}{m+1}\right).$$

Substitute these expressions into the linear system (2) to show that row j of this systems simplifies to

$$\sum_{k=0}^{m+1}{}'' \hat{u}_k \left(2\cos\left(\frac{\pi k}{m+1}\right) - 2\right)\cos\left(\frac{\pi jk}{m+1}\right) = h^2 \sum_{k=0}^{m+1}{}'' \hat{f}_k \cos\left(\frac{\pi jk}{m+1}\right),$$

in the case of $j = 0$ and $j = m+1$ use (7).

(b) Use the results from part (a) to show that the DCT of the solution to (2) for $k = 1, \ldots, m+1$ is given by

$$\hat{u}_k = \frac{h^2 \hat{f}_k}{2\cos\left(\frac{\pi k}{m+1}\right) - 2}.$$

However, for $k = 0$, the value of $\hat{u}_k$ is undefined. For this case, note that if $\hat{f}_0 = 0$ (or "numerically zero") then $\hat{u}_0$ can be chosen arbitrarily. Show how the condition $\hat{f}_0 = 0$ corresponds to the discrete compatibility condition $\mathbf{w}^T\mathbf{b} = \mathbf{w}^T\mathbf{f} = 0$ discussed in the previous problem.

(c) Put parts (a) and (b) together to explain how to obtain the solution to (2) when the right hand satisfies $\mathbf{w}^T\mathbf{f} = 0$. Also explain how one makes the solution unique by fixing the arbitrary constant to U.

(d) The MATLAB codes `dct.m` and `idct.m` on the course webpage compute the DCT and inverse DCT (these codes use the FFT to do the computation and are thus 'fast', although not as fast as they could be). Use these codes (or the equivalent codes in another language) and the procedure outlined in (a)–(c) to solve the problem from 4(c). Check your answer against the true solution $u(x) = \cos(2x)$. Plot the error in the approximate solution and report the relative two-norm of the error. Compare your solution from this problem to your solution from 3(c) above. You should find that they are the same (up to rounding errors).

(e) Extra credit (5 points): Why should the solution computed using the techniques from this problem be mathematically equivalent to the solution using the techniques from problem 3.