Study Questions Homework 4 Foundations of Computational Math 2 Spring 2022

Problem 4.1

4.1.a

Recall Simpson's **second** rule approximates the integral

$$I(f) = \int_a^b f(x)dx$$

by

$$I_{s2r}(f) = h_3 \, \frac{3}{8} \, [f_0 + 3f_1 + 3f_2 + f_3]$$

with error

$$I(f) - I_{s2r}(f) = -\frac{3}{80} \, h_3^5 \, f^{(4)} + O(h_3^6), \quad h_3 = (b-a)/3.$$

This method can be used to define a composite method, I_{cs2}, by using rule I_{s2r} on a set of intervals $[a_i, b_i]$ for $i = 1, \ldots, m$ and summing the values.

(4.1.a.i) Suppose m intervals are used each of width $H = (b-a)/m$. Determine the expression for the composite Simpson's second rule to approximate

$$I(f) = \int_a^b f(x)dx$$

Be careful with the difference between H and h_3 in your solution.

(4.1.a.ii) Suppose m intervals are used each of width $H = (b-a)/m$. Determine the expression for the error for the composite Simpson's second rule

$$E_{cs2} = I(f) - I_{cs2}$$

Be careful with the difference between H and h_3 in your solution.

4.1.b

Recall that

- The Legendre polynomials are given by the recurrence:

$$P_0(x) = 1, \; P_1(x) = x, \; P_{n+1}(x) = \frac{2n+1}{n+1} x P_n(x) - \frac{n}{n+1} P_{n-1}(x)$$

the next two are: $P_2(x) = \frac{1}{2}(3x^2 - 1), \quad P_3(x) = \frac{1}{2}(5x^3 - 3x)$

- Gauss-Legendre quadrature methods are polynomial interpolation-based quadrature methods that set the mesh points x_i, $0 \le i \le n$ to the roots of $P_{n+1}(x)$, the Legendre polynomial of degree $n+1$.

- The Chebyshev polynomials are given by the recurrence:

$$T_0(x) = 1, \;\; T_1(x) = x, \;\; T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$
$$\text{the next two are: } T_2(x) = 2x^2 - 1, \;\; T_3(x) = 4x^3 - 3x.$$

- Classical Clenshaw-Curtis (Fejér) quadrature methods are polynomial interpolation-based quadrature methods that set the mesh points x_i, $0 \le i \le n$ to the roots of $T_{n+1}(x)$, the Chebyshev polynomial of degree $n+1$.

(4.1.b.i) Determine the degrees of exactness for the two-point ($n = 1$) methods in each of these two families of polynomial interpolation-based quadrature methods applied to the definite integral

$$I(f) = \int_{-1}^{1} f(x)dx$$

(4.1.b.ii) Comment, briefly, on the main similarity and difference between Gauss-Chebyshev quadrature methods and the Classical Clenshaw-Curtis (Fejér) quadrature methods.

Problem 4.2

If $A \in \mathbb{C}^{n_1 \times n_2}$ and $B \in \mathbb{C}^{n_3 \times n_4}$ then the Kronecker product

$$M = A \otimes B \in \mathbb{C}^{n_1 n_3 \times n_2 n_4}$$

is defined in terms of blocks $M_{ij} \in \mathbb{C}^{n_3 \times n_4}$ for $1 \le i \le n_1$ and $1 \le j \le n_2$ where

$$M_{ij} = \alpha_{ij} B.$$

The Kronecker product is useful for expressing many structured matrix expressions, e.g., the Cooley-Tukey FFT/IFFT.

Let $A \in \mathbb{C}^{m \times m}$, $B \in \mathbb{C}^{n \times n}$, $x \in \mathbb{C}^{mn}$, and $y \in \mathbb{C}^{mn}$.

4.2.a. Describe an algorithm to evaluate the matrix vector product

$$y = (A \otimes B)x$$

i.e., given A, B, x determine y.

4.2.b. What is the complexity of the algorithm?

4.2.c. How does the complexity of the algorithm compare to the standard matrix-vector product computation, $y = Mx$, that ignores the structure of M.

Problem 4.3

The factored form of the Cooley-Tukey FFT

$$F_n = (A_1 A_2 \cdots A_{k-1})\, D_n P_n = \left(\prod_{i=1}^{k-1} A_i \right) D_n P_n, \tag{1}$$

where each A_i is scaled by $1/\sqrt{2}$ and has the block structure using I and Ω of the appropriate dimensions, P_n is the bit-reversal permutation matrix and $D_n = diag(F_2, \cdots, F_2)$ is a block diagonal matrix with $n/2$, 2×2 DFT matrices, was derived in the class notes by using the basic properties of the n roots of unity and writing a polynomial in the monomial basis in terms of the sum of the polynomials involving the even and odd power terms.

Given the relationship between $\omega_n = e^{i\theta_n}$ and $\mu_n = \bar{\omega}_n$, with $\theta_n = 2\pi/n$, the same proof can be repeated with μ_n replaced by ω_n to derive the IFFT as the factorization

$$F_n^H = \left(\overline{A}_1 \overline{A}_2 \cdots \overline{A}_{k-1}\right) \overline{D}_n P_n = \left(\prod_{i=1}^{k-1} \overline{A}_i \right) \overline{D}_n P_n, \tag{2}$$

where $\overline{M}$ replaces elements with their complex conjugates. This is equivalent to the factored form of the FFT with μ replaced by ω.

Recall the basic properties of the matrices F and F^H:

$$F = (F)^T, \quad F^H = \left(F^H\right)^T$$

$$F^H F = I = F F^H \rightarrow F^H = F^{-1}.$$

Show that these properties can be used to derive (2) directly from (1).

Problem 4.4

Definitions

Let $F_n \in \mathbb{C}^{n \times n}$ be the unitary matrix representing the discrete Fourier transform of length n and so $F_n^H \in \mathbb{C}^{n \times n}$ is the inverse DFT of length n. For example, for $n = 4$

$$F_4 = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \mu & \mu^2 & \mu^3 \\ 1 & \mu^2 & \mu^4 & \mu^6 \\ 1 & \mu^3 & \mu^6 & \mu^9 \end{pmatrix} \quad \text{and} \quad F_4^H = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & \omega^4 & \omega^6 \\ 1 & \omega^3 & \omega^6 & \omega^9 \end{pmatrix}$$

where $\theta = 2\pi/n$, $\omega = e^{i\theta}$ and $\mu = e^{-i\theta}$.

Let $Z_n \in \mathbb{C}^{n\times n}$ be the permutation matrix of order n such that Zv represents the circulant "upshift" of the elements of the vector v, i.e.,

$$Z_n = \begin{pmatrix} e_n & e_1 & e_2 & \dots & e_{n-1} \end{pmatrix}.$$

For example, for $n = 4$

$$Z_4 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Let $C_n \in \mathbb{C}^{n\times n}$ be a circulant matrix of order n. The circulant matrix C_n has n parameters (either the first row or first column can be viewed as these parameters). It is a Toeplitz matrix (all diagonals are constant) with the additional constraint that each row (column) is a circulant shift of the previous row (column).

For example, for $n = 4$ and using the first row as the parameters we have

$$C_4 = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_3 & \alpha_0 & \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_3 & \alpha_0 & \alpha_1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_0 \end{pmatrix}$$

Given a polynomial of degree d, a matrix polynomial is defined as follows

$$P_d(\xi) = \delta_0 + \delta_1\xi + \delta_2\xi^2 + \cdots + \delta_d\xi^d$$

$$P_d(A) = \delta_0 I + \delta_1 A + \delta_2 A^2 + \cdots + \delta_d A^d$$

$$\xi, \in \mathbb{C}, \quad \delta_i \in \mathbb{C}, \quad P_d(A), \; A \in \mathbb{C}^{n\times n}.$$

Hint: For the problems below it might be useful to consider a small dimension, e.g., $n = 4$ and then generalize the proofs and results to any n.

(4.4.a) Determine a diagonal matrix $\Lambda_n \in \mathbb{C}^{n\times n}$ i.e., nonzero elements may only appear on the main diagonal, that satisfies $Z_n = F_n^H \Lambda_n F_n$. This says that the columns of F_n^H are the eigenvectors of Z_n and the associated eigenvalues are the elements on the diagonal of Λ_n.

(4.4.b) Recall, that the set of $n \times n$ matrices is a vector space with dimension n^2. Show that the set of $n \times n$ circulant matrices, C_n, is a subspace of that vector space with dimension n. Hint: find a basis for the subspace using the results and definitions above.

(4.4.c) Show that any circulant matrix can be written

$$C_n = F_n^H \Gamma_n F_n$$

where $\Gamma_n \in \mathbb{C}^{n\times n}$ is a diagonal matrix. This says that the columns of F_n^H are the eigenvectors of C_n and the associated eigenvalues are the elements on the diagonal of Γ_n. Your proof should develop a formula for Γ_n that allows its diagonal elements to be easily evaluated and understood.

(4.4.d) Describe how you determine if C_n is a nonsingular matrix.

(4.4.e) How does this factorization of C_n result in a fast method of solving a linear system $C_n x = b$, where $x,\ b \in \mathbb{C}^n$. (Here a fast method is one that has complexity less than the $O(n^3)$ computations associated with standard factorization methods.)

Problem 4.5

4.5.a

Let $F_n \in \mathbb{C}^{n\times n}$ be the unitary matrix representing the discrete Fourier transform of length n. Justify your answers to the following. Simply giving values will receive no credit.

(i) Determine $\|F_n\|_F$.

(ii) Determine $\|F_n\|_2$.

4.5.b

Let $Z_n \in \mathbb{R}^{n\times n}$ be

$$Z_n = \begin{pmatrix} e_n & e_1 & e_2 & \dots & e_{n-1} \end{pmatrix} = \begin{pmatrix} e_2^T \\ e_3^T \\ e_4^T \\ \vdots \\ e_n^T \\ e_1^T \end{pmatrix}$$

$$\text{for example, } Z_4 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

(i) What is the computational complexity of computing the matrix vector product $w = Z_n v$, i.e., given Z_n and v compute w, where $w, v \in \mathbb{R}^n$?

(ii) Describe how you would solve the linear system $Z_n x = b$ for x with $x, b \in \mathbb{R}^n$? What is the computational complexity of your algorithm?

(iii) Describe how you would solve the linear system

$$(2\ Z_n + 3\ Z_n^2)x = b$$

for x with $x, b \in \mathbb{R}^n$? What is the computational complexity of your algorithm?

4.5.c

Recall that a function a function $f(x) \in \mathcal{L}^2[0, 2\pi]$ can be written in terms of its Generalized Fourier Series

$$f(x) = \sum_{m=-\infty}^{\infty} \alpha_m e^{imx}.$$

Given $n > 0$ and $\theta = 2\pi/n$, $f(x)$ has a discrete Fourier reconstruction, $q_n(x)$, that uses numerical quadrature and uniform samples of $f(x)$ defined by multiples of θ to compute the coefficients $\hat{\alpha}_m$ for $-n/2 \leq m \leq n/2 - 1$ defining

$$q_n(x) = \sum_{m=-n/2}^{n/2-1} \hat{\alpha}_m e^{imx}.$$

Let $g(x) = 1 + e^{ix}$ and determine a function $f(x) \neq g(x)$ such that the discrete Fourier reconstruction of $f(x)$ satisfies

$$q_n(x) = \sum_{m=-n/2}^{n/2-1} \hat{\alpha}_m e^{imx} = g(x).$$

Problem 4.6

Let x and y be two infinite sequences, i.e.,

$$x = \{\ldots \xi_{-4},\ \xi_{-3},\ \xi_{-2},\ \xi_{-1},\ \xi_0,\ \xi_1,\ \xi_2,\ \xi_3,\ \xi_4,\ \ldots\}$$

$$y = \{\ldots \eta_{-4},\ \eta_{-3},\ \eta_{-2},\ \eta_{-1},\ \eta_0,\ \eta_1,\ \eta_2,\ \eta_3,\ \eta_4,\ \ldots\}$$

The convolution $z = x * y$ is an infinite sequence with elements

$$\zeta_k = \sum_{i=-\infty}^{\infty} \eta_i \xi_{i+k}$$

Note ζ_k lines up η_0 with ξ_k and then takes the sum of pairwise products.

Now consider the structured sequences x and y where x is periodic with period n defined by the values $\mu_0, \mu_1, \ldots, \mu_{n-1}$ and y is nonzero only in n elements starting at $i = 0$ defined by the values $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$. That is

$$\ldots, \xi_{-n} = \mu_0, \xi_{-n+1} = \mu_1, \ldots, \xi_{-1} = \mu_{n-1}, \xi_0 = \mu_0, \xi_1 = \mu_1, \ldots, \xi_{n-1} = \mu_{n-1}, \xi_n = \mu_0, \ldots$$

$$\ldots, \eta_{-n} = 0, \ldots, \eta_{-1} = 0, \eta_0 = \alpha_0, \eta_1 = \alpha_1, \ldots, \eta_{n-1} = \alpha_{n-1}, \eta_n = 0, \ldots .$$

For example, for $n = 4$ we have

$$\begin{aligned} x = \{ & \;\ldots\; \xi_{-4}, \; \xi_{-3}, \; \xi_{-2}, \; \xi_{-1}, \; \xi_0, \; \xi_1, \; \xi_2, \; \xi_3, \; \xi_4, \; \ldots \; \} \\ = \{ & \;\ldots\; \mu_0, \; \mu_1, \; \mu_2, \; \mu_3, \; \mu_0, \; \mu_1, \; \mu_2, \; \mu_3, \; \mu_0, \; \ldots \; \} \\ y = \{ & \;\ldots\; \eta_{-4}, \; \eta_{-3}, \; \eta_{-2}, \; \eta_{-1}, \; \eta_0, \; \eta_1, \; \eta_2, \; \eta_3, \; \eta_4, \; \ldots \; \} \\ = \{ & \;\ldots\; 0, \; 0, \; 0, \; 0, \; \alpha_0, \; \alpha_1, \; \alpha_2, \; \alpha_3, \; 0, \; \ldots \; \} \end{aligned}$$

(4.6.a) Show that given the values that specify the structured x and y sequences the convolution $z = x * y$ is also specificed by only n values and identify the structure of the sequence z.

(4.6.b) Determine the complexity in terms of n required to compute the n values that specify the convolution z from the values that specify the structured x and y sequences and describe an algorithm that achieves this complexity. **Hint: Relate the values that specify the structured x, y to the values that specify z with a structured matrix operation.**

In your solution you may discuss the specific case $n = 4$ to simplify the presentation but make sure to indicate how the conclusions generalize to $n \neq 4$.

Problem 4.7

Consider the roots of unity needed for a radix-2 Cooley-Tukey version of the FFT of length $n = 2^t$

$$\hat{f} = F_n f = \frac{1}{\sqrt{n}} A_0 A_1 \ldots A_{t-1} P_n f$$

where P_n is the bit reversal permutation, $A_k = I_{2^k} \otimes B_{2^{t-k}}$, $k = 0, 1, \ldots, t-1$, and

$$B_r = \begin{pmatrix} I_s & \Omega_s \\ I_s & -\Omega_s \end{pmatrix}$$

$$\Omega_s = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & \mu_r & 0 & \ldots & 0 \\ 0 & 0 & \mu_r^2 & \ldots & 0 \\ & & & \ddots & \\ 0 & 0 & \ldots & 0 & \mu_r^{s-1} \end{pmatrix}, \quad B_2 = \begin{pmatrix} 1 & 1 \\ 1 & \mu_2 \end{pmatrix}, \quad \mu_r = e^{-2\pi i/r}, \ ; r = 2s$$

(4.7.a) Identify the relationships between the roots of unity needed to define each of the A_k.

(4.7.b) Describe an algorithm to compute the required roots of unity. Try to make the critical path of the computation as short as possible as a function of n since its length is the coefficient of unit roundoff in the order bound on numerical error.

Problem 4.8

Consider a Cooley-Tukey version of the FFT of length $n = 16$ that uses radix-4 rather than radix-2, i.e., at each level of the FFT, all of the DFT's of length k are split into 4 each of length $k/4$. For $n = 16$ this implies

$$\hat{f} = F_{16} f = \frac{1}{\sqrt{16}} A_0 A_1 P_{16} f$$

where P_{16} is a permutation, $A_k = I_{4^k} \otimes B_{4^{t-k}}$, $k = 0, 1, \ldots, t-1$, and B_r is appropriately modified from the radix-2 version.

(4.8.a) Derive the factorization and define the A_k's and P_{16}.

(4.8.b) Discuss the scatter form of P_{16} and its inverse permutation.

(4.8.c) Give the "wiring diagram" or compuational graph for F_{16} based on a radix-4 generalization of the radix-2 butterfly node we have described in class.