Remember to write your name! You are **not** allowed to use a calculator, the textbook, your notes, the internet, or your neighbor. To receive full credit on a problem you must show **sufficient justification for your conclusion** unless explicitly stated otherwise. You may quote any relevant theorem from the textbook or from the lectures: Don't re-prove theorems from class. Everything is real-valued unless specified otherwise. You must do problem #1. Choose two of the three remaining problems. If you submit answers to all problems, problems 1–3 will be graded. All problems are worth 33 points and you get 1 point just for turning in the exam.

Name:

---

1. Consider the fixed-point iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \omega \boldsymbol{f}(\boldsymbol{x}_k)$ for solving $\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}$. Suppose that there is an $\boldsymbol{\alpha}$ such that $\boldsymbol{f}(\boldsymbol{\alpha}) = \mathbf{0}$, and that the Jacobian $\mathbf{J}$ of $\boldsymbol{f}$ satisfies

$$\|\mathbf{J}(\boldsymbol{x})\|_2 \leq B \text{ for all } \boldsymbol{x}$$

and

$$\min_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^T\left(\mathbf{J}+\mathbf{J}^T\right)\boldsymbol{u} \geq C > 0 \text{ for all } \boldsymbol{x}.$$

(Note that this last condition is that the smallest eigenvalue of the symmetric part of $\mathbf{J}$ is greater than or equal to $(C/2) > 0$.) Prove that if $\omega = \quad C/(2B^2)$ then the iteration is locally convergent in the vicinity of the root.

Solution: We know that a fixed point exists, so we will evaluate the Jacobian of the iteration function at the fixed point and see if its spectral radius is less than 1. The Jacobian of the iteration function is

$$\mathbf{G} = \mathbf{I} - \omega\mathbf{J}.$$

We do not know anything about the eigenvalues of $\mathbf{J}$; we only know about the 2-norm of $\mathbf{J}$ and the eigenvalues of the symmetric part of $\mathbf{J}$. This suggests that we might try to show that the 2 norm of $\mathbf{G}$ is less than one, rather than dealing with the spectral radius directly. From the definition,

$$\|\mathbf{G}\|_2^2 = \max_{\|\boldsymbol{u}\|=1}\|\mathbf{G}\boldsymbol{u}\|_2^2 = \max_{\|\boldsymbol{u}\|=1}\boldsymbol{u}^T\mathbf{G}^T\mathbf{G}\boldsymbol{u} = \max_{\|\boldsymbol{u}\|=1}\boldsymbol{u}^T(\mathbf{I}-\omega\mathbf{J})^T(\mathbf{I}-\omega\mathbf{J})\boldsymbol{u}$$

$$= \max_{\|\boldsymbol{u}\|=1}\left[\boldsymbol{u}^T\boldsymbol{u} - \omega\boldsymbol{u}^T\left(\mathbf{J}+\mathbf{J}^T\right)\boldsymbol{u} + \omega^2\boldsymbol{u}^T\mathbf{J}^T\mathbf{J}\boldsymbol{u}\right].$$

Now notice that

$$\boldsymbol{u}^T\boldsymbol{u} = 1,\ \max_{\|\boldsymbol{u}\|=1}\boldsymbol{u}^T\mathbf{J}^T\mathbf{J}\boldsymbol{u} = \|\mathbf{J}\|_2^2 \leq B^2,$$

and

$$\max_{\|\boldsymbol{u}\|=1} -\boldsymbol{u}^T\left(\mathbf{J}+\mathbf{J}^T\right)\boldsymbol{u} = -\min_{\|\boldsymbol{u}\|=1}\boldsymbol{u}^T\left(\mathbf{J}+\mathbf{J}^T\right)\boldsymbol{u} \leq -C < 0.$$

Plugging these in together with the definition of ω we have

$$\|\mathbf{G}\|_2^2 \leq 1 - \frac{C^2}{2B^2} + \frac{C^2}{4B^4}B^2 = 1 - \frac{C^2}{4B^2} < 1.$$

This proves that the iteration is locally convergent. (In fact, the assumptions are strong enough to show that it is globally convergent.)

If your attempt at a proof assumes that the eigenvalues and eigenvectors of $\mathbf{J}$ are real, you will receive at best partial credit, because the problem does not specify that $\mathbf{J}$ is normal, or even diagonalizable. It's not necessary to separately prove that $1 - C^2/(4B^2) \geq 0$ because that is already implied by

$$0 \leq \|\mathbf{G}\|_2^2 \leq 1 - \frac{C^2}{2B^2} + \frac{C^2}{4B^4}B^2 = 1 - \frac{C^2}{4B^2} < 1.$$

2. Suppose that $\mathbf{A}$ is strictly diagonally dominant with positive diagonal entries and nonpositive off-diagonal entries. The Jacobi iteration is convergent for this matrix (you do not need to prove this).

   (a) (25 points) Use the Jacobi iteration to express $\mathbf{A}^{-1}$ as a series, i.e. an infinite sum of matrices.

   Solution: Let $\mathbf{A} = \mathbf{D} - \mathbf{N}$, where $\mathbf{D}$ is the diagonal part of $\mathbf{A}$. The Jacobi iteration is

   $$\boldsymbol{x}_{k+1} = \mathbf{D}^{-1}\mathbf{N}\boldsymbol{x}_k + \mathbf{D}^{-1}\boldsymbol{b}.$$

   Set $\boldsymbol{x}_0 = \mathbf{0}$ and note that the Jacobi iterates are

   $$\boldsymbol{x}_1 = \mathbf{D}^{-1}\boldsymbol{b}$$
   $$\boldsymbol{x}_2 = \mathbf{D}^{-1}\mathbf{N}\mathbf{D}^{-1}\boldsymbol{b} + \mathbf{D}^{-1}\boldsymbol{b}$$
   $$\boldsymbol{x}_3 = \mathbf{D}^{-1}\mathbf{N}\mathbf{D}^{-1}\mathbf{N}\mathbf{D}^{-1}\boldsymbol{b} + \mathbf{D}^{-1}\mathbf{N}\mathbf{D}^{-1}\boldsymbol{b} + \mathbf{D}^{-1}\boldsymbol{b}$$

   Conclude that

   $$\boldsymbol{x}_k = \left(\sum_{j=0}^{k-1}(\mathbf{D}^{-1}\mathbf{N})^j\right)\mathbf{D}^{-1}\boldsymbol{b}.$$

   Since the method is convergent we have that

   $$\mathbf{A}^{-1}\boldsymbol{b} = \lim_{k\to\infty}\boldsymbol{x}_k = \left(\sum_{j=0}^{\infty}(\mathbf{D}^{-1}\mathbf{N})^j\right)\mathbf{D}^{-1}\boldsymbol{b}$$

   which implies

   $$\left(\sum_{j=0}^{\infty}(\mathbf{D}^{-1}\mathbf{N})^j\right)\mathbf{D}^{-1} = \mathbf{A}^{-1}.$$

   (b) (8 points) Use (a) to show that the entries of $\mathbf{A}^{-1}$ are non-negative.

   Solution: Since both $\mathbf{D}$ and $\mathbf{N}$ have non-negative entries, and $\mathbf{D}^{-1}$ also has non-negative entries, every term in the series is a product of matrices with non-negative entries, so every term in the series is a matrix with non-negative entries. Adding them up yields a matrix with non-negative entries.

3. Let $x_0 < x_1 < \ldots < x_n$. Suppose that p and q are polynomials of degree at most $n-1$ such that

$$p(x_{j-1}) = f(x_{j-1}),\ \ q(x_j) = f(x_j),\ \text{for } j = 1, \ldots, n$$

for some function f whose n^{th} derivative is positive and continuous for $x \in [x_0, x_n]$. Prove that $f(x)$ is between $p(x)$ and $q(x)$ for all $x \in [x_0, x_n]$.

Solution: The reference to the n^{th} derivative being positive should suggest to you to look at the interpolation error formula, so begin by quoting this for both p and q

$$f(x) - p(x) = \frac{(x - x_0)\ldots(x - x_{n-1})}{n!} f^{(n)}(\xi),$$

$$f(x) - q(x) = \frac{(x - x_1)\ldots(x - x_n)}{n!} f^{(n)}(\eta).$$

Next consider what it means for f to be between p and q. One way to write this mathematically is

$$(f(x) - p(x))(f(x) - q(x)) \le 0$$

i.e. the errors are opposite-signed. To try to show this we start by multiplying the error formulas above:

$$(f(x) - p(x))(f(x) - q(x)) = \frac{1}{(n!)^2}(x - x_0)(x - x_1)^2 \cdots (x - x_{n-1})^2 (x - x_n) f^{(n)}(\xi) f^{(n)}(\eta).$$

Since

$$\frac{1}{(n!)^2}(x - x_1)^2 \cdots (x - x_{n-1})^2 f^{(n)}(\xi) f^{(n)}(\eta) \ge 0,$$

the sign of $(f(x) - p(x))(f(x) - q(x))$ depends on the product $(x - x_0)(x - x_n)$. For $x \in [x_0, x_n]$ the first factor is non-negative while the second is non-positive, so their product is non-negative, which completes the proof.

4. Suppose that

$$f(x) = \sum_{m=-M}^{M} c_m e^{\mathrm{i}mx}$$

for some $M > 0$.

(a) (8 points) Write the formula for the equispaced composite Trapezoid Rule approximation to the integral

$$c_m = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-\mathrm{i}mx} \mathrm{d}x$$

with $N + 1$ points from $x_0 = 0$ to $x_N = 2\pi$.

Solution: The trapezoid rule quadrature formula is

$$c_m \approx \hat{c}_m = \frac{1}{N} \sum_{j=0}^{N-1} f(x_j) e^{-\mathrm{i}mx_j} \text{ where } x_j = \frac{2\pi j}{N}.$$

Since the function is periodic, the first and last points, which each have half weight, have been combined into the first point with full weight.

(b) (25 points) Suppose that M is even. What is the minimum value of N needed to guarantee that the quadrature computes the coefficients c_m exactly for

$$-\frac{M}{2} \leq m \leq \frac{M}{2}?$$

Solution:

$$c_m \approx \hat{c}_m = \frac{1}{N}\sum_{j=0}^{N-1}(\sum_{k=-M}^{M} c_k e^{\mathrm{i}kx_j})e^{-\mathrm{i}mx_j} \text{ where } x_j = \frac{2\pi j}{N}.$$

$$c_m \approx \hat{c}_m = \frac{1}{N}\sum_{k=-M}^{M} c_k \sum_{j=0}^{N-1}\left(e^{2\pi\mathrm{i}\frac{k-m}{N}}\right)^j.$$

If $k - m = qN$ for some integer q, then the inner sum is N, because

$$e^{2\pi\mathrm{i}q} = 1$$

for integer q. Otherwise, the inner sum is geometric, with value

$$\sum_{j=0}^{N-1}\left(e^{2\pi\mathrm{i}\frac{k-m}{N}}\right)^j = \frac{1 - e^{2\pi\mathrm{i}(k-m)}}{1 - e^{2\pi\mathrm{i}\frac{k-m}{N}}} = 0.$$

In order for the quadrature result to be exact for all m in the range $-M/2$ to $M/2$, we need the geometric sum to be zero for all $k \neq m$ in the range $-M$ to M. I.e. we need

$$k - m \neq qN \text{ for all nonzero integers } q, \text{ for } k \text{ from } -M \text{ to } M \text{ and } m \text{ from } -\frac{M}{2} \text{ to } \frac{M}{2}.$$

The largest value that $k-m$ could take over the given range is $\pm 3M/2$, so we can satisfy the requirement if

$$N \geq \frac{3M}{2} + 1.$$

If you are able to recall from class that

$$\hat{c}_m = c_m + c_{m+N} + c_{m-N} + c_{m+2N} + c_{m-2N} + \dots$$

then you can reason to the same conclusion directly from this expression. This fact underlies the celebrated ‘two-thirds rule’ method for dealiasing developed by Steven Orszag.