Take Home Exam: AMSC/CMSC 666
due 5pm, Wednesday, 15 December
SOLUTIONS

(1) Let $Q_\Delta(f)$ denote quadrature over an interval by the trapeziodal rule with uniform subintervals of length Δ. Use the Euler-Maclaurin formula to extrapolate $Q_\Delta(f)$, $Q_{2\Delta}(f)$, $Q_{3\Delta}(f)$, and $Q_{6\Delta}(f)$ to obtain an eighth order accurate quadrature.

Solution. Let $I(f)$ denote the exact value of the integral. For $f \in C^{10}$ the Euler-Maclaurin asymptotic formula then states that

$$Q_\Delta(f) = I(f) + \alpha_2\delta^2 + \alpha_4\delta^4 + \alpha_6\delta^6 + O(\Delta^8)\,.$$

It follows that

$$\begin{aligned}
Q_{2\Delta}(f) &= I(f) + 4\alpha_2\delta^2 + 4^2\alpha_4\delta^4 + 4^3\alpha_6\delta^6 + O(\Delta^8) \\
&= I(f) + 4\alpha_2\delta^2 + 16\alpha_4\delta^4 + 64\alpha_6\delta^6 + O(\Delta^8)\,, \\
Q_{3\Delta}(f) &= I(f) + 9\alpha_2\delta^2 + 9^2\alpha_4\delta^4 + 9^3\alpha_6\delta^6 + O(\Delta^8) \\
&= I(f) + 9\alpha_2\delta^2 + 81\alpha_4\delta^4 + 729\alpha_6\delta^6 + O(\Delta^8)\,, \\
Q_{6\Delta}(f) &= I(f) + 36\alpha_2\delta^2 + 36^2\alpha_4\delta^4 + 36^3\alpha_6\delta^6 + O(\Delta^8) \\
&= I(f) + 36\alpha_2\delta^2 + 1296\alpha_4\delta^4 + 46656\alpha_6\delta^6 + O(\Delta^8)\,,
\end{aligned}$$

There are many ways to extrapolate. About the simplest is to set

$$Q(f) = w_1Q_\Delta(f) + w_2Q_{2\Delta}(f) + w_3Q_{3\Delta}(f) + w_6Q_{6\Delta}(f)\,,$$

where w_1, w_2, w_3, and w_6, satisfy

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 9 & 36 \\ 1 & 16 & 81 & 1296 \\ 1 & 64 & 729 & 46656 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_6 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}\,.$$

The solution of this system is

$$w_1 = \frac{1296}{840}\,, \qquad w_2 = -\frac{567}{840}\,, \qquad w_3 = \frac{112}{840}\,, \qquad w_6 = -\frac{1}{840}\,.$$

This can be obtained numerically, or analytically. □

(2) Derive the one-, two-, three-, and four-point Gaussian quadrature formulas such that

$$\int_{-1}^{1} f(x)x^2\,\mathrm{d}x = \sum_{j=1}^{n} f(x_j)\,w_j\,.$$

Give bounds on the error of these formulas.

Solution. First, the associated orthogonal monic polynomials through fourth degree are

$$p_0(x) = 1\,, \qquad p_1(x) = x\,, \qquad p_2(x) = x^2 - \tfrac{3}{5}\,,$$
$$p_3(x) = x^3 - \tfrac{5}{7}x\,, \qquad p_4(x) = x^4 - \tfrac{10}{9}x^2 + \frac{5}{21}\,.$$

The roots of the polynomials p_1, p_2, p_3, and p_4 respectively are

$$\{0\}\,, \qquad \Big\{\pm\sqrt{\tfrac{3}{5}}\Big\}\,, \qquad \Big\{0\,, \pm\sqrt{\tfrac{5}{7}}\Big\}\,,$$
$$\Big\{\pm\sqrt{\tfrac{5}{9} \pm \tfrac{2}{9}\sqrt{\tfrac{10}{7}}}\Big\}\,.$$

These are the quadrature points for the one-, two-, three-, and four-point Gaussian quadrature formulas respectively.

The one-point Gaussian quadrature formula is

$$\int_{-1}^{1} f(x)\, x^2 \mathrm{d}x \approx f(0) w_1\,,$$

where the weight w_1 is determined by

$$w_1 = \int_{-1}^{1} x^2 \, \mathrm{d}x = \tfrac{2}{3}\,.$$

Hence, $w_1 = \frac{2}{3}$.

The two-point Gaussian quadrature formula is

$$\int_{-1}^{1} f(x)\, x^2 \mathrm{d}x \approx f\big(-\sqrt{\tfrac{3}{5}}\big) w_1 + f\big(\sqrt{\tfrac{3}{5}}\big) w_2\,,$$

where the weights w_1 and w_2 are determined as follows. By symmetry one sets $w_1 = w_2 = w$. This insures that every odd function will be integrated exactly. Then w is determined by

$$2w = \int_{-1}^{1} x^2 \, \mathrm{d}x = \tfrac{2}{3}\,.$$

Hence, $w_1 = w_2 = w = \frac{1}{3}$.

The three-point Gaussian quadrature formula is

$$\int_{-1}^{1} f(x)\, x^2 \mathrm{d}x \approx f\big(-\sqrt{\tfrac{5}{7}}\big) w_1 + f(0) w_2 + f\big(\sqrt{\tfrac{5}{7}}\big) w_3\,,$$

where the weights w_1, w_2, and w_3 are determined as follows. By symmetry one sets $w_1 = w_3 = w$. This insures that every odd function will be integrated exactly. Then w and w_2 are determined by

$$w_2 + 2w = \int_{-1}^{1} x^2 \, \mathrm{d}x = \tfrac{2}{3}\,,$$
$$2\tfrac{5}{7}w = \int_{-1}^{1} x^4 \, \mathrm{d}x = \tfrac{2}{5}\,.$$

Hence, $w_1 = w_3 = w = \frac{7}{25}$ while $w_2 = \frac{8}{75}$.

The four-point Gaussian quadrature formula is

$$\int_{-1}^{1} f(x)\, x^2 \mathrm{d}x \approx f\Big(-\sqrt{\tfrac{5}{9}+\tfrac{2}{9}\sqrt{\tfrac{10}{7}}}\Big)w_1 + f\Big(-\sqrt{\tfrac{5}{9}-\tfrac{2}{9}\sqrt{\tfrac{10}{7}}}\Big)w_2$$
$$+ f\Big(\sqrt{\tfrac{5}{9}-\tfrac{2}{9}\sqrt{\tfrac{10}{7}}}\Big)w_3 + f\Big(\sqrt{\tfrac{5}{9}+\tfrac{2}{9}\sqrt{\tfrac{10}{7}}}\Big)w_4\,,$$

where the weights w_1, w_2, w_3, and w_4 are determined as follows. By symmetry one sets $w_1 = w_4 = w_+$ and $w_2 = w_3 = w_-$. This insures that every odd function will be integrated exactly. Then w_+ and w_- are determined by

$$2w_- + 2w_+ = \int_{-1}^{1} x^2 \, \mathrm{d}x = \tfrac{2}{3}\,,$$
$$2\Big(\tfrac{5}{9}-\tfrac{2}{9}\sqrt{\tfrac{10}{7}}\Big)w_- +$$
$$2\Big(\tfrac{5}{9}+\tfrac{2}{9}\sqrt{\tfrac{10}{7}}\Big)w_+ = \int_{-1}^{1} x^4 \, \mathrm{d}x = \tfrac{2}{5}\,.$$

These equations reduce to

$$w_- + w_+ = \tfrac{1}{3}\,,$$
$$w_+ - w_- = \tfrac{1}{15}\sqrt{\tfrac{7}{10}}\,.$$

Hence, $w_1 = w_4 = w_+ = \frac{1}{6} + \frac{1}{30}\sqrt{\frac{7}{10}}$ while $w_2 = w_3 = w_- = \frac{1}{6} - \frac{1}{30}\sqrt{\frac{7}{10}}$.

When $f \in C^{2n}([-1,1])$ the error of the n-point Gaussian quadrature formula can be generally bounded by

$$\big|I(f) - Q_n(f)\big| \leq \frac{1}{(2n)!}\big\|f^{(2n)}\big\|_\infty \int_{-1}^{1} p_n(x)^2 \, x^2 \mathrm{d}x\,.$$

The square integrals of the polynomials p_1, p_2, p_3, and p_4 may be computed using the fact that

$$\int_{-1}^{1} p_n(x)^2\, x^2 \,\mathrm{d}x = \int_{-1}^{1} p_n(x)\, x^{n+2} \,\mathrm{d}x\,.$$

One finds that

$$\begin{aligned}
\int_{-1}^{1} p_1(x)^2\, x^2 \mathrm{d}x &= \int_{-1}^{1} x^4 \mathrm{d}x = \frac{2}{5}\,,\\
\int_{-1}^{1} p_2(x)^2\, x^2 \mathrm{d}x &= \int_{-1}^{1} x^6 - \tfrac{3}{5}x^4 \mathrm{d}x = \frac{8}{175}\,,\\
\int_{-1}^{1} p_3(x)^2\, x^2 \mathrm{d}x &= \int_{-1}^{1} x^8 - \tfrac{5}{7}x^6 \mathrm{d}x = \frac{8}{441}\,,\\
\int_{-1}^{1} p_4(x)^2\, x^2 \mathrm{d}x &= \int_{-1}^{1} x^{10} - \tfrac{10}{9}x^8 + \tfrac{5}{21}x^6 \mathrm{d}x = \frac{128}{43,659}\,.
\end{aligned}$$

One thereby obtains the bounds

$$\begin{aligned}
\big|I(f) - Q_1(f)\big| &\leq \frac{1}{5}\big\|f^{(2)}\big\|_\infty\,,\\
\big|I(f) - Q_2(f)\big| &\leq \frac{1}{525}\big\|f^{(4)}\big\|_\infty\,,\\
\big|I(f) - Q_3(f)\big| &\leq \frac{1}{39,690}\big\|f^{(6)}\big\|_\infty\,,\\
\big|I(f) - Q_4(f)\big| &\leq \frac{1}{13,752,585}\big\|f^{(8)}\big\|_\infty\,.
\end{aligned}$$

$\square$

(3) We wish to solve $Ax = b$ iteratively where

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}.$$

Show that the Jacobi method converges while the Gauss-Seidel method does not. For what values of the parameter ω does the SOR method converge?

Solution. The matrix A decomposes as $A = D - L - U$ where

$$D = I\,, \quad L = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ -2 & -2 & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The growth matrix for the Jacobi method is

$$G_J = D^{-1}(L+U) = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix}.$$

Its characteristic polynomial is given by

$$p_J(\lambda) = \det\big(\lambda I - G_J\big) = \lambda^3 .$$

Hence, its spectrum is given by $\mathrm{sp}(G_J) = \{0\}$ and its spectral radius is $\rho(G_J) = 0$. Because $\rho(G_J) < 1$ the Jacobi method converges.

The growth matrix for the Gauss-Seidel method is

$$\begin{aligned} G_{GS} &= (D-L)^{-1}U \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}. \end{aligned}$$

Because this matrix is upper triangular, one can read off that its spectrum is given by $\mathrm{sp}(G_{GS}) = \{0, 2\}$ and that its spectral radius is $\rho(G_{GS}) = 2$. Because $\rho(G_{GS}) > 1$ the Gauss-Seidel method diverges.

The growth matrix for the SOR method is

$$G(\omega) = (D - \omega L)^{-1}[(1-\omega)D + \omega U] .$$

Then $\lambda \in \mathrm{sp}\big(G(\omega)\big)$ if and only if

$$\begin{aligned} 0 &= \det\big(\lambda I - G(\omega)\big) \\ &= \det\big(\lambda I - (D-\omega L)^{-1}[(1-\omega)D + \omega U]\big) \\ &= \det\big((D-\omega L)^{-1}\big)\det\big((\lambda+\omega-1)D - \lambda\omega L - \omega U\big) . \end{aligned}$$

Hence, $\lambda \in \mathrm{sp}\big(G(\omega)\big)$ if and only if

$$
\begin{aligned}
0 &= \det\big((\lambda+\omega-1)D - \lambda\omega L - \omega U\big) \\
&= \det\begin{pmatrix} \lambda+\omega-1 & 2\omega & -2\omega \\ \lambda\omega & \lambda+\omega-1 & \omega \\ 2\lambda\omega & 2\lambda\omega & \lambda+\omega-1 \end{pmatrix} \\
&= (\lambda+\omega-1)^3 - 4\omega^3\lambda^2 + 4\omega^3\lambda \\
&= \lambda^3 - (3-3\omega+4\omega^3)\lambda^2 + (3(1-\omega)^2+4\omega^3)\lambda - (1-\omega)^3\,.
\end{aligned}
$$

We must identify those values of $\omega \in \mathbb{R}$ for which all the roots of this cubic equation lie within the unit circle $|\lambda| < 1$.

Because $(1-\omega)^3$ is the product of these roots, a necessary condition that they all lie within the unit circle $|\lambda| < 1$ is that $|1-\omega| < 1$. This means that ω must be restricted to the interval $(0, 2)$.

Because $(3-3\omega+4\omega^3)$ is the product of these roots, a necessary condition that they all lie within the unit circle $|\lambda| < 1$ is that $|3-3\omega+4\omega^3| < 3$. Because ω is already retricted to the interval $(0, 2)$, this new requirement means ω must be restricted to the interval $(0, \frac{\sqrt{3}}{2})$.

Notice that the restriction $\omega \in (0, \frac{\sqrt{3}}{2})$ contains the result of Part (b) because the Gauss-Seidel Method is the special case $\omega = 1$. Indeed, in that case the cubic equation is

$$
0 = \lambda^3 - 4\lambda^2 + 4\lambda\,,
$$

which has one simple root $\lambda = 0$ and one double root $\lambda = 2$.

Now let us assume that $0 < \omega$ is small. An asymptotic analysis shows that the polynomial has one simple simple root and a conjugate pair of simple complex roots with the expansions

$$
\lambda = 1 - \omega - \sigma\omega(4\omega)^{\frac{1}{3}} + O\big(\omega^{\frac{5}{3}}\big)\,,
$$

where σ is one of the three cube roots of unity, $\sigma = 1$, $\sigma = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$, or $\sigma = -\frac{1}{2} - i\frac{\sqrt{3}}{2}$. It is easily checked that when ω is sufficiently small all of these roots lie within the unit circle $|\lambda| < 1$.

Because the roots of the cubic equation depend continuously on ω, and because when $\omega > 0$ is small all these roots these roots lies within the unit circle while when $\omega > \frac{\sqrt{3}}{2}$ at least one root lies outside the unit circle, there must be some $\omega \in (0, \frac{\sqrt{3}}{2}]$ such that at least one root lies on the unit circle. At such an ω there are three possibilities: either 1 is a root, -1 is a root, or there is a conjugate pair of roots $\{\sigma, \bar{\sigma}\}$ with $|\sigma| = 1$. We will consider each of these possibilities.

If 1 is a root of the cubic equation for some ω then, by setting $\lambda = 1$ in the cubic equation, we see that $\omega^3 = 0$. Therefore this possibility does not occur.

If -1 is a root of the cubic equation for some ω then, by setting $\lambda = -1$ in the cubic equation, we see that

$$(\omega - 2)^3 = 8\omega^3 .$$

The only real root of this equation is $\omega = -2$. Therefore this possibility does not occur.

The only possiblity left is that there must be some $\omega \in (0, \frac{\sqrt{3}}{2}]$ with a conjugate pair of complex roots $\{\sigma, \bar{\sigma}\}$ with $|\sigma| = 1$ and a third root λ_o in $(-1, 1)$. These roots must satisfy

$$\begin{aligned} \lambda_o + \sigma + \bar{\sigma} &= 3 - 3\omega + 4\omega^3 , \\ 1 + \lambda_o(\sigma + \bar{\sigma}) &= 3(1 - \omega)^2 + 4\omega^3 , \\ \lambda_o &= (1 - \omega)^3 . \end{aligned}$$

Upon using the third equation above to eliminate λ_o from the first two equations, we obtain

$$\begin{aligned} \sigma + \bar{\sigma} &= 3 - 3\omega + 4\omega^3 - (1 - \omega)^3 \\ &= 2 - 3\omega^2 + 5\omega^3 , \\ (1 - \omega)^3(\sigma + \bar{\sigma}) &= 3(1 - \omega)^2 + 4\omega^3 - 1 \\ &= 2 - 6\omega + 3\omega^2 + 4\omega^3 . \end{aligned}$$

Upon using the first equation above to eliminate $\sigma + \bar{\sigma}$ from the second equation, we obtain

$$(1 - \omega)^3\big(2 - 3\omega^2 + 5\omega^3\big) = 2 - 6\omega + 3\omega^2 + 4\omega^3 .$$

After expanding the left-hand side above and taking advantage of some nice cancellations, this equation becomes

$$8\omega^3 - 24\omega^4 + 18\omega^5 - 5\omega^6 = 0 .$$

We must therefore find $\omega \in (0, \frac{\sqrt{3}}{2}]$ that satisfies the cubic equation

$$\omega^3 - \tfrac{18}{5}\omega^2 + \tfrac{24}{5}\omega - \tfrac{8}{5} = 0 .$$

This equation has only one real root that can be found analytically or approximated either numerically or graphically. Provided I did not make any mistakes, the cubic formula gives this root as

$$\omega_o = \frac{6}{5} - \frac{\gamma}{5} + \frac{4}{5\gamma}\,,$$

$$\text{where} \quad \gamma = \left(44 + 20\sqrt{5}\right)^{\frac{1}{3}}\,.$$

A very rough estimate shows that this number is close to $\frac{1}{2}$.

Because ω_o is only positive value of ω that allows $\lambda \in \mathrm{sp}\big(G(\omega)\big)$ to pass through the unit circle, it is clear that for $\omega \in (0, \omega_o)$ we know that every eigenvalue of $G(\omega)$ lies inside the unit circle, while for $\omega \in [\omega_o, \infty)$ there is a pair of eigenvalues that lie outside the unit circle. We also know that for $\omega \leq 0$ there is at least one eigenvalue that lies outside the unit circle. We therefore conclude that the SOR-method converges for $\omega \in (0, \omega_o)$ and diverges otherwise.

(4) Let $A \in \mathbb{R}^{N\times N}$ be self-adjoint and positive definite with respect to a distinguished real inner product $(\,\cdot\,|\,\cdot\,)$ over $\mathbb{R}^N$. Let $b \in \mathbb{R}^N$. Define

$$f(y) = (y\,|\,Ay) - 2(b\,|\,y) \quad \text{for every } y \in \mathbb{R}^N\,.$$

Consider the steepest descent method to solve $Ax = b$:

choose an initial iterate $x^{(0)} \in \mathbb{R}^N$;

compute the initial residual $r^{(0)} = b - Ax^{(0)}$;

$$\alpha_n = \frac{\big(r^{(n)}\,|\,r^{(n)}\big)}{\big(r^{(n)}\,|\,Ar^{(n)}\big)}\,;$$

$$x^{(n+1)} = x^{(n)} + \alpha_n r^{(n)}\,;$$

$$r^{(n+1)} = r^{(n)} - \alpha_n A r^{(n)}\,.$$

Let $e^{(n)} = x^{(n)} - x$ be the error of the n^{th} iterate.

(a) Let κ be the condition number of A. Prove that

$$\frac{(y\,|\,Ay)(y\,|\,A^{-1}y)}{(y\,|\,y)^2} \leq \frac{(\kappa+1)^2}{4\kappa} \quad \text{for every nonzero } y \in \mathbb{R}^N\,.$$

Hint: Diagonalize, then maximize.

(b) Prove that

$$\frac{\big\|e^{(n+1)}\big\|_A^2}{\big\|e^{(n)}\big\|_A^2} = 1 - \frac{\big(r^{(n)}\,|\,r^{(n)}\big)}{\big(r^{(n)}\,|\,Ar^{(n)}\big)}\,\frac{\big(r^{(n)}\,|\,r^{(n)}\big)}{\big(r^{(n)}\,|\,A^{-1}r^{(n)}\big)}\,,$$

where $\|\cdot\|_A$ denotes the A-norm.

(c) Use the above inequality to derive a bound on $\left\|e^{(n)}\right\|_A$ in terms of κ and $\left\|e^{(0)}\right\|_A$. Compare the result with the similar estimate derived in class for the conjugate gradient method.

Solution of Part (a). The lower bound is easy. For example, we can use the fact that for any nonzero $y \in \mathbb{R}^N$ and any $\alpha \in \mathbb{R}$

$$\begin{aligned} 0 &\leq \big(y + \alpha A^{-1} y \,|\, y + \alpha A^{-1} y\big)_A \\ &= (y \,|\, Ay) + 2\alpha (y \,|\, y) + \alpha^2 \big(y \,|\, A^{-1} y\big) . \end{aligned}$$

Because A is positive definite and y is nonzero, it follows that $(y \,|\, Ay)$, $(y \,|\, y)$, and $(y \,|\, A^{-1}y)$ are all positive. The right-hand side above is therefore a strictly convex quadratic function of α. Minimizing this function over α yields

$$0 \leq (y \,|\, Ay) - \frac{(y \,|\, y)^2}{\big(y \,|\, A^{-1} y\big)} \,,$$

from which the lower bound follows.

To obtain the upper bound we evaluate

$$\max \Big\{ (y \,|\, Ay)(y \,|\, A^{-1} y) \;:\; y \in \mathbb{R}^N \,,\, (y \,|\, y) = 1 \Big\} .$$

To do this we use the method of Lagrange multipliers. Consider the function

$$F(y, \lambda) = \tfrac{1}{2} (y \,|\, Ay)(y \,|\, A^{-1} y) - \lambda [(y \,|\, y) - 1] .$$

One then sets the derivatives of F to zero:

$$\begin{aligned} 0 &= \nabla_y F(y, \lambda) = (y \,|\, A^{-1} y) Ay + (y \,|\, Ay) A^{-1} y - 2\lambda y \,, \\ 0 &= \partial_\lambda F(y, \lambda) = 1 - (y \,|\, y) . \end{aligned}$$

By taking the inner product of the first equation with y and using the second equation to evaluate $(y \,|\, y)$, we find that

$$\lambda = (y \,|\, Ay)(y \,|\, A^{-1} y) .$$

By multiplying the first equation by A and using the above equation to eliminate λ, it can be expressed as

$$A^2 y - 2\kappa_1 A y + \frac{\kappa_1}{\kappa_{-1}} y = 0 \,, \tag{1}$$

where the scalars κ_1 and κ_{-1} are defined by

$$\kappa_1 = (y \,|\, Ay) \,, \qquad \kappa_{-1} = (y \,|\, A^{-1} y) .$$

Because A is positive definite, both κ_1 and κ_{-1} are positive.

Equation (1) will have a solution if and only if zero is in the spectrum of the matrix $q(A)$ where $q(\lambda)$ is the quadratic polynomial given by

$$q(\lambda) = \lambda^2 - 2\kappa_1\lambda + \frac{\kappa_1}{\kappa_{-1}}\,.$$

By the Spectral Mapping Theorem

$$\mathrm{sp}\big(q(A)\big) = \big\{q(\lambda)\,:\,\lambda \in \mathrm{sp}(A)\big\}\,.$$

So there must be at least one $\lambda \in \mathrm{sp}(A)$ such that $q(\lambda) = 0$. This means that $q(\lambda)$ must have the factored form

$$q(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)\,.$$

where at least one of λ_1 and λ_2 must be in $\mathrm{sp}(A)$. By comparing this factor form with the definition of $q(\lambda)$, we read off that

$$\kappa_1 = \frac{\lambda_1 + \lambda_2}{2}\,, \qquad \frac{\kappa_1}{\kappa_{-1}} = \lambda_1\lambda_2\,.$$

Because κ_1 and κ_{-1} are positive, it follows that both λ_1 and λ_2 are positive. We can then express κ_1 and κ_{-1} in terms of λ_1 and λ_2 as

$$\kappa_1 = \frac{\lambda_1 + \lambda_2}{2}\,, \qquad \kappa_{-1} = \frac{\lambda_1 + \lambda_2}{2\lambda_1\lambda_2}\,.$$

It therefore follows from the definition of κ_1 and κ_{-1} that a unit vector y satisfying $q(A)y = 0$ must also satisfy

$$\begin{aligned} (y\,|\,Ay) &= \frac{\lambda_1 + \lambda_2}{2}\,, \\ \big(y\,|\,A^{-1}y\big) &= \frac{\lambda_1 + \lambda_2}{2\lambda_1\lambda_2}\,. \end{aligned} \tag{2}$$

Every such y will be a critical point of $F(y)$. Moreover, the set of all such y will be all the critical points of $F(y)$.

There are three cases to consider: either $\lambda_1 \in \mathrm{sp}(A)$ and $\lambda_2 \notin \mathrm{sp}(A)$, or $\lambda_2 = \lambda_1 \in \mathrm{sp}(A)$, or $\lambda_1, \lambda_2 \in \mathrm{sp}(A)$ and $\lambda_1 < \lambda_2$. In each case we seek a unit vector y such that $q(A)y = 0$ and satisfies (2). We will consider these three cases separately below.

First, consider the case where $\lambda_1 \in \mathrm{sp}(A)$ and $\lambda_2 \notin \mathrm{sp}(A)$. Because

$$0 = q(A)y = (A - \lambda_2 I)(A - \lambda_1 I)y\,,$$

while $(A - \lambda_2 I)$ is invertible, we conclude that

$$(A - \lambda_1 I)y = 0\,.$$

Hence, y must be a unit eigenvector of A associated with λ_1. A direct calculation then shows that

$$(y \,|\, Ay) = \lambda_1 \,, \qquad \big(y \,|\, A^{-1}y\big) = \frac{1}{\lambda_1} \,.$$

It follows immediately from (2) that

$$\lambda_1 = \frac{\lambda_1 + \lambda_2}{2} \,,$$

whereby $\lambda_2 = \lambda_1 \in \mathrm{sp}(A)$ — a contradiction. Therefore this case cannot occur.

Next, consider the case where $\lambda_2 = \lambda_1 \in \mathrm{sp}(A)$. Because

$$0 = q(A)y = (A - \lambda_1 I)^2 y \,,$$

the vector y must be a unit eigenvector of A associated with λ_1. A direct calculation then shows that

$$(y \,|\, Ay) = \lambda_1 \,, \qquad \big(y \,|\, A^{-1}y\big) = \frac{1}{\lambda_1} \,,$$

which is consistant with (2). Therefore every unit eigenvector of A is a critical point of $F(y)$ over the unit sphere. Its critical value is

$$(y \,|\, Ay)(y \,|\, A^{-1}y) = \lambda_1 \, \lambda_1^{-1} = 1 \,.$$

It therefore follows from our lower bound that such a critical point must be a minimum of $F(y)$.

Finally, consider the case where $\lambda_1, \lambda_2 \in \mathrm{sp}(A)$ and $\lambda_1 < \lambda_2$. Because

$$0 = q(A)y = (A - \lambda_2 I)(A - \lambda_1 I)y \,,$$

the vector y must have the form

$$y = \alpha_1 v_1 + \alpha_2 v_2 \,,$$

where $\alpha_1, \alpha_2 \in \mathbb{R}$ while v_1 and v_2 are unit eigenvectors of A associated with λ_1 and λ_2 respectively. Because v_1 and v_2 are orthogonal unit vectors while y is a unit vector, we know that

$$\alpha_1^2 + \alpha_2^2 = 1 \,.$$

A direct calculation then shows that

$$(y \,|\, Ay) = \lambda_1 \alpha_1^2 + \lambda_2 \alpha_2^2 \,,$$
$$\big(y \,|\, A^{-1}y\big) = \frac{1}{\lambda_1} \alpha_1^2 + \frac{1}{\lambda_2} \alpha_2^2 \,.$$

Therefore (2) will be satisfied provided

$$\lambda_1\alpha_1^2 + \lambda_2\alpha_2^2 = \lambda_1\frac{1}{2} + \lambda_2\frac{1}{2}\,,$$
$$\frac{1}{\lambda_1}\alpha_1^2 + \frac{1}{\lambda_2}\alpha_2^2 = \frac{1}{\lambda_1}\frac{1}{2} + \frac{1}{\lambda_2}\frac{1}{2}\,.$$

Because $0 < \lambda_1 < \lambda_2$, one sees that

$$\det\begin{pmatrix}\lambda_1 & \lambda_2\\ \lambda_1^{-1} & \lambda_2^{-1}\end{pmatrix} = \frac{\lambda_1^2 - \lambda_2^2}{\lambda_1\lambda_2} \neq 0\,.$$

We can therefore conclude that

$$\alpha_1^2 = \alpha_2^2 = \frac{1}{2}\,.$$

Therefore every vector of the form

$$y = \frac{v_1 + v_2}{\sqrt{2}}$$

is a critical point of $F(y)$ over the unit sphere whenever v_1 and v_2 are unit eigenvectors of A corresponding to different eigenvalues λ_1 and λ_2. Its critical value is

$$(y\,|\,Ay)(y\,|\,A^{-1}y) = \frac{(\lambda_1+\lambda_2)^2}{4\lambda_1\lambda_2} = \frac{\left(1+\dfrac{\lambda_2}{\lambda_1}\right)^2}{4\dfrac{\lambda_2}{\lambda_1}}\,.$$

This is an increasing function of λ_2/λ_1, so it will take its maximum value when λ_1 is the smallest eigenvalue of A while λ_2 is the largest eigenvalue of A. In that case

$$(y\,|\,Ay)(y\,|\,A^{-1}y) = \frac{(\kappa+1)^2}{4\kappa}\,.$$

As this is the largest value taken by any critical point, we conclude that

$$\max\Big\{(y\,|\,Ay)(y\,|\,A^{-1}y) \;:\; y \in \mathbb{R}^N\,,\ (y\,|\,y) = 1\Big\} = \frac{(\kappa+1)^2}{4\kappa}\,.$$

The result follows by scaling. $\square$

Solution of Part (b). Because the error of the n^{th} iterate is $e^{(n)} = x^{(n)} - x$, the residual of the n^{th} iterate is $r^{(n)} = b - Ax^{(n)}$, while $Ax = b$, we see that

$$Ae^{(n)} = A\big(x^{(n)} - x\big) = Ax^{(n)} - Ax = Ax^{(n)} - b = -r^{(n)}\,.$$

Hence, $e^{(n)} = -A^{-1}r^{(n)}$. It thereby follows from the definition of the A-norm that

$$\|e^{(n)}\|_A^2 = \big(e^{(n)} \,|\, e^{(n)}\big)_A = \big(e^{(n)} \,|\, Ae^{(n)}\big) = \big(A^{-1}r^{(n)} \,|\, r^{(n)}\big) .$$

For the steepest descent method we have

$$r^{(n+1)} = r^{(n)} - \alpha_n A r^{(n)} ,$$

where α_n is given by

$$\alpha_n = \frac{\big(r^{(n)} \,|\, r^{(n)}\big)}{\big(r^{(n)} \,|\, Ar^{(n)}\big)} .$$

Hence, we see that

$$\begin{aligned}
\|e^{(n+1)}\|_A^2 &= \big(A^{-1}r^{(n+1)} \,|\, r^{(n+1)}\big) \\
&= \big(A^{-1}[r^{(n)} - \alpha_n Ar^{(n)}] \,|\, [r^{(n)} - \alpha_n Ar^{(n)}]\big) \\
&= \big(A^{-1}r^{(n)} \,|\, r^{(n)}\big) - 2\alpha_n \big(r^{(n)} \,|\, r^{(n)}\big) + \alpha_n^2 \big(r^{(n)} \,|\, Ar^{(n)}\big) \\
&= \big(A^{-1}r^{(n)} \,|\, r^{(n)}\big) - \frac{\big(r^{(n)} \,|\, r^{(n)}\big)^2}{\big(r^{(n)} \,|\, Ar^{(n)}\big)} .
\end{aligned}$$

Upon dividing both sides above by the quantity $\big(A^{-1}r^{(n)} \,|\, r^{(n)}\big)$ while recalling that this quantity is equal to $\|e^{(n)}\|_A^2$, we obtain

$$\frac{\big\|e^{(n+1)}\big\|_A^2}{\big\|e^{(n)}\big\|_A^2} = 1 - \frac{\big(r^{(n)} \,|\, r^{(n)}\big)}{\big(r^{(n)} \,|\, Ar^{(n)}\big)} \frac{\big(r^{(n)} \,|\, r^{(n)}\big)}{\big(r^{(n)} \,|\, A^{-1}r^{(n)}\big)} ,$$

$\square$

Solution of Part (c). By the result of part (a) we know that

$$\frac{\big(r^{(n)} \,|\, r^{(n)}\big)}{\big(r^{(n)} \,|\, Ar^{(n)}\big)} \frac{\big(r^{(n)} \,|\, r^{(n)}\big)}{\big(r^{(n)} \,|\, A^{-1}r^{(n)}\big)} \geq \frac{4\kappa}{(\kappa+1)^2} .$$

When this is combined with the result from part (b) we obtain

$$\begin{aligned}
\frac{\big\|e^{(n+1)}\big\|_A^2}{\big\|e^{(n)}\big\|_A^2} &\leq 1 - \frac{4\kappa}{(\kappa+1)^2} = \frac{(\kappa+1)^2 - 4\kappa}{(\kappa+1)^2} \\
&= \frac{(\kappa-1)^2}{(\kappa+1)^2} .
\end{aligned}$$

Hence, taking square roots yields

$$\big\|e^{(n+1)}\big\|_A \leq \frac{\kappa - 1}{\kappa + 1} \big\|e^{(n)}\big\|_A$$

By induction we therefore arrive at the convergence estimate

$$\|e^{(n)}\|_A \leq \left(\frac{\kappa-1}{\kappa+1}\right)^n \|e^{(0)}\|_A .$$

The similar estimate derived in class for the conjugate gradient method is

$$\|e^{(n)}\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \|e^{(0)}\|_A .$$

For large κ this convergence factor behaves like

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 1 - \frac{2}{\sqrt{\kappa}} + O\big(\kappa\big) ,$$

while for large κ the steepest descent convergence factor behaves like

$$\frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa} + O\big(\kappa^{-2}\big) .$$

Because

$$\left(1 - \frac{2}{\kappa} + O\big(\kappa^{-2}\big)\right)^{\kappa^{-\frac{1}{2}}} \sim 1 - \frac{2}{\sqrt{\kappa}} + O\big(\kappa\big) ,$$

it would therefore take on the order of $\kappa^{-\frac{1}{2}}$ iterations of the steepest descent to obtain the same estimate on the error as that for one iteration of the conjugate gradient method. $\square$

(5) Let A be the symmetric tridiagonal real matrix

(3)
$$A = \begin{pmatrix} a_0 & b_1 & 0 & \cdots & 0 \\ b_1 & a_1 & b_2 & \ddots & \vdots \\ 0 & b_2 & a_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{pmatrix} .$$

Show that A is irreducible if and only if every b_m is nonzero.

Solution. If $b_m = 0$ for some $m < n$ then A has the reducible form

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} ,$$

where $A_1 \in \mathbb{R}^{m\times m}$ and $A_2 \in \mathbb{R}^{(n-m)\times(n-m)}$ are given by

$$A_1 = \begin{pmatrix} a_0 & \ddots & \\ \ddots & \ddots & b_{m-1} \\ & b_{m-1} & a_{m-1} \end{pmatrix} , \quad A_2 = \begin{pmatrix} a_m & b_{m+1} & \\ b_{m+1} & \ddots & \ddots \\ & \ddots & a_n \end{pmatrix} ,$$

where all terms off the three main diagonals are zero. Therefore A is not irreducible.

Now suppose every b_m is nonzero. The graph associated with A is

$$0 \quad \leftrightarrow \quad 1 \quad \leftrightarrow \quad \cdots \quad \leftrightarrow \quad n-1 \quad \leftrightarrow \quad n$$

Because there is a directed path between any two nodes on this graph, A is therefore irreducible. $\square$

(6) Let A be an irreducible symmetric tridiagonal real matrix of the form (3). Let $\{p_m(x)\}_{m=0}^{n+1}$ be the sequence of polynomials generated by

$$p_0(x) = 1\,, \qquad p_1(x) = (x - a_0)\,,$$
$$p_{m+1}(x) = (x - a_m)p_m(x) - b_m^2 p_{m-1}(x) \quad \text{for } m = 1, \cdots, n\,.$$

Let $\pi_0 = 1$, and $\pi_m = b_m \pi_{m-1}$ for every $m = 1, \cdots, n$. Let $q_m(x) = p_m(x)/\pi_m$ for every $m = 0, \cdots, n$.

(a) Show that $p_{n+1}(x)$ has $n+1$ simple roots $\{x_k\}_{k=0}^{n+1}$.

(b) Show that $V^{-1}AV$ is diagonal where

$$V = \begin{pmatrix} q_0(x_0) & q_0(x_1) & q_0(x_2) & \cdots & q_0(x_n) \\ q_1(x_0) & q_1(x_1) & q_1(x_2) & \cdots & q_1(x_n) \\ q_2(x_0) & q_2(x_1) & q_2(x_2) & \cdots & q_2(x_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_n(x_0) & q_n(x_1) & q_n(x_2) & \cdots & q_n(x_n) \end{pmatrix}.$$

Solution of Part (a). To show that $p_{n+1}(x)$ has $n+1$ simple roots, we prove more. Specifically, we will prove by induction that for every $m = 1, \cdots, n+1$ the m^{th} degree polynomial $p_m(x)$ defined above m simple roots that strictly interlace with the $m-1$ simple roots of $p_{m-1}(x)$.

It is clear that $p_0(x) = 1$ has no roots while $p_1(x) = (x - a_0)$ has one simple root. The interlacing is therefore trivially true. The assertion is thereby holds for $m = 1$.

We now suppose the assertion holds for m. If we denote the roots of $p_m(x)$ by

$$x_1^{(m)} < x_2^{(m)} < \cdots < x_{m-1}^{(m)} < x_m^{(m)}\,,$$

and the roots of $p_{m-1}(x)$ by

$$x_1^{(m-1)} < x_2^{(m-1)} < \cdots < x_{m-2}^{(m-1)} < x_{m-1}^{(m-1)}\,.$$

then the fact these strictly interlace means

(4)
$$\begin{aligned} &x_1^{(m)} < x_1^{(m-1)} < x_2^{(m)} < x_2^{(m-1)} < x_3^{(m)} < \cdots \\ &\qquad\qquad \cdots < x_{m-2}^{(m-1)} < x_{m-1}^{(m)} < x_{m-1}^{(m-1)} < x_m^{(m)} \,. \end{aligned}$$

When this fact is combined with the fact that $p_{m-1}(x) \sim x^{m-1}$ as $|x| \to \infty$ then one obtains

(5)
$$\operatorname{sign}\big(p_{m-1}\big(x_k^{(m)}\big)\big) = (-1)^{m-k} \quad \text{for every } k = 1, \cdots, m \,.$$

We now use this fact to do a sign analysis of $p_{m+1}(x)$.

The defining relation of $p_{m+1}(x)$ along with the fact $x_k^{(m)}$ is a root of $p_m(x)$ yields

$$p_{m+1}\big(x_k^{(m)}\big) = -b_m^2 p_{m-1}\big(x_k^{(m)}\big) \,.$$

Because A is irreducible, the previous problem shows that $b_m \neq 0$. This fact combined with the above relation implies

$$\operatorname{sign}\big(p_{m+1}\big(x_k^{(m)}\big)\big) = (-1)^{m-k+1} \quad \text{for every } k = 1, \cdots, m \,.$$

This sign analysis along with the fact that $p_{m+1}(x) \sim x^{m+1}$ as $|x| \to \infty$, shows that $p_{m+1}(x)$ must have at least one root in each of the $n+1$ intervals

$$\big(-\infty, x_1^{(m)}\big) \,, \quad \big(x_1^{(m)}, x_2^{(m)}\big) \,, \quad \cdots \,, \quad \big(x_{m-1}^{(m)}, x_m^{(m)}\big) \,, \quad \big(x_m^{(m)}, \infty\big) \,.$$

Therefore $p_{m+1}(x)$ is an $(m+1)^{th}$ degree polynomial with $m+1$ simple roots that interlace with the m simple roots of $p_m(x)$. $\square$

Solution of Part (b). To show that $V^{-1}AV$ is diagonal, first observe that the recursion relations defining the polynomials $p_m(x)$ can be expressed as

(6)
$$\begin{pmatrix} x-a_0 & -1 & 0 & \cdots & 0 \\ -b_1^2 & x-a_1 & -1 & \ddots & \vdots \\ 0 & -b_2^2 & x-a_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -b_n^2 & x-a_n \end{pmatrix} \begin{pmatrix} p_0(x) \\ p_1(x) \\ p_2(x) \\ \cdots \\ p_n(x) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ p_{n+1}(x) \end{pmatrix} .$$

Let Π be the diagonal matrix defined by

$$\Pi = \begin{pmatrix} \pi_0 & 0 & \cdots & 0 \\ 0 & \pi_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \pi_n \end{pmatrix} .$$

Observe that

$$A = \Pi^{-1} \begin{pmatrix} a_0 & 1 & 0 & \cdots & 0 \\ b_1^2 & a_1 & 1 & \ddots & \vdots \\ 0 & b_2^2 & a_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & b_n^2 & a_n \end{pmatrix} \Pi \,, \qquad \begin{pmatrix} p_0(x) \\ p_1(x) \\ p_2(x) \\ \cdots \\ p_n(x) \end{pmatrix} = \Pi \begin{pmatrix} q_0(x) \\ q_1(x) \\ q_2(x) \\ \cdots \\ q_n(x) \end{pmatrix} .$$

Hence, equation (6) can therefore be expressed as

$$A \begin{pmatrix} q_0(x) \\ q_1(x) \\ q_2(x) \\ \cdots \\ q_n(x) \end{pmatrix} = x \begin{pmatrix} q_0(x) \\ q_1(x) \\ q_2(x) \\ \cdots \\ q_n(x) \end{pmatrix} - \frac{1}{\pi_n} \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ p_{n+1}(x) \end{pmatrix} .$$

Now let $\{x_k\}_{k=0}^{n+1}$ be the $n+1$ simple roots of $p_{n+1}(x)$ established by Part (a). The above relation then shows that $AV = V\Lambda$ where Λ is the diagonal matrix

$$\Lambda = \begin{pmatrix} x_0 & 0 & \cdots & 0 \\ 0 & x_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & x_n \end{pmatrix} .$$

The result will therefore follow upon showing that V is invertible.

Suppose V is not invertible. Then there exists a nonzero vector w such that $w^T V = 0$. Let $w = (w_0, w_1, \cdots, w_n)^T$. Then

$$0 = w^T V = \begin{pmatrix} q(x_0) & q(x_1) & \cdots & q(x_n) \end{pmatrix} ,$$

where $q(x)$ is the polynomial defined by

$$q(x) = \sum_{m=0}^{n} w_m q_m(x) \,.$$

Because $q(x)$ is a polynomial of degree n or less that vanishes at $n+1$ points, it must be identically zero. But because each $p_m(x)$ is a monic polynomial of degree m, the polynomials $\{p_m(x)\}$ are linear independent. Because $q_m(x) = p_m(x)/\pi_m$, the polynomials $\{q_m(x)\}$ are also linear independent. It follows that $w_m = 0$ for every $m = 0, \cdots, n$. But this contradicts the fact that w is nonzero. Therefore V is invertible. $\square$

(7) Given any self-adjoint matrix $A \in \mathbb{R}^{N\times N}$ and any unit vector $u \in \mathbb{R}^N$, use the Lanczos algorithm to construct an orthogonal matrix Q such that the first column of Q is u and that $Q^T A Q$ is tridiagonal.

Solution. The Lanczos algorithm constructs a sequence of vectors $p^{(n)}$ as

$$
\begin{aligned}
p^{(1)} &= Ap^{(0)} - \kappa_0 p^{(0)} \,, \\
p^{(n+1)} &= Ap^{(n)} - \kappa_n p^{(n)} - \mu_n p^{(n-1)} \quad \text{for } n = 1, 2, \cdots \,,
\end{aligned}
\tag{7}
$$

where the coefficients κ_n and μ_n are given by

$$
\kappa_n = \frac{\big(p^{(n)} \,|\, Ap^{(n)}\big)}{\big(p^{(n)} \,|\, p^{(n)}\big)} \qquad \text{for } n = 0, 1, \cdots \,, \tag{8}
$$

$$
\mu_n = \frac{\big(p^{(n)} \,|\, p^{(n)}\big)}{\big(p^{(n-1)} \,|\, p^{(n-1)}\big)} \qquad \text{for } n = 1, 2 \cdots \,. \tag{9}
$$

The algorithm halts as soon as $p^{(n)} = 0$ for some n. The vectors $p^{(n)}$ satisfy the orthogonality relation

$$
\big(p^{(m)} \,|\, p^{(n)}\big) = 0 \quad \text{for every } m < n \,. \tag{10}
$$

It folows from (9) that

$$
\big\|p^{(n)}\big\|^2 = \pi_n \big\|p^{(0)}\big\|^2 \,,
$$

where $\pi_0 = 1$ and $\pi_n = \mu_n \pi_{n-1}$, which will be positive until $p^{(n)} = 0$ for some n.

Now apply the Lanczos algorithm with $p^{(0)} = u$ to construct $p^{(n)}$ until $p^{(n_1)} = 0$. For $n = 0, \cdots, n_1 - 1$ set

$$
u^{(n)} = \frac{1}{\sqrt{\pi_n}}\, p^{(n)} \,.
$$

If $n_1 = N + 1$ then you are done. Otherwise let $u^{(n_1)}$ be any unit vector that is orthogonal to $\big\{u^{(0)}, \cdots, u^{(n_1-1)}\big\}$ and apply the Lanczos algorithm with $p^{(0)} = u^(n_1)$ to construct $p^{(n)}$ until $p^{(n_2)} = 0$. For $n = 0, \cdots, n_2 - 1$ set

$$
u^{(n_1+n)} = \frac{1}{\sqrt{\pi_n}}\, p^{(n)} \,.
$$

Repeat this until $n_1 + \cdots + n_m = N + 1$.

(8) Recall that $A \in \mathbb{C}^{N\times N}$ is called *normal* whenever $A^*A = AA^*$. Show that A is normal and invertible if and only if there exists a unitary matrix U and a self-adjoint, positive definite matrix P such that $A = UP = PU$.

Solution. First suppose there exists a unitary matrix U and a self-adjoint, positive definite matrix P such that $A = UP = PU$. Because both U and P are invertible, it follows that $A = UP = PU$ is also invertible. Moreover, because

$$A^*A = (UP)^*UP = PU^*UP = P^2\,,$$

while

$$AA^* = PU(PU)^* = PUU^*P = P^2\,,$$

it follows that $A^*A = P^2 = AA^*$. Therefore A is normal and invertible.

Now suppose A is normal and invertible. Because A is normal there exists a unitary matrix $V \in \mathbb{C}^{N\times N}$ and a diagonal matrix Λ such that $A = V\Lambda V^*$. Then

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_N \end{pmatrix}\,,$$

where the λ_j are the eigenvalues of A. Because A is invertible every λ_j is nonzero.

Let $|\Lambda|$ and Σ be the diagonal matrices given by

$$|\Lambda| = \begin{pmatrix} |\lambda_1| & 0 & \cdots & 0 \\ 0 & |\lambda_2| & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & |\lambda_N| \end{pmatrix}\,, \qquad \Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_N \end{pmatrix}\,,$$

where $\sigma_j = \lambda_j/|\lambda_j|$ for $j = 1, \cdots, N$. Clearly, $|\Lambda|$ is self-adjoint and positive definite, Σ is unitary, while

$$\Lambda = \Sigma|\Lambda| = |\Lambda|\Sigma\,.$$

Let $P = V|\Lambda|V^*$ and $U = V\Sigma V^*$. Clearly, P is self-adjoint and positive definite and U is unitary. Moreover,

$$PU = V|\Lambda|V^*V\Sigma V^* = V\Lambda V^* = A\,,$$

and

$$UP = V\Sigma V^*V|\Lambda|V^* = V\Lambda V^* = A\,,$$

so that $A = UP = PU$. $\square$

(9) Let $A \in \mathbb{R}^{N\times N}$ be normal and invertible. Let $\{A_n\}_{n=0}^{\infty}$ be the sequence of $N \times N$ matrices constructed recursively by the QR-Method: $A_0 = A$, $A_n = Q_n R_n$, and $A_{n+1} = R_n Q_n$, where every Q_n is orthogonal and every R_n is upper triangular with positive diagonal entries. Show that every A_n is normal. (Hint: The result of the previous problem might be helpful.)

Solution. We will prove that every A_n is normal and invertible by induction on n. Because $A_0 = A$, the assertion holds for $n = 0$ by hypothesis. Now suppose the assertion holds for n. We will show it holds for $n + 1$.

Because $A_n \in \mathbb{R}^{N\times N}$ is invertible there exists unique matrices Q_n and R_n such that Q_n is orthogonal, R_n is upper triangular with positive diagonal entries, and

$$A_n = Q_n R_n \,.$$

Because $A_n \in \mathbb{R}^{N\times N}$ is normal and invertible, by applying the result of the previous problem to the real setting, there exists an orthogonal matrix U_n and a symmetric, positive definite matrix P_n such that

$$U_n P_n = P_n U_n = A_n = Q_n R_n \,.$$

Because $Q_n^{-1} = Q_n^T$, the above relations lead to the formulas

$$R_n = Q^T U_n P_n \,, \quad \text{and} \quad R_n = Q^T P_n U_n \,.$$

Because $A_{n+1} = R_n Q_n$, the above formulas show that

$$\begin{aligned} A_{n+1} &= R_n Q_n = Q_n^T P_n U_n Q_n = Q_n^T P_n Q_n Q_n^T U_n Q_n \\ &= \big(Q_n^T P_n Q_n\big)\big(Q_n^T U_n Q_n\big) = P_{n+1} U_{n+1} \,, \\ A_{n+1} &= R_n Q_n = Q_n^T U_n P_n Q_n = Q_n^T U_n Q_n Q_n^T P_n Q_n \\ &= \big(Q_n^T U_n Q_n\big)\big(Q_n^T P_n Q_n\big) = U_{n+1} P_{n+1} \,. \end{aligned}$$

where we have defined $U_{n+1} = Q_n^T U_n Q_n$ and $P_{n+1} = Q_n^T P_n Q_n$. It is clear that U_{n+1} is orthogonal and that P_{n+1} is symmetric and positive definite. Because the above calculation shows that $A_{n+1} = U_{n+1} P_{n+1} = P_{n+1} U_{n+1}$, the result of the previous problem implies that A_{n+1} is normal and invertible. $\square$

Remark. This result is one of the steps in the proof that the QR-method converges when A is normal.

(10) Let $H_0 \in \mathbb{R}^{N\times N}$ and $H(t)$ satisfy the isospectral flow initial-value problem

$$\frac{\mathrm{d}H}{\mathrm{d}t} = JH - HJ\,, \qquad H(0) = H_0\,,$$

where $J(t) \in \mathbb{R}^{N\times N}$ such that $J(t)^T = -J(t)$ for every $t \in \mathbb{R}$. Show that if H_0 is normal then so is $H(t)$ for every $t \in \mathbb{R}$.

Solution. We will show that $H(t)^T H(t) = H(t)H(t)^T$ for every $t \in \mathbb{R}$. By taking the transpose of the isospectral flow initial-value problem one sees that $H(t)^T$ satisfies

$$\begin{aligned}\frac{\mathrm{d}H^T}{\mathrm{d}t} &= \big(JH - HJ\big)^T = H^T J^T - J^T H^T \\ &= JH^T - H^T J\,, \qquad H(0)^T = H_0^T\,.\end{aligned}$$

Upon combining the initial-value problems governing $H(t)$ and $H(t)^T$ one sees that $H(t)^T H(t)$ is governed by

$$\begin{aligned}\frac{\mathrm{d}H^T H}{\mathrm{d}t} &= \frac{\mathrm{d}H^T}{\mathrm{d}t}H + H^T\frac{\mathrm{d}H}{\mathrm{d}t} \\ &= \big(JH^T - H^T J\big)H + H^T(JH - HJ) \\ &= JH^T H - H^T JH + H^T JH - H^T HJ \\ &= JH^T H - H^T HJ\,, \qquad H(0)^T H(0) = H_0^T H_0\,.\end{aligned}$$

Similarly, one sees that $H(t)H(t)^T$ is governed by

$$\begin{aligned}\frac{\mathrm{d}HH^T}{\mathrm{d}t} &= \frac{\mathrm{d}H}{\mathrm{d}t}H^T + H\frac{\mathrm{d}H^T}{\mathrm{d}t} \\ &= (JH - HJ)H^T + H\big(JH^T - H^T J\big) \\ &= JHH^T - HJH^T + HJH^T - HH^T J \\ &= JHH^T - HH^T J\,, \qquad H(0)H(0)^T = H_0 H_0^T\,.\end{aligned}$$

Because H_0 is normal one knows that $H_0^T H_0 = H_0 H_0^T$, whereby $H(t)^T H(t)$ and $H(t)H(t)^T$ are governed by the same initial-value problem. Because the initial-value problem has a unique solution, it follows that $H(t)^T H(t) = H(t)H(t)^T$ for every $t \in \mathbb{R}$. Hence, $H(t)$ is normal for every $t \in \mathbb{R}$. $\square$