Remember to write your name! You are allowed to use a calculator. **You are not allowed to use the textbook or your notes or your neighbor**. To receive full credit on a problem you must show **sufficient justification for your conclusion** unless explicitly stated otherwise. You may cite any theorem from Atkinson or from the lectures unless explicitly stated otherwise.

You must do the first problem. You must pick **only** two of the remaining problems. Each problem is 15 points; there are 45 points total.

---

1. Quadrature

(a) Let $p_*(x)$ be the minimax approximation to f of degree at most n, then the quadrature integrates this exactly. The quadrature error is thus

$$|I[f] - I_n[f]| = |I[f] - I_n[f] - I_n[p_*] + I_n[p_*]| = |I[f] - I_n[f] - I[p_*] + I_n[p_*]| \le +|I_n[f - p*]|$$

Bound the terms separately:

$$|I[f - p_*]| = |\int_a^b f(x) - p_*(x)\mathrm{d}x| \le \int_a^b |f(x) - p_*(x)|\mathrm{d}x \le \rho_n(f)(b-a)$$

$$|I_n[f - p_*]| = |\sum_i w_{i,n}(f(x_{i,n}) - p_*(x_{i,n}))| \le \sum_i |w_{i,n}||f(x_{i,n}) - p_*(x_{i,n})|$$

$$\le \rho_n(f)\sum_i |w_{i,n}| = \rho_n(f)\sum_i w_{i,n} = \rho_n(f)(b-a)$$

where $\rho_n(f)$ is the minimax error and the last few equalities follow from the fact that $w_{i,n} \ge 0$ and the quadrature integrates $f(x) = 1$ exactly. Since $\rho_n \to 0$ for continuous functions f, we have convergence.

(b) The error is

$$|\int_a^b f(x)\mathrm{d}x - \sum_i w_{i,n}(f(x_i) + \epsilon_{i,n})| \le |I[f] - I_n[f]| + |\sum_i w_{i,n}\epsilon_{i,n}| \le |I[f] - I_n[f]| + \sum_i w_{i,n}|\epsilon_{i,n}|$$

$$\le |I[f] - I_n[f]| + \epsilon\sum_i w_{i,n} = |I[f] - I_n[f]| + \epsilon(b-a).$$

The first term $\to 0$ as $n \to \infty$ by (a), which gives the desired result.

Note: there is no guarantee that the errors $\epsilon_{i,n}$ are values of some integrable function $g(x_{i,n}) = \epsilon_{i,n}$, so answers based on that assumption are wrong (but received some partial credit if otherwise correct).

(c) The error is

$$|\int_a^b f(x)\mathrm{d}x - \frac{h}{2}\sum_i (f(x_i) + f(x_{i+1}) + \epsilon_{i,n} + \epsilon_{i+1,n})| \le |I[f] - T_n[f]| + \frac{h}{2}|\sum_i \epsilon_{i,n} + \epsilon_{i+1,n}| \le$$

$$|I[f] - T_n[f]| + \frac{h}{2}\sum_i 2\epsilon = |I[f] - T_n[f]| + \epsilon(b-a)$$

where $T_n[f]$ is the trapezoid rule. The first term $\to 0$ as $n \to \infty$ since the trapezoid rule converges for twice continuously differentiably functions, leaving the desired result.

2. Linear Systems

(a) Let $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ be the system of equations (square), and let $\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}$ be the diagonal, lower-triangular, and upper-triangular parts of $\mathbf{A}$. The Jacobi iteration has the form

$$\boldsymbol{x}_{k+1} = \mathbf{D}^{-1}\boldsymbol{b} - \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\boldsymbol{x}_k.$$

The error equation is

$$\boldsymbol{e}_{k+1} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\boldsymbol{e}_k.$$

If any norm of the iteration matrix is less than 1 then the iteration converges. Now consider the infinity norm, which is the max absolute row sum. The infinity norm of the iteration matrix is

$$\max_i \sum_{j \neq i} \frac{|a_{i,j}|}{|a_{i,i}|}.$$

The fact that the linear system is strictly diagonally dominant implies that the above is less than 1, so the error converges to 0.

(b) Consider that when you change the order of the inner loop, you are simultaneously permuting the rows of the system *and* the columns, i.e. you are applying Gauss-Seidel to the system

$$\mathbf{PAP}^T\mathbf{P}\boldsymbol{x} = \mathbf{PAP}^T\boldsymbol{y} = \mathbf{P}\boldsymbol{b}$$

where $\mathbf{P}$ is the permutation matrix that re-orders the rows. The new system is still symmetric positive definite, so the Gauss-Seidel iteration converges to a re-ordered version of the solution $\boldsymbol{y} = \mathbf{P}\boldsymbol{x}$.

To be more precise, when you reorder the rows you multiply the system from the left by the permutation matrix $\mathbf{P}$. This yields the system (e.g.)

$$a_{n,1}x_1 + a_{n,2}x_2 + \ldots + a_{n,n}x_n = b_n$$

$$a_{n-1,1}x_1 + a_{n-1,2}x_2 + \ldots + a_{n-1,n}x_n = b_{n-1}$$

$$\vdots$$

$$a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n = b_1.$$

The first step of the Gauss-Seidel loop would modify x_n which is not on the diagonal, so we should re-order the columns

$$a_{n,n}x_n + a_{n,n-1}x_{n-1} + \ldots + a_{n,1}x_1 = b_n$$

$$a_{n-1,n}x_n + a_{n-1,n-1}x_{n-1} + \ldots + a_{n-1,1}x_1 = b_{n-1}$$

$$\vdots$$

$$a_{1,n}x_n + a_{1,n-1}x_{n-1} + \ldots + a_{1,1}x_1 = b_1.$$

Now we should re-label the unknowns so that the first equation corresponds to the first unknown, e.g. $y_1 = x_n$, $y_2 = x_{n-1}$, etc. So permuting the order is the same as applying Gauss-Seidel to the equation $\mathbf{PAP}^T\boldsymbol{y} = \mathbf{P}\boldsymbol{b}$ where $\boldsymbol{y} = \mathbf{P}\boldsymbol{x}$. The system is still SPD so GS still converges.

3. Rootfinding/Nonlinear Equations

(a) Any solution is a solution of the fixed-point problem $x = h(x) = y + \delta_t f(x)$. The iteration function $h(x)$ is globally Lipshitz with constant $2\delta_t$

$$|h(x_0) - h(x_1)| \leq 2\delta_t |x_0 - x_1| \, \forall \, x_0, x_1 \in \mathbb{R}.$$

If we take $\delta_t < 1/2$ then the map is a contraction on $\mathbb{R}$, which means it must have a unique fixed point. Suppose there are 2 fixed points $\alpha_0 = h(\alpha_0)$ and $\alpha_1 = h(\alpha_1)$. Then we must have

$$|h(\alpha_0) - h(\alpha_1)| = |\alpha_0 - \alpha_1| \leq 2\delta_t |\alpha_0 - \alpha_1|.$$

If we choose $\delta_t < 1/2$ then it is not possible to have $\alpha_0 \neq \alpha_1$, i.e. there is a unique solution.

Atkinson doesn't state the theorem for $\mathbb{R}$, just for finite intervals, so grading will be lenient for existence: show that the map is a contraction for $\delta_t < 1/2$ and state that this implies existence and uniqueness.

(b) Notice that since $f(\alpha) = 0$, α must be a fixed point of

$$x_{k+1} = x_k + \delta_t f(x_{k+1})$$

. In part (a) we showed that this equation implicitly defines an iteration $x_{k+1} = g(x_k)$, and we now assume that g is smooth. We can obtain conditions for convergence by examining $g'(\alpha)$. Plugging in our notation

$$g(x_k) = x_k + \delta_t f(g(x_k)).$$

Take the derivative, then evaluate at α:

$$g'(\alpha) = 1 + \delta_t f'(g(\alpha))g'(\alpha) = 1 + \delta_t f'(\alpha)g'(\alpha).$$

$$g'(\alpha) = \frac{1}{1 - \delta_t f'(\alpha)}$$

Theorem 2.7 from Atkinson guarantees that the iteration will converge to α for 'close enough' initial conditions provided that $|g'(\alpha)| < 1$, i.e.

$$f'(\alpha) < 0 \text{ or } f'(\alpha) > \frac{2}{\delta_t}$$

(it is assumed that $\delta_t > 0$.)

Part (b) is the backwards Euler iteration for the ODE $x'(t) = f(x)$. We have shown that if α is a stable equilibrium ($f'(\alpha) < 0$) then the backwards Euler iteration will converge to it for any stepsize δ_t as long as the initial condition is close enough, i.e. backwards Euler behaves qualitatively like the true system for any δ_t. Conversely, if α is an unstable equilibrium $f'(\alpha) > 0$, then backwards Euler will still converge to the equilibrium if δ_t is too large, i.e. if δ_t is too large then backwards Euler has the exact opposite behavior from the true system.

4. Interpolation

(a) The dimension of the space of cubic splines with $n+1$ nodes is $n+3$.

(b) The following are all acceptable

- 'Natural' splines set the second derivative to 0 at the endpoints
- 'Not-a-knot' splines make the third derivative continuous at the nodes just inside each boundary
- 'Complete' cubic splines have the same first derivative as the function at the endpoints
- If the function is periodic then the spline can also be forced to be periodic

(c) Let $\phi_k(x)$, $k = 1, \ldots, n+3$ be any basis, e.g. the monomials plus truncated power functions. Seek to expand the desired basis $\varphi_j(x)$ in the available basis $\phi_k(x)$

$$\varphi_j(x) = \sum_i c_{j,k}\phi_k(x).$$

Now impose the cardinality conditions on $\varphi_j(x)$ for $j \leq n+1$

$$\varphi_j(x_i) = \sum_i c_{j,k}\phi_k(x_i) = \delta_{i,j}, i = 0, \ldots, n$$

This linear system has $n+1$ equations and $n+3$ unknowns $c_{j,k}$ where $j = 1, \ldots, n+3$. It is also a spline interpolation problem: find the spline function $\varphi_j(x)$ that interpolates the data $f_j(x_i) = \delta_{i,j}$. We know that a particular solution exists, since any of the answers from part (b) will yield a solution. To be specific, let $\varphi_j(x)$ be natural splines.

Up to this point we've shown the existence of $\varphi_j(x)$ for $j \leq n+1$ that satisfy the cardinality conditions. These functions are certainly linearly independent since $\varphi_j(x_i) = \delta_{i,j}$. It remains to complete the basis by finding $\varphi_j(x)$ for $j > n+1$. Let $\hat{\varphi}(x)$ be the *complete* spline that solves the interpolation problem $\hat{\varphi}(x_i) = \delta_{i,0}$ with Hermite data $\hat{\varphi}'(x_0) = \varphi'_{n+1}(x_0) - 1$, $\hat{\varphi}'(x_n) = \varphi'_1(x_n) - 1$, and let $\tilde{\varphi}(x)$ be the *complete* spline that solves the interpolation problem $\tilde{\varphi}(x_i) = \delta_{i,0}$ with Hermite data $\tilde{\varphi}'(x_0) = \varphi'_1(x_0) + 1$, $\tilde{\varphi}'(x_n) = \varphi'_1(x_n) + 1$. Define

$$\varphi_{n+1}(x) = \varphi_1(x) - \hat{\varphi}(x), \quad \varphi_{n+2}(x) = \varphi_1(x) - \tilde{\varphi}(x).$$

By construction these satisfy the remaining cardinality condition $\varphi_j(x_i) = 0$ for $j > n+1$. They are also independent of each other, since $\varphi'_{n+1}(x_0) = 1$ while $\varphi'_{n+2}(x_0) = -1$. This is clearly just one way to construct such a cardinal spline basis.

Several people mentioned using the Lagrange basis to construct the spline basis. In general this won't work because the Lagrange polynomials have degree n, which is too high unless $n = 3$.

5. Approximation The weighted norm of the error is

$$\int_0^1 xw(x)(f(x)-p(x))^2\mathrm{d}x = \int_0^1 xw(x)(f(x))^2\mathrm{d}x - 2\int_0^1 xw(x)f(x)p(x)\mathrm{d}x + \int_0^1 xw(x)(p(x))^2\mathrm{d}x.$$

Expand $p(x)$ in the basis and insert into the expression above

$$p(x) = \sum_i c_i\phi_i(x) \Rightarrow$$

$$\int_0^1 xw(x)(f(x)-p(x))^2\mathrm{d}x =$$

$$\int_0^1 xw(x)(f(x))^2\mathrm{d}x - 2\sum_i c_i\int_0^1 xw(x)f(x)\phi_i(x)\mathrm{d}x + \sum_i\sum_j c_ic_j\int_0^1 xw(x)\phi_i(x)\phi_j(x)\mathrm{d}x..$$

We can write this as

$$\int_0^1 xw(x)(f(x)-p(x))^2\mathrm{d}x = d - 2\boldsymbol{c}^T\boldsymbol{f} + \boldsymbol{c}^T\mathbf{A}\boldsymbol{c}$$

where $\boldsymbol{c}$ is a vector with entries c_i, where d is a constant, and where the matrix $\mathbf{A}$ is symmetric positive definite with entries

$$(\mathbf{A})_{i,j} = \int_0^1 xw(x)\phi_i(x)\phi_j(x)\mathrm{d}x.$$

We know that the ϕ_k form an orthogonal basis of increasing degree, so they must satisfy a three-term recurrence, i.e.

$$x\phi_i(x) = a_i\phi_{i+1}(x) + b_i\phi_i(x) + c_i\phi_{i-1}(x).$$

This shows that the matrix $\mathbf{A}$ is tridiagonal since

$$(\mathbf{A})_{i,j} = \int_0^1 w(x)(a_i\phi_{i+1}(x) + b_i\phi_i(x) + c_i\phi_{i-1}(x))\phi_j(x)\mathrm{d}x = a_i\delta_{i+1,j} + b_i\delta_{i,j} + c_i\delta_{i-1,j}.$$

(The above expression also assumes that the ϕ_k are orthonormal, without loss of generality.) The unique minimizer of the quadratic error is obtained by solving for the critical point

$$\mathbf{A}\boldsymbol{c} = \boldsymbol{f}$$

which can be accomplished using Gaussian Elimination in $\mathcal{O}(n)$ operations.