EXAM #

CSEM Prelim Exam
Area B
May 29, 2018
Part I – Num Anly: Linear Algebra
(CSE 383C)

Please explain your answers.

1. **30 points.** Let $A \in \mathbb{R}^{3\times 2}$. We wish to determine A.

   Let $\sigma_{\max} = 2$ be the largest singular value of A. Let $\sigma_{\min} = 1$ be the smallest singular value of A. Let $A^T e_3 = 0$, where e_3 is the third column of the 3-by-3 identity matrix. Let $Ae_2 = \sigma_{\max} u$, where e_2 is the first column of the 2-by-2 identity matrix, and u is a column vector with entries $[1/\sqrt{2}, 1\sqrt{2}, 0]$.

   The information above is not quite sufficient for determining A.

   (a) Explain why.

   (b) Introduce the *minimum* number of additional assumptions that are required to be able to fully determine all the entries of A and then state A (either in full or factored form).

2. **70 points.** Let $A \in \mathbb{R}^{m\times m}$ be symmetric *semi-positive* definite matrix (i.e., $x^T Ax \geq 0$ for all x such that $\|x\| > 0$). We seek stable algorithms for solving $Ax = b$.

   (a) What will happen if apply the Cholesky factorization algorithm on A?

   (b) Suppose V is a basis for the null space of A. Assume that $b \in \text{Range}(A)$. Using V and the Cholesky factorization, give an algorithm for solving $Ax = b$.

   (c) What is the complexity of your algorithm (number of floating point operations) as a function of m and the dimension of the null space of A?

   (d) Assume now that $b \notin \text{Range}(A)$. Can you use your algorithm to solve $\min_x \|Ax-b\|_2^2$?

   (e) What is the stability bound $\|\delta x\|/\|x\|$ of your algorithm, where δx minimizes $\|A(x+\delta x) - (b+\delta b)\|_2^2$?g Here b and $\delta b \notin \text{Range}(A)$ and x minimizes $\|Ax - b\|_2^2$.

   (f) What is its stability of your algorithm for the above case in the presence of round-off errors?

   (g) Assume A is sparse. Suggest an alternative iterative algorithm (that exploits the sparsity of A) for solving $Ax = b$. State its complexity. Under what conditions is this new algorithm faster than your original Cholesky factorization algorithm?

Part II – Num Anly: Diff Equations
(CSE 383L)

3. Consider the following ordinary differential equation, two-point boundary value problem.

$$y''(x) = f(y'(x), y(x)) \qquad (1)$$

(a) With the initial conditions $y(0)=1.\ y'(0)=2$ rewrite as first order system and approximate by implicit Euler. What is the local truncation error and is the algorithm 0-stable (Dahlquist stable)?

(b) With the boundary conditions $y(0)=1.\ y'(1)=2$ rewrite as first order system and approximate by trapezoidal rule. Describe briefly how Richardson extrapolation can be used to improve the order of accuracy

(c) Rewrite the problem (1) on variational form when f is linear and with the boundary conditions of (b). Determine a finite element algorithm for its approximation.

4. Consider the following elliptic PDE,

$$\begin{aligned}
&-\nabla\cdot a(x,y)\nabla u + b\cdot\nabla u + cu = f(x,y), \quad 0<x<1,\ 0<y<1,\ 0<a\le a(x,y)\le A, c\ge 0,\\
&u = g(x,y),\ x=0 \text{ and } x=1, 0<y<1,\\
&u_y + su = 0, y=0 \text{ and } y=1, 0<x<1,
\end{aligned}$$

(a) Rewrite the equation on weak form.

(b) State and prove the fundamental error estimate for the finite element method that the error measured in a suitable norm between the PDE and FEM solutions are modulo a constant bounded by the error between the PDE solution and any function in a suitable space.

(c) Show that the relevant bilinear form in (a) for $c=g=b=s=0$ is continuous and coersive and that the relevant linear form is continuous.

5. A nonlinear scalar, viscous conservation law with has the form,

$$u(x,t)_t + f(u(x,t))_x = bu(x,t)_{xx}, \quad b>0$$

(a) Use von Neumann analysis when $f(u)=au$ to set up the conditions, which determine necessary and sufficient conditions for the spatial and temporal step sizes to guarantee L_2 stability for explicit Euler in time and central 2^{nd} order differencing in space.

(b) Devise an upwind finite difference method for the equation above when $f'(u) > 0, b = 0$ and show that the method is consistent and on conservation form.

(c) Give conditions on the spatial and temporal step sizes to guarantee the scheme in (b) is monotone.

Part II – Stat/Disc Meth Sci Comp
(CSE 383M)

3. [25 points]

Consider points on the real line $\mathbb{R}$. Let $\mathcal{H}$ be the family of all unit-length closed intervals in $\mathbb{R}$. (In other words, $\mathcal{H}$ consists of all intervals in the form $[a, a+1]$ where $a \in \mathbb{R}$.) The *growth function* of $\mathcal{H}$ is defined as

$$H[m] = \max_{|S|=m} |H[S]|, \qquad \text{where } H[S] = \{h \cap S \colon h \in \mathcal{H}\}.$$

A set S is *shattered* by $\mathcal{H}$ if $|\mathcal{H}[S]| = 2^{|S|}$. The *VC-dimension* of $\mathcal{H}$ is defined as $D = \max\{m \colon \mathcal{H}[m] = 2^m\}$.

(a) Find the VC-dimension D of $\mathcal{H}$ and justify your answer.

(b) Are all point sets of $\mathbb{R}$ of size D shattered by $\mathcal{H}$? Explain your answer.

(c) Show that $\mathcal{H}[50] \geq 100$.

(a) $D = 2$.

To show that $D \geq 2$, it suffices to show that the set $S = \{1, 2\}$ can be shattered.

$$\begin{aligned} [-1, 0] \cap S &= \emptyset \\ [0, 1] \cap S &= \{1\} \\ [1, 2] \cap S &= \{2\} \\ [2, 3] \cap S &= \{1, 2\} \end{aligned}$$

To show that $D < 3$, consider $S = \{x_1, x_2, x_3\}$. We may assume that $x_1 < x_2 < x_3$. If $x_1, x_3 \in [a, a+1]$, then $x_2 \in [a, a+1]$. Therefore, $\{x_1, x_3\} \notin \mathcal{H}[S]$.

(b) No. The point set $S = \{0, 2\}$ is not shattered of $\mathcal{H}$.

If $0 \in [a, a+1]$, then $a \leq 0$. So, $2 \notin [a, a+1]$. Therefore, $\{0, 1\} \notin \mathcal{H}[S]$.

(c) Consider $S = \{1, 2, \ldots, 50\}$.

$$\begin{aligned} [-1, 0] \cap S &= \emptyset \\ [i - \tfrac{1}{2}, i + \tfrac{1}{2}] \cap S &= \{i\} \qquad \text{for all } 1 \leq i \leq 50 \\ [i, i+1] \cap S &= \{i, i+1\} \qquad \text{for all } 1 \leq i \leq 49 \end{aligned}$$

Therefore, $\mathcal{H}[m] \geq |\mathcal{H}[S]| \geq 1 + 50 + 49 = 100$.

Area B
May 2018
Solutions

4. [25 points]

Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^d$ be data points on the unit sphere, with labels $\ell_1, \ldots, \ell_n \in \{\pm 1\}$. Define the maximum margin as

$$\gamma_{\max} = \max_{\substack{\mathbf{v}\in\mathbb{R}^d \\ \|\mathbf{v}\|_2=1}} \min_{1\le i\le n} (\mathbf{v}^\mathsf{T}\mathbf{a}_i)\ell_i.$$

Flip each label independently and randomly with probability 0.1, such that

$$\ell_i' = \begin{cases} -\ell_i & \text{with probability } 0.1 \\ \ell_i & \text{with probability } 0.9 \end{cases}.$$

Using these noisy labels and parameter $C > 0$, the support vector machine with hinge loss computes $\mathbf{w} \in \mathbb{R}^d$ and $\xi_1, \xi_2, \ldots, \xi_n$ that

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & (\mathbf{w}^\mathsf{T}\mathbf{a}_i)\,\ell_i' \ge 1 - \xi_i \qquad \forall 1 \le i \le n \\ \text{and} \quad & \xi_i \ge 0 \qquad \forall 1 \le i \le n. \end{aligned}$$

Suppose $\gamma_{\max} \ge 0.5$. Show that

$$\Pr\left[\|\mathbf{w}\|_2^2 \ge \gamma_{\max}^{-2} + 0.9Cn\right] \le O\left(\frac{1}{n}\right).$$

Define

$$X_i = \begin{cases} 1 & \text{if } \ell' = -\ell \\ 0 & \text{if } \ell' = \ell \end{cases}.$$

Then, X_i is Bernoulli random variable with paramter 0.1. By Chebyshev Inequality,

$$\Pr\left[\sum_{i=1}^n X_i \ge 0.1n + 0.1n\right] \le \frac{0.09n}{(0.1n)^2} \le O\left(\frac{1}{n}\right).$$

So, it suffices to show that $\sum_{i=1}^n X_i \le 0.2n$ implies $\|\mathbf{w}\|_2^2 \le \gamma_{\max}^{-2} + 0.9Cn$.

Let $\mathbf{v}_{\max} \in \mathbb{R}^d$ such that $\|\mathbf{v}_{\max}\|_2 = 1$ and

$$\min_{1\le i\le n} (\mathbf{v}_{\max}^\mathsf{T}\mathbf{a}_i)\ell_i = \gamma_{\max}.$$

Let $\mathbf{w}' = \gamma_{\max}^{-1}\mathbf{v}_{\max}$. Let $\xi_i' = 3X_i$. If $\ell_i = \ell_i'$, then

$$((\mathbf{w}')^\mathsf{T}\mathbf{a}_i)\ell_i' = ((\mathbf{w}')^\mathsf{T}\mathbf{a}_i)\ell_i = \gamma_{\max}^{-1}(\mathbf{v}_{\max}^\mathsf{T}\mathbf{a}_i)\ell_i \ge \gamma_{\max}^{-1}\gamma_{\max} = 1 = 1 - 0 = 1 - \xi_i'.$$

If $\ell_i \ne \ell_i'$, then

$$\begin{aligned} ((\mathbf{w}')^\mathsf{T}\mathbf{a}_i)\ell_i' = -((\mathbf{w}')^\mathsf{T}\mathbf{a}_i)\ell_i &= -\gamma_{\max}^{-1}(\mathbf{v}_{\max}^\mathsf{T}\mathbf{a}_i)\ell_i \\ &\ge -\gamma_{\max}^{-1}\|\mathbf{v}_{\max}\|_2 \cdot \|\mathbf{a}_i\| \cdot |\ell_i| \ge -2(1)(1)(1) = 1 - 3 = 1 - \xi_i'. \end{aligned}$$

Also, $\xi_i' \ge 0$. So, $(\mathbf{w}', \{\xi_i'\}_{i=1}^n)$ satisfies the constraints of the minimization problem. Thus,

$$\|w\|_2^2 \le \|w\|_2^2 + C\sum_{i=1}^n \xi_i \le \|w'\|_2^2 + C\sum_{i=1}^n \xi_i' = \gamma_{\max}^{-2} + 0.6Cn.$$

Area B
May 2019

CSEM Preliminary Exam - Area B

May 28, 2019

Part I - CSE 383C - Num Anly: Linear Algebra

This is a closed-book exam. Please explain all your answers.

1. [**40 points.**] Let $M \in \mathbb{R}^{m\times m}$ and $N \in \mathbb{R}^{n\times n}$ be symmetric positive definite matrices. We use M to define an inner product function in $\mathbb{R}^{m\times m}$ as follows: $< y, z >_M := y^T Mz$. We use N to define an inner product function in $\mathbb{R}^{n\times n}$ as follows: $< x, w >_N := x^T Nw$. Using these inner products we wish to *redefine* the SVD of a matrix $A \in \mathbb{R}^{m\times n}$ using M-inner product for the column space and the N-inner product for the row space.

   (a) Give the formula that defines the i_{th} singular value of A and its associated left and right singular vectors in the space defined by the M and N inner products.

   (b) What are the orthogonality conditions that the left and right singular vector *matrices* satsify?

   (c) You have a routine that computes the standard SVD of A (in the canonical inner product). How can you use this routine to compute the new redefined SVD of A?

2. [**30 points.**] Let $A \in \mathbb{R}^{n\times n}$ be non-singular. Let P, L, U be the partial pivoting LU decomposition of A such that $PA = LU$. P is a permutation, L is triangular with $L_{ii} = 1$. Assume $|L_{ij}| \le 1$. U is upper triangular with elements U_{ij}. Show that $\text{cond}_\infty(A) \ge \|A\|_\infty / \min_j |U_{jj}|$, where $\text{cond}_\infty(A)$ is the infinity-norm condition number of A.

3. [**30 points.**] Let $A \in \mathbb{R}^{m\times n}$ and consider the following algorithm for computing the first singular value σ_1 (assume $\sigma_1 > \sigma_2$) and the corresponding left and right singular vectors u_1 and v_1:

   choose $q_0 \in \mathbb{R}^n, \quad q \neq 0$
   for $k = 1, \ldots,$
   $\quad p_k = Aq_{k-1}$
   $\quad p_k = p_k / \|p_k\|_2$
   $\quad q_k = A^T p_k$
   $\quad q_k = q_k / \|q_k\|_2$
   end

   Explain what's the logic of the method, how we obtain u_1, σ_1 and v_1, and how it can fail.

Area B
May 2019

Part II - CSE 383L
Numerical Analysis: Differential Equations

4. Consider the following weak form of a PDE:

Find $u \in \mathcal{S} = \mathcal{V}$ such that for all $w \in \mathcal{V}$,

$$a(w, u) = (w, f) \tag{1}$$

where $\mathcal{V}$ is a suitably defined Hilbert space (for example, a Sobolev space), and $a(\cdot, \cdot)$ and $(\cdot, \cdot)$ are inner products.

Part 1

Define the natural norm, $|||\cdot|||$ induced by $a(\cdot, \cdot)$ and state the Cauchy-Schwarz inequality in terms of $a(\cdot, \cdot)$ and $|||\cdot|||$.

Part 2

Consider a Galerkin finite element method for (1):

Find $u^h \in \mathcal{S}^h = \mathcal{V}^h$ such that for all $w^h \in \mathcal{V}^h$,

$$a(w^h, u^h) = (w^h, f)$$

and $\mathcal{V}^h$ is an appropriate finite element subspace of $\mathcal{V}$ for the weak form.

Prove the "best approximation property," that is, show

$$|||e||| \leqslant |||U^h - u||| \quad \forall U^h \in \mathcal{V}^h$$

where $e = u^h - u$ is the error in the Galerkin finite element solution.

Area B
May 2019

T.J.R. Hughes
June 2, 2019.

Problem 4.

Part 1 $\quad |||w||| = +\left(a(w,w)\right)^{1/2}$

$$a(v,w) \le |a(v,w)| \le |||v|||\ |||w|||$$

Part 2 You can do this part at least three different ways. Here is one way:

Let
$$\begin{aligned} e &= u^h - u \\ &= u^h - U^h + U^h - u \\ &= \quad e^h \quad + \quad \eta \end{aligned}$$

where U^h is an "interpolate" of u.
Note $e \in \mathcal{V}$, $e^h \in \mathcal{V}^h$, $\eta \in \mathcal{V}$.

$$|||e|||^2 = a(e,e)$$

$$= a(e^h + \eta, e)$$

0 Galerkin orthogonality

$$= \cancel{a(e^h, e)} + a(\eta, e)$$

$$\le |||\eta|||\ \cancel{|||e|||} \qquad \text{Cauchy-Schwarz}$$

$$\therefore |||e||| \le |||\eta||| = |||U^h - u||| \quad \forall U^h \in \mathcal{V}^h$$

5. Consider the non-dimensional form of the advection-diffusion equation,

$$Pu_{,x} = u_{,xx} \quad \text{on } (0,1)$$

with boundary conditions

$$u(0) = 1,$$
$$u(1) = 0,$$

where $P > 0$ is the Péclet number. To solve this BVP, employ the following finite difference method (FDM):

$$\frac{P(u_{A+1} - u_{A-1})}{2h} = \frac{(u_{A+1} - 2u_A + u_{A-1})}{h^2}$$

where $A = 1, 2, \ldots, A_{\max} - 1$ and boundary conditions

$$u_0 = 1$$
$$u_{A_{\max}} = 0.$$

Part 1

Look for solutions of the FDM in the form $u_A = c_1\zeta_1^A + c_2\zeta_2^A$. Determine ζ_1, ζ_2, c_1, and c_2.

Part 2

Show that the solution exhibits spurious oscillations when the mesh Péclet number

$$P^h = P\frac{h}{2} > 1.$$

Part 3

The FDM can also be viewed as a Galerkin finite element method (FEM) in which piecewise-linear C^0-continuous basis functions are employed. With this, one can show

$$||e||_1 \leqslant \widetilde{C}(1 + P)h^p||u||_q.$$

What are the values of p and q?

Area B
May 2019

Problem 5

Part 1 You can do this part either before or after determining $\zeta_{1,2}$. Here is the solution "before":

$$u_A = c_1 \zeta_1^A + c_2 \zeta_2^A \,, \quad u_0 = 1 \,, \quad u_{A_{max}} = 0.$$

$$1 = c_1 \zeta_1^0 + c_2 \zeta_2^0 = c_1 + c_2$$

$$c_2 = 1 - c_1 .$$

$$0 = u_{A_{max}} = c_1 (\zeta_1)^{A_{max}} + c_2 (\zeta_2)^{A_{max}}$$

$$c_1 = \zeta_2^{A_{max}} \Big/ \left(\zeta_2^{A_{max}} - \zeta_1^{A_{max}} \right)$$

$$= 1 \Big/ \underbrace{\left(1 - (\zeta_1 / \zeta_2)^{A_{max}} \right)}_{= \alpha}$$

$$u_A = \frac{1}{\alpha} \zeta_1^{\cancel{A}} + \left(1 - \frac{1}{\alpha}\right) \zeta_2^A$$

$$= \frac{1}{\alpha} \left(\zeta_1^A - \zeta_2^A \right) + \zeta_2^A$$

$$= \frac{\zeta_1^A - \zeta_2^A}{1 - (\zeta_1/\zeta_2)^{A_{max}}} + \zeta_2^A .$$

Area B
May 2019

Problem 5

Part 2

$$P^h$$

$$\frac{Ph}{2}(u_{A+1}-u_{A-1}) = u_{A+1} - 2u_A + u_{A-1}$$

$$(P^h-1)u_{A+1} + 2u_A - (P^h+1)u_{A-1} = 0$$

$$(P^h-1)\zeta^2 + 2\zeta - (P^h+1) = 0$$

$$\zeta_{1,2} = \left(-\cancel{2} \pm \sqrt{\cancel{4}+\cancel{4}(P^h-1)(P^h+1)}\right)/\cancel{2}(P^h-1)$$

$$= \left(-1 \pm \sqrt{\cancel{1}+(P^h)^2-\cancel{1}}\right)/(P^h-1)$$

$$= (-1 \pm P^h)/(P^h-1)$$

$$= (-1+P^h)/(P^h-1),\ (-1-P^h)/(P^h-1)$$

$$= 1,\ -(P^h+1)/(P^h-1)$$

$$= 1,\ (1+P^h)/(1-P^h)$$

If $P^h > 1$, $\zeta_2 < 0$, and ζ_2^A oscillates.

Part 3. $\|e\|_1 \le (1+P)\|U^h - u\|_1 \quad \forall U^h \in \mathcal{V}^h$

$\exists U^h \in \mathcal{V}^h$ (an "interpolate") $\ni$

$$\|U^h - u\|_s \le \tilde{C} h^{k+1-s} \|u\|_{k+1}$$

$k=1, s=1$. Therefore $p=1$, $q=2$.

6. Consider a central difference method in space and in time for the second-order wave equation

$$u_{,tt} = c^2 u_{,xx}$$

with periodic boundary conditions, viz.,

$$\frac{u_A^{n+1} - 2u_A^n + u_A^{n-1}}{\Delta t^2} = \frac{c^2(u_{A+1}^n - 2u_A^n + u_{A-1}^n)}{h^2}$$

where $u_A^h \approx u(x_A, t_n)$, etc. We can write this equation in non-dimensional fashion in terms of the Courant number, $C = \frac{c\Delta t}{h}$ as follows.

$$u_A^{n+1} - 2u_A^n + u_A^{n-1} = C^2(u_{A+1}^n - 2u_A^n + u_{A-1}^n) \tag{$*$}$$

Part 1

Is the method implicit or explicit?

Part 2

Perform a Von Neumann stability analysis and prove the method is Von Neumann stable if $C \leqslant 1$.

Part 3

What result of Fourier analysis enables one to show that Von Neumann stability implies ℓ_2-stability?

Part 4

Introduce the local truncation error in the usual way for ($*$), namely $\Delta t^2 \mathcal{T}(\Delta t, h)$. Show that $\mathcal{T}(\Delta t, h) = O(\Delta t^p) + O(\Delta t^q)$ and determine p and q.

Area B
May 2019

Problem 6

Part 1 Explicit

Part 2 $u_A^{n+1} - 2u_A^n + u_A^{n-1} = C^2\left(u_{A+1}^n - 2u_A^n + u_{A-1}^n\right)$

$$\hat{u}^{n+1} - 2\hat{u}^n + \hat{u}^{n-1} = C^2\big(\underbrace{e^{-i\xi} - 2 + e^{+i\xi}}_{-2(1-\cos\xi)}\big)\hat{u}^n$$

$$= -4\sin^2\left(\tfrac{\xi}{2}\right)$$

$$\zeta^2 - 2\zeta + 1 = -4C^2\sin^2\left(\tfrac{\xi}{2}\right)\zeta$$

$$\zeta^2 - 2\big(\underbrace{1 - 2C^2\sin^2\left(\tfrac{\xi}{2}\right)}_{=\alpha}\big)\zeta + 1 = 0.$$

$$\zeta_{1,2} = \left(\cancel{2}\alpha \pm \sqrt{\cancelto{1}{4}\alpha^2 - \cancelto{1}{4}}\right)/\cancel{2}$$

Assume $C \leq 1$. Then $\alpha^2 - 1 = (1 - 2C^2\sin^2(\frac{\xi}{2})) - 1$
$\leq 0 \ \forall \xi \in (0, 2\pi)$.

Hence, we can write:
$\zeta_{1,2} = \alpha \pm i\sqrt{1-\alpha^2}$ and

$|\zeta_{1,2}|^2 = \zeta_{1,2}\bar{\zeta}_{1,2} = \alpha^2 + (1-\alpha^2) = 1$. Thus, when $C \leq 1$, the method is VN stable.

Note: We can show that the result is sharp, in that if $C > 1$, it is VN unstable, as done in class.

Part 3 Parseval identity

Area B
May 2019

Problem 6

Part 4 Be careful to distinguish between small c (wave speed) and large $C = c\Delta t/h$ (Courant number).

$$(1)\quad \Delta t^2 T = u(x_A, t_{n+1}) - 2u(x_A, t_n) + u(x_A, t_{n-1})$$
$$- C^2\big(u(x_{A+1}, t_n) - 2u(x_A, t_n) + u(x_{A-1}, t_n)\big)$$

$$(2)\quad u(x_A, t_{n\pm1}) = u(x_A, t_n) \pm \cancel{\Delta t\, u_{,t}(x_A, t_n)} + \frac{1}{2}\Delta t^2 u_{,tt}(x_A, t_n) \pm \cancel{\frac{\Delta t^3}{3!} u_{,ttt}}(x_A, t_n) + O(\Delta t^4)$$

$$(3)\quad u(x_{A\pm1}, t_n) = u(x_A, t_n) \pm \cancel{h\, u_{,x}(x_A, t_n)} + \frac{1}{2}h^2 u_{,xx}(x_A, t_n) \pm \cancel{\frac{h^3}{3!} u_{,xxx}}(x_A, t_n) + O(h^4)$$

$(2) \,\&\, (3) \longrightarrow (1) \Rightarrow$

$$\Delta t^2 T = \Delta t^2 \overbrace{\big(u_{,tt}(x_A, t_n) - c^2 u_{,xx}(x_A, t_n)\big)}^{=0} + O(\Delta t^4) + O(\Delta t^2 h^2)$$

$$\therefore\ T = O(\Delta t^2) + O(h^2)$$

Note: All odd derivative terms cancel.

CSEM Preliminary Exam
AREA B
Part I - CSE 383C Num Anly: Linear Algebra

Question 1: (40p) Consider a function `myfun` that takes as input an $m \times n$ matrix $\mathbf{A}$ with real entries, and computes two matrices $\mathbf{B}$ and $\mathbf{C}$ through the following procedure:

$$
\begin{array}{l}
[\mathbf{B}, \mathbf{C}] = \texttt{myfun}(\mathbf{A}) \\
\quad [m, n] = \texttt{size}(\mathbf{A}) \\
\quad \mathbf{B} = \mathbf{A} \\
\quad \mathbf{C} = \mathbf{I}_n \\
\quad \textbf{for } i = 1 : \min(m, n) \\
\qquad \mathbf{x} = \mathbf{B}(i, i:n)^{\mathrm{t}} \\
\qquad \mathbf{H} = \texttt{householder}(n, \mathbf{x}) \\
\qquad \mathbf{B} = \mathbf{B}\mathbf{H}^{\mathrm{t}} \\
\qquad \mathbf{C} = \mathbf{H}\mathbf{C} \\
\quad \textbf{end}
\end{array}
$$

The matrix $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\mathbf{Z}^{\mathrm{t}}$ denotes the transpose of $\mathbf{Z}$, and the function `householder` builds a Householder reflector. To be precise, given a $k \times 1$ vector $\mathbf{x}$ and an integer n such that $k \le n$, the matrix $\mathbf{H} = \texttt{householder}(n, \mathbf{x})$ is the $n \times n$ Householder reflector such that for any vector $\mathbf{y}$ of size $(n-k) \times 1$, we have

$$
\mathbf{H} \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \|\mathbf{x}\| \\ \mathbf{0}_{k-1,1} \end{bmatrix},
$$

where $\mathbf{0}_{k-1,1}$ is the zero vector of size $(k-1) \times 1$, and where $\|\mathbf{x}\|$ is the Euclidean norm.

(a) (10p) Describe how the matrices $\mathbf{B}$ and $\mathbf{C}$ relate to $\mathbf{A}$, and any particular properties they may have (such as being diagonal / orthogonal / triangular / ...).

(b) (10p) Can any numerical problems arise in the execution of the function `myfun`? In other words, is there a risk that you may encounter division by zero, or excessive amplification of round off errors?

(c) (10p) As m and n grow, the number of floating point operations required by `myfun` grows as $O(m^a n^b)$ for some positive numbers a and b. Determine a and b. Is it possible to slightly rearrange the execution of `myfun` to improve its asymptotic complexity without changing its output (assuming exact arithmetic)?

(d) (10p) Suppose that $\mathbf{A}$ is an $m \times n$ matrix of rank n. Describe how you can use `myfun` to solve an inconsistent linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ in a least squares sense.

Area B
2020

Question 2: (20p) Let $f : V \to W$ be a function from the vector space $V = \mathbb{R}^n$ to the vector space W of all matrices of size $n \times n$. We define the *relative condition number* $\kappa(\mathbf{x})$ of f at the point $\mathbf{x}$ to be the quantity

$$\kappa(\mathbf{x}) = \lim_{\delta \to 0} \sup_{\|\mathbf{y}\|_V \leq \delta} \frac{\|f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x})\|_W \, \|\mathbf{x}\|_V}{\|f(\mathbf{x})\|_W \, \|\mathbf{y}\|_V}.$$

We use the max-norm on V, so that $\|\mathbf{x}\|_V = \max_{1 \leq i \leq n} |x_i|$. In this question, you are asked to evaluate $\kappa(\mathbf{x})$ for the function

$$f(\mathbf{x}) = \mathbf{U}\,\mathbf{D}(\mathbf{x})\,\mathbf{Q}^*$$

where $\mathbf{U}$ and $\mathbf{Q}$ are two fixed unitary matrices of size $n \times n$, and where $\mathbf{D}(\mathbf{x})$ is the $n \times n$ matrix whose diagonal entries are given by the entries of $\mathbf{x}$.

(a) (10p) Evaluate $\kappa(\mathbf{x})$ when $\|\cdot\|_W$ is the spectral norm on W, so that

$$\|\mathbf{A}\|_W = \sup_{\|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2,$$

where $\|\cdot\|_2$ is the Euclidean norm.

(b) (10p) Evaluate $\kappa(\mathbf{x})$ when $\|\cdot\|_W$ is the Frobenius norm on W, so that

$$\|\mathbf{A}\|_W = \left(\sum_{i,j=1}^{n} |\mathbf{A}(i,j)|^2 \right)^{1/2}.$$

Question 3: (40p) Let $\mathbf{A}$ be a real matrix of size $m \times n$. You know that $m > n$ and that $\mathbf{A}$ has rank n. Suppose that you are interested in computing the real number

$$c = \inf_{\|\mathbf{x}\| = 1} \|\mathbf{A}\mathbf{x}\|,$$

and also a vector $\mathbf{w} \in \mathbb{R}^n$ such that

$$c = \|\mathbf{A}\mathbf{w}\|.$$

(a) (10p) Describe how you would determine c and $\mathbf{w}$ if you had access to a function that computes the singular value decomposition (SVD) of $\mathbf{A}$.

(b) (10p) Describe how you would determine c and $\mathbf{w}$ if you had access to a function that computes the eigenvalue decomposition (EVD) of a symmetric matrix (but *not* a function that computes the SVD). If you had access to routines for computing both the SVD and the EVD, then which would you choose?

(c) (10p) Suppose now that you do not have access to functions for computing either the EVD or the SVD of a matrix. However, you can compute standard factorizations of a matrix (QR, LU, Cholesky, etc.) and you can perform standard operations such as the matrix-matrix product, solving a linear system, and so on. Describe an easy to implement and computationally efficient iterative method for computing approximations to c and $\mathbf{w}$. Comment on what properties of $\mathbf{A}$ determine the speed of convergence of the iteration.

(d) (10p) Suppose now that $\mathbf{A}$ is large but very sparse. In other words, you cannot afford to compute any full factorizations, but you can affordably compute matrix-vector products involving $\mathbf{A}$. Would it be possible to use the Lanczos iteration to estimate c and $\mathbf{w}$? Either describe how this would work, or explain why this method is not applicable.

Area B
2020

Part II - CSE 383L Num Anly: Differential Equations

4. The 2-point boundary value problem,

$$\begin{cases} \dfrac{d^4y}{dx^4} = f(x), \qquad 0 < x < 1 \\ y(0) = y(1) = y'(0) = y'(1) = 0 \end{cases}$$

can be seen as an approximation of a clamped beam.
(a) Rewrite the problem on variational form and propose suitable spaces for the continuous problem and corresponding discrete FEM approximation. Discuss convergence.
(b) Rewrite the problem as a first order system and determine the corresponding trapezoidal rule FDM approximation.
(c) Describe how this first order system can be solved by initial value techniques (shooting).

5. Consider the heat equation,

$$\begin{cases} \dfrac{\partial u}{\partial t} = \nabla \cdot \sigma(x,y)\nabla u, \qquad 0 < \sigma_1 < \sigma(x,y) \le \sigma_2, \qquad 0 < x < 1, 0 < y < 1, \\ u(x,0) = u(x,1) = 0, \quad 0 \le x \le 0 \\ \dfrac{\partial u}{\partial x}(0,y) = \dfrac{\partial u}{\partial x}(1,y) = 0, \quad 0 < y < 1 \\ u(x,y,0) = u_0(x,y), \quad 0 < x < 1, 0 < y < 1 \end{cases}$$

(a) Formulate an implicit Euler-in-time FEM approximation of this problem based on an appropriate variational formulation.
(b) Outline a convergence proof based on the properties of the bilinear and linear forms.
(c) Determine the convergence condition (CFL number) for a FDM approximation based on forward Euler in time and centered difference in space. Use von Neumann analysis and assume periodic boundary conditions with constant conductivity σ.

6. The following nonlinear hyperbolic conservation law is given,

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x} f(u) = 0, \qquad 0 \le f'(u) \le C, \qquad t > 0, 0 < x < 1 \\ \textit{periodic boundary conditions} \\ u(x,0) = u_0(x), 0 < x < 1 \end{cases}$$

(a) Formulate an explicit first order finite volume approximation
(b) Show that the scheme is on discrete conservation form and give conditions on the step sizes such that the scheme is monotone.
(c) Formulate a P1 discontinuous Galerkin (DG) approximation with appropriate interface conditions.

Area B
May 2021

CSEM Area B Preliminary Exam

May 26, 2021, about any 3 hours from 9:00 a.m. to 3:00 p.m.

Open notes, open book(s), open internet.

Work Part I and either Part II or Part III, but not both.

Part I, Numerical Analysis: Linear Algebra

1. (10 points) Let $A \in \mathbb{R}^{m\times m}$ have the property that $\|Ax\|_2 = \|x\|_2$ for all $x \in \mathbb{R}^m$. Use the Singular Value Decomposition Theorem to show that A must be a unitary matrix.

2. (10 points) Let A be a symmetric positive definite matrix. Prove or disprove that $f : \mathbb{R}^m \to \mathbb{R}$ defined by

$$f(x) = \sqrt{x^T Ax}$$

is a vector norm. You may invoke knowledge you have about various p-norms and other relevent knowledge from the course.

3. (20 points) Let $A \in \mathbb{R}^{m\times n}$ have linearly independent columns.

You have encountered a number of different algorithms for computing the QR factorization of such a matrix. You will build on that knowledge in this question.

(a) Use permutation matrices and what you know about the QR factorization to prove that there exists a QL factorization of A:

$$A = \widehat{Q}L$$

where $\widehat{Q}$ is an $m \times n$ matrix with orthonormal columns and L is a lower triangular matrix of appropriate size.

(b) Develop an algorithm inspired by Modified Gram-Schmidt for computing this factorization. It should overwrite matrix A with $\widehat{Q}$ and also compute L. Your algorithm may not use any additional temporary space. Give enough detail so that you can discuss why each step in your algorithm is well defined.

(c) Briefly, say a few words about what techniques you would use instead if your algorithm is to produce highly orthonormal columns in $\widehat{Q}$ when floating point arithmetic is employed.

What is it fundamentally about these techniques that improves the orthogonality of the resulting matrix $\widehat{Q}$?

Part II, Numerical Analysis: Differential Equations

1. Consider an ODE initial value problem and the corresponding, so called, θ-method,

$$y'(t) = f(y(t)), \quad t > 0, \quad y(0) = y_0,$$
$$y_{n+1} = y_n + h\big(\theta\, f(y_n) + (1-\theta)\, f(y_{n+1})\big), \quad y_0 \text{ given}, \ t_n = nh, \ 0 < \theta < 1.$$

(a) Determine 0-stability (Dahlquist stability), A-stability and the order of the method as a function of θ.

(b) Even if the θ-method is not strictly symplectic for $\theta = 1/2$, show that this θ value is the only one for which, as a complex valued function, $|y(t)| = |y_0|$, $t > 0$, when $f(y) = iy$.

(c) The approximation y_{n+1} is computed using $\theta = 1$ and step size h. Also starting with y_n assume there is another approximation of $y(t_{n+1}) \approx \tilde{y}$, based on the same method but with step size $h/2$. Discuss how these two approximations (y_{n+1} and $\tilde{y}$) of $y(t_{n+1})$ can be used for an a posteriori error estimate.

2. Given the parabolic differential equation for $u(x,t)$,

$$u_t = au_{xx} + bu_x + cu + f(x), \quad 0 < x < 1, \ t > 0, \ a > 0, \ c < 0,$$
$$u(x,0) = u_0(x), \quad u(0,t) = 0, \ u_x(1,t) = 0.$$

(a) Sketch an implicit Euler in time finite element approximation.

(b) Sketch a finite difference approximation based on forward differencing in time and centered differencing in space.

(c) For theory show coercivity of the relevant bilinear form for the stationary problem (no u_t term) based on limits on $|b|$ for the method given in (a). Also apply von Neumann stability analysis for the method in (b) in the simplified case of $b = 0$ and periodic boundary conditions.

3. Consider the following nonlinear hyperbolic conservation law for $u(x,t)$,

$$u_t + f(u)_x = 0, \quad 0 < x < 1, \ t > 0, \ f'(u) > 0,$$
$$u(x,0) = u_0(x), u(0,t) = 0.$$

(a) Write the equation on weak form and formulate a general Discontinuous Galerkin (DG) method based of forward differencing (Euler) in time.

(b) Give realistic interface conditions (numerical fluxes).

(c) Show that the method is essentially explicit (only inversion of block diagonal matrix) for discontinuous, piecewise linear (P1) elements.

Part III, Foundations of Machine Learning and Data Science

Theorem 1: Random Projection Theorem

Let $\mathbf{v} \in \mathbb{R}^d$ be fixed. Draw r vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$ as i.i.d. standard spherical Gaussian vectors $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $i = 1, 2, \dots, r$. Consider the random linear map $f : \mathbb{R}^d \to \mathbb{R}^r$ given by $f(\mathbf{x}) = (\langle \mathbf{u}_1, \mathbf{x} \rangle, \dots, \langle \mathbf{u}_r, \mathbf{x} \rangle)$. For any $\epsilon \in (0, 1)$,

$$\mathbf{Prob}\left(\left|\|f(\mathbf{v})\|_2 - \sqrt{r}\|\mathbf{v}\|_2\right| \geq \epsilon\sqrt{r}\|\mathbf{v}\|_2\right) \leq 3e^{-\frac{r\epsilon^2}{96}}$$

where the probability is taken with respect to the random draws of the vectors $\mathbf{u}_1, \dots, \mathbf{u}_r$.

The first two problems consider the version of the Random Projection Theorem above.

1. In the Random Projection Theorem, the statement is not true if the first two conditions "Let $\mathbf{v} \in \mathbb{R}^d$ be fixed. Draw r vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$ as i.i.d. standard spherical Gaussian vectors $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$," are replaced by "Draw r vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r$ as i.i.d. standard spherical Gaussian vectors $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Let $\mathbf{v} \in \mathbb{R}^d$ be an arbitrary vector." Show this by constructing a counterexample.

2. From the Random Projection Theorem, prove the following statement.

Let $V = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p\}$ be a fixed set of p points in $\mathbb{R}^d$. If $r \geq 19200\log(2p)$, there exists a linear map $f : \mathbb{R}^d \to \mathbb{R}^r$ with the property that for all $\mathbf{v}_j, \mathbf{v}_k \in V$,

$$.9\|\mathbf{v}_j - \mathbf{v}_k\|_2 \leq \frac{1}{\sqrt{r}}\|f(\mathbf{v}_j) - f(\mathbf{v}_k)\|_2 \leq 1.1\|\mathbf{v}_j - \mathbf{v}_k\|_2$$

State clearly and justify all of the steps in the derivation.

The next two problems consider the following set up. Consider a collection of data points $\mathbf{x}_1, \dots, \mathbf{x}_n$ in $\mathbb{R}^d$ which can be partitioned into two sets $\{S_1, S_2\}$ such that

$$\begin{aligned}\max_{\mathbf{x}\in S_1}\|\mathbf{x} - \mu_{S_1}\|_2 &\leq \min_{\mathbf{y}\in S_1, \mathbf{z}\in S_2}\|\mathbf{y} - \mathbf{z}\|_2 \\ \max_{\mathbf{x}\in S_2}\|\mathbf{x} - \mu_{S_2}\|_2 &\leq \min_{\mathbf{y}\in S_1, \mathbf{z}\in S_2}\|\mathbf{y} - \mathbf{z}\|_2 \end{aligned} \tag{1}$$

where $\mu_{S_1} \in \mathbb{R}^d$ is the average of the points in S_1 and $\mu_{S_2} \in \mathbb{R}^d$ is the average of the points in S_2. Consider the data matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ whose columns are the data points $\mathbf{x}_1, \dots, \mathbf{x}_n$. Letting $\mu = \frac{1}{n}\sum_{j=1}^n \mathbf{x}_j \in \mathbb{R}^d$ be the average of *all* the points, consider the centered matrix $\mathbf{X}_c = [\mathbf{x}_1 - \mu, \mathbf{x}_2 - \mu, \dots, \mathbf{x}_n - \mu]$ and its Gram matrix $\mathbf{X}_c^T\mathbf{X}_c \in \mathbb{R}^{n\times n}$.

3. Suppose that the two clusters are of the same size: $|S_1| = |S_2| = \frac{n}{2}$. Prove that under the separability condition (1), the leading eigenvector $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$ of $\mathbf{X}_c^T\mathbf{X}_c$ has the property that $\text{sign}(v_j) \neq \text{sign}(v_k)$ whenever $\mathbf{x}_j$ and $\mathbf{x}_k$ belong to different clusters. (Hint: show if $\text{sign}(v_j) = \text{sign}(v_k)$ for some $\mathbf{x}_j$ and $\mathbf{x}_k$ belonging to different clusters, this contradicts the fact that $\mathbf{v} = \arg\max_{\mathbf{u}:\|\mathbf{u}\|_2=1}\|\mathbf{X}_c\mathbf{u}\|_2$)

4. Use the answer from the previous problem to describe a clustering algorithm which takes as input a collection of data points $\mathbf{x}_1, \dots, \mathbf{x}_n$ and outputs a partition of the data points into two clusters.

CSEM Area B Preliminary Exam

Numerical Analysis: Linear Algebra (CSE 383C)
Numerical Analysis: Differential Equations (CSE 383L)

May 27, 2022, 9:00 a.m. to 12:00 noon
200 points total

1. [30 pts.] In this problem, we seek to fit a polynomial $p(x) = \sum_{j=1}^{n} c_j x^{j-1}$ to a given set of measured points $\left(x_i,\, y_i\right)_{i=1}^{m}$ such that $x_1 < x_2 < \cdots < x_m$ and $n < m$. To be precise, we seek to determine a polynomial p that minimizes the error

$$E = \sum_{i=1}^{m} \left|p(x_i) - y_i\right|^2.$$

(a) Consider first the case of fitting the points to a quadratic $p(x) = c_1 + c_2 x + c_3 x^2$. Describe how you can formulate the minimization problem using a linear algebraic framework, and how you would go about solving the problem that you formulated. In this part of the question, you do not need to worry about computational efficiency or sensitivity to round-off errors.

(b) Now suppose that you want to fit an $n-1$'th order polynomial $p(x) = \sum_{j=1}^{n} c_j x^{j-1}$. Describe how the associated linear algebraic problem can be solved in different ways, using the QR, singular value, and Cholesky decompositions. Discuss the advantages and disadvantages of the different methods when m and n are both large.

2. [40 pts.] Let $\mathbf{A}$ be an $n \times n$ symmetric non-singular matrix that is partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11}$ is an invertible matrix. The matrix $\mathbf{A}$ admits a factorization of the form

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^*$$

where

$$\mathbf{L} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{I} \end{bmatrix} \qquad \text{and} \qquad \mathbf{D} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{22} \end{bmatrix}.$$

Please motivate your answers to the questions below.

(a) Specify the matrices $\mathbf{L}_{21}$ and $\mathbf{B}_{22}$.

(b) Now consider the special case where $\mathbf{A}_{11} = [e]$ for some real number $e \in (0, 1)$, and

$$\mathbf{A} = \begin{bmatrix} e & 1 \\ 1 & 0 \end{bmatrix}.$$

Specify the corresponding factors $\mathbf{D}$ and $\mathbf{L}$, the condition numbers $\kappa(\mathbf{A})$ and $\kappa(\mathbf{D})$ as functions of e, and finally the limit

$$\lim_{e \searrow 0} \frac{\kappa(\mathbf{D})}{\kappa(\mathbf{A})}.$$

(Use the standard definition $\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ where $\|\mathbf{A}\| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|$.)

(c) Is it possible to factorize the matrix $\mathbf{A} = \begin{bmatrix} e & 1 \\ 1 & 0 \end{bmatrix}$ as $\mathbf{A} = \mathbf{C}\mathbf{C}^*$, where $\mathbf{C}$ is lower triangular?

3. [30 pts.] Let $\mathbf{A}$ be an $n \times n$ matrix, and define the function

$$r(\mathbf{x}) = \mathbf{x}^* \mathbf{A} \mathbf{x}, \qquad \mathbf{x} \in \mathbb{R}^{n\times 1}.$$

Let $\mathbf{v}$ be an eigenvector of $\mathbf{A}$ of unit length with an associated eigenvalue λ, so that

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

For full credit, please motivate your answers to questions (b) and (c).

(a) Evaluate the gradient $\nabla r(\mathbf{x})$.

(b) Suppose that $(\mathbf{x}_j)_{j=1}^{\infty}$ is a sequence of unit vectors that converge to $\mathbf{v}$ at the rate $O(j^{-3})$ as $j \to \infty$. In other words, $\|\mathbf{x}_j\| = 1$ and $\|\mathbf{v} - \mathbf{x}_j\| = O(j^{-3})$. Please specify the limit of the sequence $(r(\mathbf{x}_j))_{j=1}^{\infty}$, as well as the speed of convergence (in the worst case).

(c) Does your answer to question (b) change if you know that $\mathbf{A}$ is Hermitian, so that $\mathbf{A}^* = \mathbf{A}$?

4. [25 pts.] Consider the ODE $u' = f(u)$ and the generic, explicit, second order Runge-Kutta method for parameter $\alpha \neq 0$ with Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}.$$

(a) Combine the Runge-Kutta stages to write the method for u^{n+1} in terms of u^n in a single line over the time step from t^n to t^{n+1}, $\Delta t = t^{n+1} - t^n > 0$.

(b) Show that the local truncation error is at least $\mathcal{O}(\Delta t^3)$, but not in general $\mathcal{O}(\Delta t^4)$.

5. [30 pts.] Let $\Omega \subset \mathbb{R}^3$ be a domain, $\mathbf{e} = (1,1,1)/\sqrt{3}$, and consider the variational problem: Find $u \in H^1(\Omega)$ such that

$$(\nabla u, \nabla v) + (\mathbf{e} \cdot \nabla u, v) + (u, v) = (f(u), v) \quad \forall v \in H^1(\Omega).$$

(a) Find the partial differential equation and boundary condition corresponding to this variational form.

(b) Suppose that f is Lipschitz with constant $L \leq 1/4$ and that we use the finite element method to approximate u by $u_h \in V_h \subset H^1(\Omega)$. Prove the error estimate

$$\|u - u_h\|_{H^1} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1}$$

for some constant $C > 0$.

6. [15 pts.] Perform a von Neumann linear stability analysis for the backward Euler scheme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{h}\left[f(u_{j+1}^{n+1}) - f(u_{j-1}^{n+1})\right]$$

to show that the scheme is unconditionally stable.

7. [30 pts.] Let $E = [0,h]^2$ and suppose that $v \in C^1(E)$.

(a) Show that

$$\left(\int_0^h |v(0,y)|^2 \, dy\right)^{1/2} \leq C\left[h^{-1/2}\|v\|_{L^2(E)} + h^{1/2}\|\nabla v\|_{L^2(E)}\right].$$

(b) Prove the inverse inequality: If v is a polynomial of degree $k \geq 0$, then there is some $C > 0$ independent of v and h such that

$$\|\nabla v\|_{L^2(E)} \leq Ch^{-1}\|v\|_{L^2(E)}.$$

CSEM Area B Preliminary Exam

Numerical Analysis: Linear Algebra (CSE 383C)
Foundational Techniques in Machine Learning & Data Science (CSE 382M)

May 27, 2022, 9:00 a.m. to 12:00 noon
200 points total

1. [30 pts.] In this problem, we seek to fit a polynomial $p(x) = \sum_{j=1}^{n} c_j x^{j-1}$ to a given set of measured points $\left(x_i,\, y_i\right)_{i=1}^{m}$ such that $x_1 < x_2 < \cdots < x_m$ and $n < m$. To be precise, we seek to determine a polynomial p that minimizes the error

$$E = \sum_{i=1}^{m} \left| p(x_i) - y_i \right|^2.$$

(a) Consider first the case of fitting the points to a quadratic $p(x) = c_1 + c_2 x + c_3 x^2$. Describe how you can formulate the minimization problem using a linear algebraic framework, and how you would go about solving the problem that you formulated. In this part of the question, you do not need to worry about computational efficiency or sensitivity to round-off errors.

(b) Now suppose that you want to fit an $n-1$'th order polynomial $p(x) = \sum_{j=1}^{n} c_j x^{j-1}$. Describe how the associated linear algebraic problem can be solved in different ways, using the QR, singular value, and Cholesky decompositions. Discuss the advantages and disadvantages of the different methods when m and n are both large.

2. [40 pts.] Let $\mathbf{A}$ be an $n \times n$ symmetric non-singular matrix that is partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11}$ is an invertible matrix. The matrix $\mathbf{A}$ admits a factorization of the form

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^*$$

where

$$\mathbf{L} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{I} \end{bmatrix} \qquad \text{and} \qquad \mathbf{D} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{22} \end{bmatrix}.$$

Please motivate your answers to the questions below.

(a) Specify the matrices $\mathbf{L}_{21}$ and $\mathbf{B}_{22}$.

(b) Now consider the special case where $\mathbf{A}_{11} = [e]$ for some real number $e \in (0,1)$, and

$$\mathbf{A} = \begin{bmatrix} e & 1 \\ 1 & 0 \end{bmatrix}.$$

Specify the corresponding factors $\mathbf{D}$ and $\mathbf{L}$, the condition numbers $\kappa(\mathbf{A})$ and $\kappa(\mathbf{D})$ as functions of e, and finally the limit

$$\lim_{e \searrow 0} \frac{\kappa(\mathbf{D})}{\kappa(\mathbf{A})}.$$

(Use the standard definition $\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ where $\|\mathbf{A}\| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|$.)

(c) Is it possible to factorize the matrix $\mathbf{A} = \begin{bmatrix} e & 1 \\ 1 & 0 \end{bmatrix}$ as $\mathbf{A} = \mathbf{C}\mathbf{C}^*$, where $\mathbf{C}$ is lower triangular?

3. [30 pts.] Let $\mathbf{A}$ be an $n \times n$ matrix, and define the function

$$r(\mathbf{x}) = \mathbf{x}^* \mathbf{A} \mathbf{x}, \qquad \mathbf{x} \in \mathbb{R}^{n \times 1}.$$

Let $\mathbf{v}$ be an eigenvector of $\mathbf{A}$ of unit length with an associated eigenvalue λ, so that

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

For full credit, please motivate your answers to questions (b) and (c).

(a) Evaluate the gradient $\nabla r(\mathbf{x})$.

(b) Suppose that $(\mathbf{x}_j)_{j=1}^{\infty}$ is a sequence of unit vectors that converge to $\mathbf{v}$ at the rate $O(j^{-3})$ as $j \to \infty$. In other words, $\|\mathbf{x}_j\| = 1$ and $\|\mathbf{v} - \mathbf{x}_j\| = O(j^{-3})$. Please specify the limit of the sequence $(r(\mathbf{x}_j))_{j=1}^{\infty}$, as well as the speed of convergence (in the worst case).

(c) Does your answer to question (b) change if you know that $\mathbf{A}$ is Hermitian, so that $\mathbf{A}^* = \mathbf{A}$?

4. [30 pts.] Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a fixed matrix, and suppose $\mathbf{X} = (X_{j,k}) \in \mathbb{R}^{n\times r}$ is a random matrix with independent and identically distributed entries $X_{j,k}$ satisfying $\mathbb{E}(X_{j,k}) = 0$ and $\mathrm{Var}(X_{j,k}) = 1/r$.

(a) Show that $\mathbb{E}\|\mathbf{AX}\|_F^2 = \|\mathbf{A}\|_F^2$.

(b) Use Markov's inequality to give a nontrivial upper bound on the probability of the event

$$\|\mathbf{AX}\|_F^2 \geq 10\,\|\mathbf{A}\|_F^2.$$

(c) In the case where $\mathbf{X} \in \mathbb{R}^{n\times r}$ has independent and identically distributed entries $X_{j,k} \sim \mathcal{N}(0, 1/r)$, give a bound on the number of columns r in $\mathbf{X} \in \mathbb{R}^{n\times r}$ to guarantee that with probability at least $1 - \frac{1}{10}$,

$$\frac{1}{2}\|\mathbf{A}\|_F^2 \leq \|\mathbf{AX}\|_F^2 \leq \frac{3}{2}\|\mathbf{A}\|_F^2.$$

5. [40 pts.] Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a full-rank matrix with $m \geq n$ and indexed by column vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$. Suppose that $\mathbf{A}$ has singular value decomposition $\mathbf{A} = \mathbf{U\Sigma V}^T$ and distinct singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_n$, and let $\mathbf{b} \in \mathbb{R}^m$ be a fixed vector. For each optimization problem below, provide expressions for the solution set in terms of $\mathbf{A}, \mathbf{U}, \mathbf{\Sigma}$, $\mathbf{V}$, and $\mathbf{b}$.

(a) $\displaystyle\max_{\mathbf{v}\in\mathbb{R}^n:\,\|\mathbf{v}\|_2=1} \|\mathbf{Av}\|_2$

(b) $\displaystyle\min_{\mathbf{v}\in\mathbb{R}^n:\,\|\mathbf{v}\|_2=1} \|\mathbf{Av}\|_2$

(c) $\displaystyle\min_{\mathbf{u}\in\mathbb{R}^n} \|\mathbf{Au} - \mathbf{b}\|_2^2$

(d) $\displaystyle\min_{\mathbf{w}\in\mathbb{R}^m} \sum_{j=1}^{n} \|\mathbf{a}_j - \mathbf{w}\|_2^2$

6. [30 pts.] Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a full-rank matrix with $m \geq n$ and indexed by column vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$, and fix $\mathbf{b} \in \mathbb{R}^m$. Consider $\Phi \in \mathbb{R}^{r\times m}$ such that $\Phi\mathbf{A}$ is full-rank, and

$$\mathbf{x}_{opt} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|_2^2, \qquad \tilde{\mathbf{x}}_{opt} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\Phi\mathbf{Ax} - \Phi\mathbf{b}\|_2^2$$

(a) Show that if

$$(1-\delta)\|\mathbf{Ax}_{opt} - \mathbf{b}\|_2^2 \leq \|\Phi(\mathbf{Ax}_{opt} - \mathbf{b})\|_2^2 \leq (1+\delta)\|\mathbf{Ax}_{opt} - \mathbf{b}\|_2^2,$$
$$(1-\delta)\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \leq \|\Phi(\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b})\|_2^2 \leq (1+\delta)\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2,$$

then

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \leq \left(\frac{1+\delta}{1-\delta}\right)\|\mathbf{Ax}_{opt} - \mathbf{b}\|_2^2. \tag{1}$$

(b) If Φ is a random matrix whose entries are i.i.d. $\mathcal{N}(0, 1/r)$, give a bound on the number of rows r needed to ensure the bound (1) with high probability.

Solutions to NALA part of Area B prelim, May 2022

Question 1:

(a) Form the $m \times 3$ matrix $\mathbf{A}$ with entries

$$\mathbf{A}(i,j) = x_i^{j-1}$$

and the $m \times 1$ vector $\mathbf{y} = [y_i]_{i=1}^m$. Then

$$E = \sum_{i=1}^{m} \big|p(x_i) - y_i\big|^2 = \sum_{i=1}^{m} \Big|\sum_{j=1}^{n} c_j x_i^{j-1} - y_i\Big|^2 = \sum_{i=1}^{m} \big|\mathbf{A}(i,:)\mathbf{c} - y_i\big|^2 = \|\mathbf{A}\mathbf{c} - \mathbf{y}\|^2,$$

where $\mathbf{c} = [c_1, c_2, c_3]^*$. To solve the minimization problem, you could for instance form the *normal equations*

$$\mathbf{A}^*\mathbf{A}\mathbf{c} = \mathbf{A}^*\mathbf{y}.$$

In the present case, the normal equations form a 3×3 non-singular system that can be solved using, e.g., Gaussian elimination.

(b) This problem can be formulated in a way that is entirely analogous to the method in (a), but involving an $m \times n$ matrix $\mathbf{A}$,

$$\min_{\mathbf{c} \in \mathbb{C}^{n \times 1}} \|\mathbf{A}\mathbf{c} - \mathbf{y}\|.$$

Let us discuss three ways of solving it:

Normal equations: Form the $n \times n$ matrix $\mathbf{N} = \mathbf{A}^*\mathbf{A}$, then form the Cholesky factorization $\mathbf{N} = \mathbf{R}^*\mathbf{R}$, and finally evaluate $\mathbf{c} = \mathbf{R}^{-1}(\mathbf{R}^*)^{-1}\mathbf{y}$ via two triangular solves. This method is quite efficient in practice, but leads to loss of accuracy when $\mathbf{A}$ is ill-conditioned, since $\kappa(\mathbf{N}) = (\kappa(\mathbf{A}))^2$. In the present case, $\mathbf{A}$ will get very ill-conditioned as n grows, so this method is not recommended.

QR factorization: Form the QR factorization $\mathbf{A} = \mathbf{Q}\mathbf{R}$, and then determine $\mathbf{c}$ via $\mathbf{c} = \mathbf{R}^{-1}(\mathbf{Q}^*\mathbf{y})$, where $\mathbf{R}^{-1}$ is applied via a triangular solve. This technique is computationally efficient, and reasonably numerically stable.

SVD factorization: Form the singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^*$, and then determine $\mathbf{c}$ via $\mathbf{c} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^*\mathbf{y}$. This method is the most numerically stable, and can easily be stabilized even further by ignoring all singular modes associated with singular values below a certain threshold. It is reasonably computationally efficient, although slightly slower than the methods relying on QR or the normal equations.

Question 2:

(a) Multiplying the factors together, we get

$$\mathbf{LDL}^* = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{L}_{21}^* \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \cdots = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{11}\mathbf{L}_{21}^* \\ \mathbf{L}_{21}\mathbf{A}_{11} & \mathbf{B}_{22} + \mathbf{L}_{21}\mathbf{A}_{11}\mathbf{L}_{21}^* \end{bmatrix}.$$

To ensure the product equals $\mathbf{A}$, we need $\mathbf{L}_{21}\mathbf{A}_{11} = \mathbf{A}_{21}$, which implies that

$$\mathbf{L}_{21} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}. \tag{1}$$

We must also have

$$\mathbf{A}_{22} = \mathbf{B}_{22} + \mathbf{L}_{21}\mathbf{A}_{11}\mathbf{L}_{21}^*. \tag{2}$$

Combining (2) and (1), we get

$$\mathbf{B}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}.$$

(b) Using the formulas we derived in (a), we immediately get

$$\begin{bmatrix} e & 1 \\ 1 & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ 1/e & 1 \end{bmatrix}}_{=\mathbf{L}} \underbrace{\begin{bmatrix} e & 0 \\ 0 & -1/e \end{bmatrix}}_{=\mathbf{D}} \underbrace{\begin{bmatrix} 1 & 1/e \\ 0 & 1 \end{bmatrix}}_{=\mathbf{L}^*}.$$

The matrix $\mathbf{D}$ clearly has the singular values e and $1/e$, so

$$\kappa(\mathbf{D}) = \frac{\sigma_{\max}}{\sigma_{\min}} = \frac{1/e}{e} = 1/e^2.$$

In order to evaluate $\kappa(\mathbf{A})$, we first compute the eigenvalues of $\mathbf{A}$:

$$\lambda_{1,2} = \frac{e}{2} \pm \sqrt{\frac{e^2}{4} + 1}.$$

Since the singular values of a symmetric matrix are the absolute values of the eigenvalues, we find that

$$\kappa(\mathbf{A}) = \frac{\sqrt{\frac{e^2}{4} + 1} + \frac{e}{2}}{\sqrt{\frac{e^2}{4} + 1} - \frac{e}{2}}.$$

We see that as $e \to 0$, we have

$$\kappa(\mathbf{D}) \to \infty, \qquad \text{and} \qquad \kappa(\mathbf{A}) \to 1.$$

In consequence

$$\lim_{e \searrow 0} \frac{\kappa(\mathbf{D})}{\kappa(\mathbf{A})} = \infty.$$

(c) No, since $\mathbf{C}^*\mathbf{C}$ is a non-negative matrix, and $\mathbf{A}$ is not.

To be precise, observe that for any vector $\mathbf{x}$, we have

$$\mathbf{x}^*\mathbf{C}^*\mathbf{C}\mathbf{x} = (\mathbf{C}\mathbf{x})^*\mathbf{C}\mathbf{x} = \|\mathbf{C}\mathbf{x}\|^2 \geq 0,$$

so $\mathbf{C}^*\mathbf{C}$ is non-negative. But for $\mathbf{x} = [1, -1]^*$, we have

$$\mathbf{x}^*\mathbf{A}\mathbf{x} = e - 2 < 0.$$

Question 3:

(a) Since $r(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$, we find that

$$\frac{\partial r}{\partial x_k} = \sum_{i=1}^{n} a_{ik}x_i + \sum_{j=1}^{n} a_{kj}x_j,$$

and so

$$\nabla r(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}^*\mathbf{x}.$$

(b) The answer is that

$$r(\mathbf{x}_j) = \lambda + O(\|\mathbf{v} - \mathbf{x}_j\|) = \lambda + O(j^{-3}), \qquad \text{as } j \to \infty.$$

To justify this claim, observe first that r is continuous since it is simply a quadratic function in the components of $\mathbf{x}$, so clearly

$$\lim_{j\to\infty} r(\mathbf{x}_j) = r(\lim_{j\to\infty} \mathbf{x}_j) = r(\mathbf{v}) = \mathbf{v}^*\mathbf{A}\mathbf{v} = \mathbf{v}^*(\lambda\mathbf{v}) = \lambda\|\mathbf{v}\|^2 = \lambda.$$

To prove the $O(j^{-3})$ rate, use the bound on the gradient from (a),

$$|r(\mathbf{v}) - r(\mathbf{x}_j)| = |\nabla r(\mathbf{v}) \cdot (\mathbf{v} - \mathbf{x}_j) + O(\|\mathbf{v} - \mathbf{x}_j\|^2) \le (\|\mathbf{A}^*\mathbf{v}\| + \|\mathbf{A}\mathbf{v}\|)\|\mathbf{v} - \mathbf{x}_j\| + O(\|\mathbf{v} - \mathbf{x}_j\|^2) = O(j^{-3}).$$

(c) The answer is that

$$r(\mathbf{x}_j) = \lambda + O(\|\mathbf{v} - \mathbf{x}_j\|^2) = \lambda + O(j^{-6}), \qquad \text{as } j \to \infty.$$

To motivate this, observe that when $\mathbf{A}^* = \mathbf{A}$, we have $\nabla r(\mathbf{v}) = 2\mathbf{A}\mathbf{v} = 2\lambda\mathbf{v}$. This means that the second term in the Taylor expansion, $\nabla r(\mathbf{v}) \cdot (\mathbf{v} - \mathbf{x}_j)$, vanishes *quadratically* as $j \to \infty$, since the vectors $\mathbf{v}$ and $\mathbf{v} - \mathbf{x}_j$ become orthogonal as $\mathbf{x}_j \to \mathbf{v}$ along the surface of the unit sphere. To be precise,

$$\nabla r(\mathbf{v}) \cdot (\mathbf{v} - \mathbf{x}_j) = 2\lambda\mathbf{v} \cdot (\mathbf{v} - \mathbf{x}_j) = \lambda(2\|\mathbf{v}\|^2 - 2\mathbf{v} \cdot \mathbf{x}_j) = \lambda(\|\mathbf{v}\|^2 - 2\mathbf{v} \cdot \mathbf{x}_j + \|\mathbf{x}_j\|^2) = \lambda\|\mathbf{v} - \mathbf{x}_j\|^2,$$

where in the second to last equality we used that $\|\mathbf{v}\| = \|\mathbf{x}_j\|$ (they are both unit length).

Area B - May 2022
Solutions to Part II - CSE 383L
Problems 4 - 7 (shown as 1-4 here)

(1) $u' = f(u)$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array} \qquad \alpha \neq 0.$$

(a) $v_1 = u^n + \alpha \Delta t\, f(u^n)$

$$u^{n+1} = u^n + \Delta t\left[\left(1-\tfrac{1}{2\alpha}\right) f(u^n) + \tfrac{1}{2\alpha} f(v_1)\right]$$

$$= u^n + \Delta t\left[\left(1-\tfrac{1}{2\alpha}\right) f(u^n) + \tfrac{1}{2\alpha} f(u^n + \alpha\Delta t f(u^n))\right]$$

(b) By Taylor's thm

$$f(u^n + \alpha\Delta t f^n) = f^n + (f')^n \alpha \Delta t f^n + \tfrac{1}{2}(f'')^n (\alpha \Delta t f^n)^2 + O(\Delta t^3)$$

$$u' = f,\quad u'' = f'u' = f'f,\quad u''' = (f'f)' = f''f^2 + (f')^2 f$$

$\Rightarrow$

$$f(u^n + \alpha\Delta t f^n) = (u')^n + (u'')^n \alpha\Delta t + \tfrac{1}{2}(f'')^n (f^n)^2 \alpha^2 \Delta t^2 + O(\Delta t^3)$$

So

$$u^{n+1} = u^n + \Delta t\left\{\left(1-\tfrac{1}{2\alpha}\right)(u')^n + \tfrac{1}{2\alpha}\left[(u')^n + (u'')^n \alpha \Delta t + \tfrac{1}{2}(f'')^n (f^n)^2 \alpha^2 \Delta t^2 + O(\Delta t^3)\right]\right\}$$

$$= u^n + \Delta t (u')^n + \tfrac{1}{2}(u'')^n \Delta t^2 + \underbrace{\tfrac{1}{4}\alpha (f'')^n (f^n)^2 \Delta t^3}_{\neq \frac{1}{6}(u''')^n \Delta t^3} + O(\Delta t^4)$$

This is the Taylor expansion of u, up to $O(\Delta t^3)$.

$e = \frac{1}{\sqrt{3}}(1,1,1)$, $\|e\| = 1$.

(2) $u \in H^1$: $(\nabla u, \nabla v) + (e\cdot\nabla u, v) + (u,v) = (f(u), v)$ $\forall v \in H^1$.

(a) IBP:

$(\nabla u, \nabla v) = -(\Delta u, v) + \langle \nabla u \cdot \nu, v\rangle$

$\Rightarrow$ PDE is

$-\Delta u + e\cdot\nabla u + u = f(u)$

BC is $\nabla u \cdot \nu = 0$.

(b) $|f(u) - f(v)| \leq \frac{1}{4}|u-v|$

$(\nabla u_h, \nabla v) + (e\cdot\nabla u_h, v) + (u_h, v) = (f(u_h), v)$

$\Rightarrow$

$(\nabla(u-u_h), \nabla v) + (e\cdot\nabla(u-u_h), v) + (u-u_h, v)$

$= (f(u) - f(u_h), v)$ $\forall v \in V_h$

Let $v \longmapsto v - u_h$ for $v \in V_h$. Then

$(\nabla(u-u_h), \nabla(u-u_h)) + (e\cdot\nabla(u-u_h), u-u_h) + (u-u_h, u-u_h)$

$= (\nabla(u-u_h), \nabla(u-v)) + (e\cdot\nabla(u-u_h), u-v)$

$+ (u-u_h, u-v) + (f(u)-f(u_h), u-u_h)$

$+ (f(u) - f(u_h), v-u)$

$\Rightarrow$

$\|u-u_h\|^2_{H^1} \leq \varepsilon\{\|\nabla(u-u_h)\|^2 + \|u-u_h\|^2\}$

$+ C\{\|\nabla(u-v)\| + \|u-v\|\}$

$+ \frac{1}{4}\|u-u_h\|^2 - \frac{1}{2}\|\nabla(u-u_h)\|^2 - \frac{1}{2}\|u-u_h\|^2$

$\Rightarrow$

$\|u-u_h\|_{H^1} \leq C\|u-v\|_{H^1}$

Take infimum.

Area B - May 2022
Solutions to Part II - CSE 383L
Problems 4 - 7 (shown as 1-4 here)

(3) $u_j^{n+1} = u_j^n - \frac{\Delta t}{h}\left[f(u_{j+1}^{n+1}) - f(u_{j-1}^{n+1})\right]$

Take $f(u) = au$, $\lambda = \frac{a\Delta t}{h}$,

$$\hat{u} = \sum u_j e^{ij\omega}$$

$\Rightarrow$

$$\hat{u}^{n+1} = \hat{u}^n - \lambda\left[\hat{u}^{n+1} e^{i\omega} - \hat{u}^{n+1} e^{-i\omega}\right]$$

$$= \hat{u}^n - 2i\lambda \sin\omega \; \hat{u}^{n+1}$$

$\Rightarrow$

$$\hat{u}^{n+1} = \underbrace{\frac{1}{1+2i\lambda\sin\omega}}_{Q(\omega)} \hat{u}^n$$

Now

$$|Q(\omega)|^2 = \frac{1}{1+4\lambda^2\sin^2\omega} \leq 1.$$

④ $E = [0,h]^2$, $v \in C^1(E)$

(a) By Taylor:

$$v(0,y) = v(x,y) - \int_0^x \frac{\partial v}{\partial x}(\xi, y)\, d\xi$$

$\Rightarrow$

$$\int_0^h |v(0,y)|^2\, dy = \int_0^h \left| v(x,y) - \int_0^x \frac{\partial v}{\partial x}(\xi,y)\, d\xi \right|^2 dy$$

$$\leq 2 \int_0^h \left(|v(x,y)|^2 + \left|\int_0^x \frac{\partial v}{\partial x}(\xi,y)\, d\xi\right|^2 \right) dy$$

[i.e., $(a+b)^2 \leq 2(a^2+b^2)$]

$$\leq 2 \int_0^h |v(x,y)|^2\, dy + 2\int_0^h \int_0^x \left|\frac{\partial v}{\partial x}\right|^2 d\xi \int_0^x 1^2\, d\xi\, dy$$

[by Cauchy Schwartz]

$$\leq 2 \int_0^h |v(x,y)|^2\, dy + 2 \int_0^h \int_0^h |\nabla v|\, dx\, dy\; h$$

Integrate in x over $[0,h]$ to see

$$h \int_0^h |v(0,y)|^2\, dy \leq 2 \int_0^h \int_0^h |v|^2\, dx\, dy + 2 \int_0^h \int_0^h |\nabla v|\, dx\, dy\; h^2$$

$$\Rightarrow \sqrt{\int_0^h |v(0,y)|^2\, dy} \leq C \left\{ h^{-1} \|v\|_{L^2}^2 + h \|\nabla v\|_{L^2}^2 \right\}^{1/2}$$

$$\leq C \left\{ h^{-1/2} \|v\|_{L^2} + h^{1/2} \|\nabla v\|_{L^2} \right\}$$

[i.e., $\sqrt{a^2+b^2} \leq a+b$ when $a,b \geq 0$]

(b) $$\int_E |\nabla v|^2\, dx = \int_{[0,1]^2} h^{-2} |\hat\nabla \hat v|^2 h^2\, d\hat x$$

$$\leq \hat C h^{-2} \int_{[0,1]^2} |\hat v|^2 h^2\, d\hat x \qquad \text{[by equiv. of norms]}$$

$$= \hat C h^{-2} \int_E |v|^2\, dx$$

$\Rightarrow$

$$\|\nabla v\|_{L^2(E)} \leq C h^{-1} \|v\|_{L^2(E)}.$$

CSEM Area B
Part I - CSE 383C
May 2023

EXAM #____

Summer 2023, Numerical Linear Algebra, Preliminary exams

This is a closed-book exam. Please explain all your answers.
A single double-sided cheat sheet is allowed.

Givens rotations and QR factorizations (70 points)

If you haven't read about Givens rotations do not worry. This problem is self contained and uses material covered in the class.

Let $G_2(\theta) \in \mathbb{R}^{2\times 2}$ be the 2D rotation matrix defined by

$$G_2(\theta) := \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}, \text{ where } c := \cos(\theta),\ s := \sin(\theta). \tag{1}$$

For any vector $u \in \mathbb{R}^2$, we can find θ such that $G_2 u = \|u\|_2 e_1$, where e_1 is the first canonical basis vector. In other words, we can use $G_2(\theta)$ to rotate u to align with the horizontal axis.

A generic Givens rotation matrix $G^{(i)}(\theta) \in \mathbb{R}^{n\times n}$ is defined as follows. Let I be the n-by-n identity matrix. We set $G^{(i)} = I$ and then we set the i_{th} 2-by-2 diagonal block of $G^{(i)}$ to $G^{(i)}_{i:i+1,i:i+1} = G_2(\theta)$. That is:

$$G^{(i)} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & c & -s & & \\ & & s & c & & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix}, \text{ where } G^{(i)}_{i,i+1} = -s, \quad G^{(i)}_{i,i} = c, \text{ etc }. \tag{2}$$

1. Show that that $G^{(i)}$ is a unitary matrix.
2. Let $u \in \mathbb{R}^n$. Show that we can find a Givens rotation matrix G such that $v = Gu$ and the $i+1$ element of v satisfies $v_{i+1} = 0$.
3. What is the work complexity and floating point error in computing $v = Gu$ from the above question?
4. Explain what steps you would follow to estimate the floating point error in computing v_i? Please just give a very rough sketch of how you would get to the solution.
5. Assuming no pivoting is used, give and algorithm that uses $n-1$ Givens rotations to compute the QR factorization of a a real square upper Hessenberg matrix.
6. Let $A = I + uv^T$, where $u, v \in \mathbb{R}^n$. Show that there exists a series of Givens rotations $G^{(i)}$, $i = n, \ldots, 2$ that convert u to $\|u\|_2 e_1$ and the matrix A to an upper Hessenberg matrix.
7. Given the results above, given an $O(n^2)$ work and storage algorithm for computing the QR factorization of $I + uv^T$.

8. How does this algorithm compare with the modified Gram-Schmidt in terms of work complexity and accuracy? For accuracy just give one or two sentences based on your analysis for Question 4.

9. Let $A \in \mathbb{R}^{m\times n}$, with $m > n$. Let Q, R be the reduced QR factors of A. Let $B = A + uv^T$ where $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$. Given the Q, R matrices and assuming $u \in \text{span}(A)$, give an $O(mn)$ algorithm to compute the QR factorization of B.

10. Now relax the assumption that $u \in \text{span}(A)$. Given the reduced QR factors of A, suggest an $O(mn)$ algorithm for computing the QR factorization of B.

Review questions (30 points)

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix.

1. Suggest a dense direct solver for solving for $Ax = b$ and state its work complexity.

2. Suggest an iterative solver for solving $Ax = b$ and state its work complexity assuming the cost a matrix-vector multiplication with A is $O(n)$.

3. Assume that A is sparse, numerically rank deficient, b is polluted by noise and n is very large. Is solving $Ax = b$ possible? Suggest a solver for finding an exact or an approximate solution to $Ax = b$.

4. Suppose we're interested in computing the 10 largest eigenvalues of A. Suggest and algorithm to compute them.

5. Suppose we're interested in computing the 10 smallest eigenvalues of A. Suggest and algorithm to compute them.

CSEM Area B
Part II - CSE 383L
May 2023

EXAM #____

Prelim, CSEM, Area B, Numerical Differential equations

1. Let the ordinary differential equation, $u' = f(u)$ be approximated by,

$$u_{n+1} = u_n + h\theta f(u_n) + h(1-\theta)f(u_{n+1}), \qquad 0 \le \theta \le 1$$

(a) Investigate the order of the local truncation error (the order of the convergence of the method in terms of powers of h) for different values of θ.

(b) Rewrite the algorithm on the standard form of Runge-Kutta methods after replacing u_{n+1} by $u_n + hf(u_n)$ and determine if the algorithm is 0-stable (Dahlquist stable).

(c) Assume $f(u) = \lambda u$ and determine if $h\lambda = -1$ is in the region of absolute stability for the original algorithm with $\theta = 1$. Is there a value of θ for which the method after the modification given in (b) above is A-stable.

2. (a) Rewrite the strong form the elliptic PDE,

$$-\nabla \cdot a(x)\nabla u + b(x)u = f(x), \qquad x \in \Omega \subset R^2, a(x) \ge a > 0, b(x) \ge b > 0,$$

$$\frac{\partial u}{\partial n} = \alpha u + \beta, x \in \partial\Omega$$

on weak (variational) form suitable for a FEM formulation. Define coercivity and continuity of the related bilinear and linear forms.

(b) Investigate coercivity for the bilinear form.

(c) Redo (a) when $\Omega = \sum_n \Omega_n$ with the purpose of applying Discontinuous Galerkin (DG). Include interface terms but you do not need to elaborate on numerical fluxes. Also determine the number of degrees of freedom if a square domain Ω in divided into 4 squares (2 by 2) with Q1 (bilinear) elements for the DG setting and also the same for conforming elements in FEM.

3. Consider the initial boundary value problem,

$$u_t + f(u)_x = \varepsilon u_{xx}, \qquad 0 < x < 1, t > 0, f'(u) > 0, \varepsilon > 0$$
$$u(x,0) = u_0(x), \;\; 0 < x < 1, \;\text{periodic bondary conditions.}$$

(a) Construct an explicit finite volume scheme based on upwind differencing of the nonlinear term and centered differencing of the second order term. (It can also be seen as finite difference scheme on conservation form.)

(b) For $f(u) = 0$, use von Neumann analysis to investigate L_2 stability.

(c) For $\varepsilon = 0$, give conditions on the step sizes such that the Lax-Friedrich scheme below is monotone,

$$u_j^{n+1} = \left(u_{j+1}^n + u_{j-1}^n\right)/2 + \left(\frac{\Delta t}{2\Delta x}\right)\left(f\left(u_{j+1}^n\right) - f\left(u_{j-1}^n\right)\right).$$

EXAM #____

Area B Prelim Exam - Data Science

Date: May 12, 2023

Instructions: *For this portion of the Area B Prelim Exam, you are allowed to use two double-sided A4-sized "cheat sheets" (typed or handwritten) that you prepared yourself containing material from CSE 382M. This portion of the exam is intended to take about 90 minutes. Write legibly and* ***give sufficient justification for your answers.*** *If there is a part of a question you cannot answer, you may assume it and proceed to the next part of the question. The maximum possible total score is 60 points.* ***Print your prelim exam number above.***

Problem	**Points**	**Max**
1		20
2		20
3		20
Total		60

Problem 1. *Clustering and kernel methods*

We are given a dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq \mathbb{R}^d$. Fix a positive integer k. Recall Lloyd's algorithm for clustering $\mathcal{X}$ into k parts:

- initialize a partition of $\mathcal{X}$ as $C_1^{(0)} \sqcup \ldots \sqcup C_k^{(0)}$
- compute means $\mu_\alpha^{(t)} = \frac{1}{|C_\alpha^{(t)}|} \sum_{\mathbf{x}_j \in C_\alpha^{(t)}} \mathbf{x}_j$ for $\alpha = 1, \ldots, k$
- compute clusters $C_\alpha^{(t+1)} = \left\{\mathbf{x}_i \in \mathcal{X} : \left\|\mathbf{x}_i - \mu_\alpha^{(t)}\right\|_2 \leq \left\|\mathbf{x}_i - \mu_\beta^{(t)}\right\|_2 \text{ for all } \beta\right\}$ for $\alpha = 1, \ldots, k$
- repeat the two steps above

We assume throughout this problem that we encounter no degenerate behavior, i.e. that there are never ties in minimal distances and no part becomes empty.

(a) [6 pts] Let $t \geq 1$. Explain why the convex hulls of $C_\alpha^{(t)}$ for $\alpha = 1, \ldots, k$ must be disjoint.

(Hint: How does $C_\alpha^{(t)}$ relate to the Voronoi cell of $\mu_\alpha^{(t-1)}$?)

(b) [2 pts] Draw a dataset $\mathcal{X}$ in the plane $\mathbb{R}^2$ where there are obviously two "natural" clusters, yet Lloyd's algorithm could never converge to that partition.

(c) **[4 pts]** Justify the formula

$$\left\|\mathbf{x}_i - \mu_\alpha^{(t)}\right\|_2^2 \;=\; \langle \mathbf{x}_i, \mathbf{x}_i \rangle \;-\; \frac{2}{|C_\alpha^{(t)}|} \sum_{\mathbf{x}_j \in C_\alpha^{(t)}} \langle \mathbf{x}_i, \mathbf{x}_j \rangle \;+\; \frac{1}{|C_\alpha^{(t)}|^2} \sum_{\mathbf{x}_j, \mathbf{x}_{j'} \in C_\alpha^{(t)}} \langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle.$$

(d) [6 pts] Let $\varphi : \mathbb{R}^d \to \mathbb{R}^D$ be a feature map with $D \gg d$. Describe a variant of Lloyd's algorithm that applies to the points $\varphi(\mathcal{X}) = \{\varphi(\mathbf{x}_1), \ldots, \varphi(\mathbf{x}_n)\} \subseteq \mathbb{R}^D$, but that avoids any explicit operations on D-dimensional vectors. Assume the kernel matrix $(K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ has been precomputed, where $K(\mathbf{x}, \mathbf{y}) := \langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle$.
(Hint: Apply part (c) appropriately in $\mathbb{R}^D$.)

(e) [2 pts] In the kernel-based variant of Lloyd's algorithm, *roughly* how many flops does each iteration cost? Express your answer in big $\mathcal{O}$ notation.

END OF PROBLEM 1

Problem 2. *Gradient descent and a simple neural network*

(a) Suppose $\ell : \mathbb{R}^d \to \mathbb{R}$ is a C^2 loss function bounded below by $\ell_* \in \mathbb{R}$. To minimize ℓ we can use gradient descent with constant step size $\alpha > 0$:

- initialize $\mathbf{w}_0 \in \mathbb{R}^d$
- set $\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla \ell(\mathbf{w}_t)$
- repeat the step above

Assume there is a constant $C > 0$ such that the Hessian $\nabla^2 \ell(\mathbf{w})$ has maximal eigenvalue upper-bounded by C for each $\mathbf{w} \in \mathbb{R}^d$. Set the step size as $\alpha = \frac{1}{C}$.

(i) **[6 pts]** Prove $\ell(\mathbf{w}_{t+1}) \leq \ell(\mathbf{w}_t) - \frac{1}{2C} \|\nabla \ell(\mathbf{w}_t)\|_2^2$.

(Hint: Use Taylor's theorem, $\ell(\mathbf{w}_{t+1}) = \ell(\mathbf{w}_t) + \nabla \ell(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{1}{2}(\mathbf{w}_{t+1} - \mathbf{w}_t)^\top \nabla^2 \ell(\mathbf{w}_*)(\mathbf{w}_{t+1} - \mathbf{w}_t)$ for some $\mathbf{w}_*$ on the line segment between $\mathbf{w}_{t+1}$ and $\mathbf{w}_t$, with the bound on the maximal eigenvalue of $\nabla^2 \ell(\mathbf{w}_*)$.)

(ii) **[6 pts]** Deduce gradient descent converges in the sense that

$$\min_{t=0,\ldots,T-1} \|\nabla \ell(\mathbf{w}_t)\|_2^2 \quad \leq \quad \frac{2C\big(\ell(\mathbf{w}_0) - \ell_*\big)}{T}.$$

(Hint: Rearrange $\ell(\mathbf{w}_{t+1}) \leq \ell(\mathbf{w}_t) - \frac{1}{2C}\|\nabla \ell(\mathbf{w}_t)\|_2^2$, and sum over t.)

(b) In a classification task, we are given 1D training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \{0, 1\}$. We want to predict the probability of being classified as 1 given x. We consider a simple one-neuron model:

$$f(w, x) = \sigma(wx)$$

where $w \in \mathbb{R}$ is a trainable parameter and $\sigma(s) = \frac{e^s}{1+e^s}$ is the sigmoidal activation function. Use the following training loss, which is a negative log-likelihood:

$$\ell(w) \;=\; -\sum_{k:y_k=1} \log(\sigma(wx_k)) \;-\; \sum_{k:y_k=0} \log(1 - \sigma(wx_k)). \qquad (\star)$$

(i) **[4 pts]** Calculate the second derivative $\ell''(w)$.

(ii) **[4 pts]** Give a constant step size $\alpha > 0$ so that gradient descent applied to the loss $(\star)$ is guaranteed to converge at a rate of $\mathcal{O}(1/T)$ as in (a)(ii). *(Hint: Your answer should depend on $x_1, \ldots, x_n$.)*

END OF PROBLEM 2

Problem 3. *Randomized projections and linear subspaces*

Let $\mathcal{L} \subseteq \mathbb{R}^d$ be a fixed subspace of dimension k passing through the origin. Let $\delta \in (0,1)$ and $\eta \in (0,1)$. Over the course of this problem, you will prove that if

$$m \geq \frac{ck\log(12/\delta) + c\log(3/\eta)}{\delta^2}$$

and $\Phi \in \mathbb{R}^{m\times d}$ is a Gaussian matrix with i.i.d. $\mathcal{N}(0,1/m)$-entries, then the event

$$(1-\delta)\|\mathbf{x}\|_2 \;\leq\; \|\Phi\mathbf{x}\|_2 \;\leq\; (1+\delta)\|\mathbf{x}\|_2 \quad \text{for all } \mathbf{x} \in \mathcal{L} \tag{$\star$}$$

holds with probability at least $1-\eta$ (where $c > 0$ is a universal constant).

(a) [2 pts] Let $\mathbb{S} = \{\mathbf{x} \in \mathcal{L} : \|\mathbf{x}\|_2 = 1\}$ be the $(k-1)$-dimensional unit sphere in $\mathcal{L}$. Explain why $(\star)$ is equivalent to $1-\delta \leq \|\Phi\mathbf{x}\|_2 \leq 1+\delta$ for all $\mathbf{x} \in \mathbb{S}$.

(b) [4 pts] Let $\varepsilon \in (0,1)$ and $\gamma > 0$ (to be specified later). We introduce the concept of an ε-net on the sphere. This is a subset $\mathcal{N} \subseteq \mathbb{S}$ such that for all $\mathbf{x} \in \mathbb{S}$ there exists $\mathbf{y} \in \mathcal{N}$ with $\|\mathbf{x}-\mathbf{y}\|_2 \leq \varepsilon$. You may assume that we can fix $\mathcal{N}$ to have size $|\mathcal{N}| \leq (3/\varepsilon)^k$. Using the random projection theorem and a union bound, give a nontrivial lower bound on the probability of the following event:

$$1-\gamma \;\leq\; \|\Phi\mathbf{y}\|_2 \;\leq\; 1+\gamma \quad \text{for all } \mathbf{y} \in \mathcal{N}. \tag{$\star\star$}$$

(Hint: The random projection theorem says that for fixed $\mathbf{y} \in \mathbb{S}$ *we have* $\mathbb{P}\big(\big|\|\Phi\mathbf{y}\|_2 - 1\big| \geq \gamma\big) \;\leq\; 3\exp(-Cm\gamma^2)$ *where* $C > 0$ *is a universal constant.)*

(c) [4 pts] Let $\sigma_{\max} := \max_{\mathbf{x}\in\mathbb{S}} \|\Phi\mathbf{x}\|_2$. Prove $\sigma_{\max} \le 1 + \gamma + \varepsilon\sigma_{\max}$ so long as the event $(\star\star)$ holds.

(Hint: Let $\sigma_{\max}$ be attained at $\mathbf{x}_{\max} \in \mathbb{S}$. Choose $\mathbf{y} \in \mathcal{N}$ so that $\|\mathbf{x}_{\max} - \mathbf{y}\|_2 \le \varepsilon$. Then write $\Phi\mathbf{x}_{\max} = \Phi\mathbf{y} + \Phi(\mathbf{x}_{\max} - \mathbf{y})$. How can you bound the norm?)

(d) [2 pts] From part (c), deduce $\sigma_{\max} \le 1 + \frac{\varepsilon+\gamma}{1-\varepsilon}$ if event $(\star\star)$ holds.

(e) [4 pts] Let $\sigma_{\min} := \min_{\mathbf{x}\in\mathbb{S}} \|\Phi\mathbf{x}\|_2$ be attained at $\mathbf{x}_{\min} \in \mathbb{S}$. Similarly to the above, show if $(\star\star)$ holds then $\sigma_{\min} \geq 1 - \gamma - \sigma_{\max}\varepsilon$. Deduce $\sigma_{\min} \geq 1 - \frac{\varepsilon+\gamma}{1-\varepsilon}$.

(f) [4 pts] Finish the proof of the theorem about $(\star)$ at the start of the problem. *(Hint: Choose $\varepsilon = \frac{\delta}{4}$ and $\gamma = \frac{\delta}{2}$. Verify $\frac{\varepsilon+\gamma}{1-\varepsilon} \leq \delta$. How big should m be so the probability from part (b) is at least $1-\eta$?)*

END OF EXAM

Area B Prelim - PART ONE - CSE 383C

Summer 2024, Area B: Numerical Linear Algebra, Preliminary exam

This is a closed-book exam. Please explain all your answers.
A single double-sided cheat sheet is allowed.

1. [**12 points.**] Let

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

   (a) What are the singular values of A?

   (b) What is a basis for $\mathsf{range}(A)$?

   (c) What is a basis for $\mathsf{range}(A^T)$?

2. [**12 points.**] Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ be given. Considering storage and computation time, suggest an algorithm for solving $\min_x \|Ax - b\|_2$ for the following cases:

   (a) $m = n$, A is full rank dense.

   (b) $m = n$, A is full rank sparse symmetric.

   (c) $m > n$, $n = O(1)$, A is full rank dense.

   (d) $m > n$, A has unknown rank.

   (e) $m > n$, A is sparse with unknown rank.

   (f) $m < n$, $m = O(1)$, A is full rank.

3. [**6 points.**] Consider $A \in \mathbb{R}^{n \times n}$, and $A = LU$ is its unpivoted LU factorization. Propose an algorithm to compute the (i, j) entry of A^{-1} in $O((n-j)^2+(n-1)^2)$ floating-point operations.

4. [**15 points.**] Determine whether the following statements are true or false. Assume that we use IEEE arithmetic, that a, b, and c are normalized floating point numbers and that no exceptions occur at the stated operations.

   (a) $\text{round}\{a + b\} = \text{round}\{b + a\}$.
   (b) $\text{round}\{b - a\} = -\text{round}\{a - b\}$.
   (c) $\text{round}\{a + a\} = \text{round}\{2a\}$.
   (d) $\text{round}\{(a + b) + c\} = \text{round}\{a + (b + c)\}$.
   (e) $a \leq \text{round}\{(a + b)/2\} \leq b$, assuming $a \leq b$.

5. [**10 points.**] Let

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

We want to solve $Ax = b$ with a matrix splitting scheme in which $A = M - N$, and $M =$ the lower triangular part of A.

(a) Does this iterative scheme converge?

(b) Assume $b = 0$ and $x_0 \neq 0$ be the initial guess. Let e_0 the initial error. How many iterative steps k do we need to approximately have $\|e_k\|_\infty / \|e_0\|_\infty \leq 10^{-5}$?

6. [**25 points.**] Let $A \in \mathbb{R}^{n \times n}$ with $A = A^T$. Let $\mathsf{dim}(\mathsf{null}(A)) = 1$ and $Av = 0$ with $\|v\|_2 = 1$.

   (a) Propose an iterative method for solving $\min_x \|Ax - b\|$ that does not deteriorate the inherent conditioning of the problem.

   (b) What the expected perturbation in the solution x given perturbations in b?

   (c) What is the work complexity of the method assuming A is sparse with $O(n)$ non-zero values?

7. [**5 points.**] Let $v \in \mathbb{R}^n$ with $\|v\| = 1$. What are the eigenvalues and eigenvectors of $H = I - 2uu^T$?

8. [**15 points.**] Let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ be the eigenvalues of an $n \times n$ real symmetric matrix A.

   (a) To which of the eigenvalues of A is possible for the power method to converge by using an appropriately chosen shift σ?

   (b) In each such case, what value for the shift gives the most rapid convergence?

   (c) Answer the same two questions for the inverse iteration method.

CSEM Area B
Part II - CSE 382M
May 2024
(w/ solutions)

EXAM #: ____________________

Area B Prelim Exam - Data Science

Date: May 17, 2024

Instructions*: For this portion of the Area B Prelim Exam, you are allowed to use one double-sided A4-sized "cheat sheet" (typed or handwritten) that you prepared yourself containing material from CSE 382M. This portion of the exam is intended to take about 90 minutes. Write legibly and* ***give sufficient justification for your answers.*** *If there is a part of a question you cannot answer, you may assume it and proceed to the next part of the question. The maximum possible total score is 60 points.* ***Print your prelim exam number above.***

Problem	**Points**	**Max**
1		20
2		20
3		20
Total		60

Problem 1. *Spectral clustering and kernel methods*

We are given a dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$. Fix a positive integer k. Recall the k-means objective as a function of a partition $C_1 \sqcup \ldots \sqcup C_k = [n]$ is

$$f(C_1, \ldots, C_k) \;:=\; \sum_{i=1}^{k} \sum_{j \in C_i} \|\mathbf{x}_j - \frac{1}{|C_i|} \sum_{j' \in C_i} \mathbf{x}_{j'}\|_2^2,$$

where C_i are assumed to be nonempty.

(a) [6 pts] Show that

$$f(C_1, \ldots, C_k) \;=\; \sum_{j=1}^{n} \|\mathbf{x}_j\|_2^2 \;-\; \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{j,j' \in C_i} \langle \mathbf{x}_{j'}, \mathbf{x}_{j''} \rangle.$$

Expanding the squares and summing up the three terms, compute

$$\begin{aligned}
f(C_1, \ldots, C_k) &= \sum_{i=1}^{k} \sum_{j \in C_i} \Big(\|\mathbf{x}_j\|_2^2 - \frac{2}{|C_i|} \sum_{j' \in C_i} \langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle + \frac{1}{|C_i|^2} \sum_{j',j'' \in C_i} \langle \mathbf{x}_{j'}, \mathbf{x}_{j''} \rangle \Big) \\
&= \sum_{j=1}^{n} \|\mathbf{x}_j\|_2^2 - \sum_{i=1}^{k} \frac{2}{|C_i|} \sum_{j,j' \in C_i} \langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle + \sum_{i=1}^{k} \frac{|C_i|}{|C_i|^2} \sum_{j',j'' \in C_i} \langle \mathbf{x}_{j'}, \mathbf{x}_{j''} \rangle \\
&= \sum_{j=1}^{n} \|\mathbf{x}_j\|_2^2 \;-\; \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{j,j' \in C_i} \langle \mathbf{x}_{j'}, \mathbf{x}_{j''} \rangle.
\end{aligned}$$

Here we renamed the dummy indices j', j'' as j, j' in the rightmost sum in the second line to merge with the middle sum there.

(b) [2 pts] Deduce that the k-means problem

$$\min_{C_1 \sqcup \ldots \sqcup C_k = [n]} f(C_1, \ldots, C_k)$$

is equivalent to

$$\max_{C_1 \sqcup \ldots \sqcup C_k = [n]} \langle \mathbf{G}_{\mathcal{X}}, \mathbf{Z}_{C_1, \ldots, C_k} \rangle,$$

where $\mathbf{G}_{\mathcal{X}} \in \mathbb{R}^{n \times n}$ is the Gram matrix of $\mathcal{X}$ with (j, j') entry given by $\langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle$, and $\mathbf{Z}_{C_1, \ldots, C_k} \in \mathbb{R}^{n \times n}$ is the so-called weighted adjacency matrix of the partition with (j, j') entry given by $\frac{1}{|C_i|}$ if $\mathbf{x}_j, \mathbf{x}_{j'}$ are in the same cluster and 0 otherwise.

We use the result from part (a). Since $\sum_{j=1}^{n} \|\mathbf{x}_j\|_2^2$ is independent of the partition, and the other sum is being subtracted in the formula in part (a), minimizing $f(C_1, \ldots, C_k)$ is equivalent to maximizing $\sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{j,j' \in C_i} \langle \mathbf{x}_{j'}, \mathbf{x}_{j''} \rangle$.

The latter cost function equals $\sum_{j,j'=1}^{n} (\mathbf{G}_{\mathcal{X}})_{j,j'} (\mathbf{Z}_{C_1,\ldots,C_k})_{j,j'}$ *by definition of the Gram matrix and weighted adjacency matrix, hence* k*-means is equivalent to*

$$\max_{C_1 \sqcup \ldots \sqcup C_k = [n]} \langle \mathbf{G}_{\mathcal{X}}, \mathbf{Z}_{C_1,\ldots,C_k} \rangle.$$

(c) [6 pts] Explain how the formulation in part (b) is relaxed to the following:

$$\max_{\substack{\mathbf{Q} \in \mathbb{R}^{n \times k} \\ \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_k}} \langle \mathbf{G}_{\mathcal{X}}, \mathbf{Q}\mathbf{Q}^\top \rangle.$$

What is the solution $\mathbf{Q}_*$ to the relaxation? *(A proof for this second sentence is not needed, if the relevant fact from linear algebra is stated clearly.)* How can we use Lloyd's algorithm to obtain from $\mathbf{Q}_*$ a partition $C_1 \sqcup \ldots \sqcup C_k$ of $[n]$?

Define $\mathbf{Q}_{C_1,\ldots,C_k} \in \mathbb{R}^{n \times k}$ *to have* (j,i) *entry* $\frac{1}{\sqrt{|C_i|}}$ *if* $\mathbf{x}_j \in C_i$ *and* 0 *otherwise, and notice* $\mathbf{Z}_{C_1,\ldots,C_k} = \mathbf{Q}_{C_1,\ldots,C_k} \mathbf{Q}_{C_1,\ldots,C_k}^\top$ *and* $\mathbf{Q}_{C_1,\ldots,C_k}^\top \mathbf{Q}_{C_1,\ldots,C_k} = \mathbf{I}_k$*. Therefore* $\mathbf{Z}_{C_1,\ldots,C_k}$ *is an orthogonal projector onto a rank-*k *linear subspace. Though* $\mathbf{Z}_{C_1,\ldots,C_k}$ *satisfies other (combinatorial) constraints, we drop them to obtain a larger search space and so a relaxation of* k*-means:*

$$\max_{\substack{\mathbf{Q} \in \mathbb{R}^{n \times k} \\ \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_k}} \langle \mathbf{G}_{\mathcal{X}}, \mathbf{Q}\mathbf{Q}^\top \rangle.$$

The relaxation is computationally tractable because its solution $\mathbf{Q}_*$ *is given by the leading* k *eigenvectors of* $\mathbf{G}_{\mathcal{X}}$*. This holds since for any symmetric matrix* $\mathbf{M} \in \mathbb{R}^{n \times n}$*, the leading* k *eigenvectors of* $\mathbf{M}$ *solve* $\max_{\mathbf{Q} \in \mathbb{R}^{n \times k}, \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_k} \langle \mathbf{M}, \mathbf{Q}\mathbf{Q}^\top \rangle$.

The unrelaxed solution $\mathbf{Q}_{C_1,\ldots,C_k}$ *has rows which are scaled indicator functions of the clusters. Therefore it makes sense to cluster the rows of the relaxed solution* $\mathbf{Q}_*$ *using Lloyd's algorithm, to obtain a valid clustering from* $\mathbf{Q}_*$*. In particular, we run Lloyd's algorithm in* $\mathbb{R}^k$ *rather than* $\mathbb{R}^d$*.*

(d) [6 pts] For clusters with complicated geometry, in theory it is advantageous to use a feature map $\varphi : \mathbb{R}^d \to \mathbb{R}^D$, and apply the procedure in part (c) to the dataset $\varphi(\mathcal{X}) = \{\varphi(\mathbf{x}_1), \ldots, \varphi(\mathbf{x}_n)\} \subset \mathbb{R}^D$. If $D \gg d$, explain why a naive implementation of this is impractical. Describe how an efficient implementation could be achieved, at least for convenient choices of φ. Sketch what flop and storage costs might be achieved in the efficient version. *(Hint: kernel methods.)*

A naive implementation would involve operations on D*-dimensional vectors in* $\varphi(\mathcal{X})$*, which is impractical if* $D \gg d$ *(especially so if* $D = \infty$*). However, the procedure in part (b) only needs access to the Gram matrix of the dataset.*

Thus, to run it on $\varphi(\mathcal{X})$ we just need $\mathbf{G}_{\varphi(\mathcal{X})} \in \mathbb{R}^{n\times n}$ given by $(\mathbf{G}_{\varphi(\mathcal{X})})_{j,j'} = \langle \varphi(\mathbf{x}_j), \varphi(\mathbf{x}'_j)\rangle$. For nice feature maps φ, the kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $K(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle$ can be evaluated in $\mathcal{O}(d)$ flops and bytes. In this case, the Gram matrix costs $\mathcal{O}(n^2 d)$ flops to form and $\mathcal{O}(n^2)$ space to store, which greatly improve over naive methods. For the rest, it costs $\mathcal{O}(n^2 k)$ flops to compute the top k-eigenvectors from $\mathbf{G}_{\varphi(\mathcal{X})}$, and $\mathcal{O}(nk^2)$ flops per Lloyd iteration.

Problem 2. *Gradient descent and the PL inequality*

(a) [6 pts] Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a C^2 loss function which is bounded below. To minimize f we can use gradient descent with constant step size $\alpha > 0$:

- initialize $\mathbf{w}_0 \in \mathbb{R}^d$
- set $\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla f(\mathbf{w}_t)$
- repeat the step above.

Assume there is a constant $L > 0$ such that $\nabla^2 f(\mathbf{w})$ has maximal eigenvalue upper-bounded by L for all $\mathbf{w} \in \mathbb{R}^d$. Set the step size as $\alpha = \frac{1}{L}$. Please prove

$$f(\mathbf{w}_{t+1}) \ \le \ f(\mathbf{w}_t) - \frac{1}{2L}\|\nabla f(\mathbf{w}_t)\|_2^2.$$

(Hint: Taylor's theorem.)

Taylor's theorem for twice continuously differentiable functions states

$$f(\mathbf{w}_{t+1}) = f(\mathbf{w}_t) + (\mathbf{w}_{t+1} - \mathbf{w}_t)^\top \nabla f(\mathbf{w}_t) + \frac{1}{2}(\mathbf{w}_{t+1} - \mathbf{w}_t)^\top \nabla^2 f(\bar{\mathbf{w}})(\mathbf{w}_{t+1} - \mathbf{w}_t)$$

for some $\bar{\mathbf{w}} \in \mathbb{R}^d$ on the line segment connecting $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$. Plugging in the spectral bound on the Hessians and the definition of $\mathbf{w}_{t+1}$ yields

$$\begin{aligned}
f(\mathbf{w}_{t+1}) \ &\le \ f(\mathbf{w}_t) + (\mathbf{w}_{t+1} - \mathbf{w}_t)^\top \nabla f(\mathbf{w}_t) + \frac{1}{2}L\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&= \ f(\mathbf{w}_t) + \big(-\tfrac{1}{L}\nabla f(\mathbf{w}_t)\big)^\top \nabla f(\mathbf{w}_t) + \frac{1}{2}L\big\| -\tfrac{1}{L}\nabla f(\mathbf{w}_t)\big\|_2^2 \\
&= \ f(\mathbf{w}_t) - \frac{1}{2L}\|\nabla f(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

(b) [6 pts] In addition, assume there is a constant $\mu \in (0, L]$ such that for all $\mathbf{w} \in \mathbb{R}^d$,

$$\frac{1}{2}\|\nabla f(\mathbf{w})\|_2^2 \ \ge \ \mu(f(\mathbf{w}) - f_*),$$

where $f_* \in \mathbb{R}$ is the minimum value of f. Prove that gradient descent must converge *linearly* to the global minimum, i.e.,

$$f(\mathbf{w}_t) - f_* \;\leq\; \kappa^t (f(\mathbf{w}_0) - f_*),$$

for some constant $\kappa \in [0, 1)$. Please express κ in terms of L and μ.

Inserting the Polyak-Łojasiewicz inequality into the result from part (a),

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t) - \frac{\mu}{L}(f(\mathbf{w}_t) - f_*).$$

Subtracting f_* from both sides and factorizing gives

$$f(\mathbf{w}_{t+1}) - f_* \;\leq\; f(\mathbf{w}_t) - \frac{\mu}{L}(f(\mathbf{w}_t) - f_*) - f_* \;=\; (1 - \frac{\mu}{L})(f(\mathbf{w}_t) - f_*).$$

Recursing on t we have

$$f(\mathbf{w}_t) - f_* \;\leq\; (1 - \frac{\mu}{L})^t (f(\mathbf{w}_0) - f_*).$$

This shows linear convergence to the global minimum with rate $\kappa = 1 - \frac{\mu}{L}$.

(c) [6 pts] As an example, consider overdetermined least squares regression:

$$\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) \qquad \text{with} \quad f(\mathbf{w}) = \tfrac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2,$$

where $\mathbf{X} \in \mathbb{R}^{m \times d}$, $\mathbf{y} \in \mathbb{R}^m$ and $m \gg d$. Is there a constant L satisfying the L-Lipschitz gradient condition in part (a)? If so, what is L? Is there a constant μ satisfying the PL inequality condition in part (b)? If so, what is μ? Please justify your answers.

We compute expressions for the function $f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{y}^\top \mathbf{X}\mathbf{w} + \|\mathbf{y}\|_2^2$, the gradient $\nabla f(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y}$ and the Hessian $\nabla^2 f(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}$. Because the Hessian is independent of $\mathbf{w}$, clearly there exists an L as in part (a). We take it to be the largest eigenvalue of $\mathbf{X}^\top \mathbf{X}$, i.e., $L = \lambda_{\max}(\mathbf{X}^\top \mathbf{X})$ if $\mathbf{X} \neq 0$. For the PL inequality, we note that $\nabla f(\mathbf{w}) = \mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$ implies

$$\|\nabla f(\mathbf{w})\|_2 \geq \sigma_{\min}(\mathbf{X}^\top)\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2,$$

where $\sigma_{\min}$ denotes the minimal singular value. Therefore, $\frac{1}{2}\|\nabla f(\mathbf{w})\|_2^2 \geq \frac{1}{2}\sigma_{\min}(\mathbf{X}^\top)^2 \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = \lambda_{\min}(\mathbf{X}^\top \mathbf{X}) f(\mathbf{w})$. Hence so long as $\mathbf{X}$ is full-rank, there exists μ as in part (b). In this case we can take it to be the smallest eigenvalue of $\mathbf{X}^\top \mathbf{X}$, i.e., $\mu = \lambda_{\min}(\mathbf{X}^\top \mathbf{X})$.

(d) [2 pts] In the context of the example in part (c), what would be stochastic gradient descent with batch size b? What is the computational cost of an SGD iteration compared to that of a GD iteration? *(An answer in big $\mathcal{O}$ notation is fine.)*

Stochastic gradient descent would randomly select one equation or a mini-batch of b equations from the overdetermined linear least squares from part (c) to determine each gradient step (possibly over multiple epochs). Here equations correspond to the rows of $\mathbf{X}$, and the selection of b equations corresponds to selecting a submatrix $\tilde{\mathbf{X}} \in \mathbb{R}^{b\times d}$ of $\mathbf{X}$. The stochastic gradient would be

$$\tilde{\nabla} f(\mathbf{w}) = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\mathbf{w} - \tilde{\mathbf{X}}^\top \mathbf{y},$$

which consists of three mat-vec multiplications. It costs $\mathcal{O}(bd)$ flops to evaluate, and dominates the SGD update $\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha\tilde{\nabla} f(\mathbf{w}_t)$. Meanwhile, the full GD iteration is the same but with $\mathbf{X}$ instead of $\tilde{\mathbf{X}}$. Therefore it costs $\mathcal{O}(md)$ flops.

Problem 3. *MCMC sampling and Metropolis-Hastings*

(a) [2 pts] Consider a strongly connected Markov chain on n states with transition matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ and initial distribution $\mathbf{p}(\mathbf{0}) \in \mathbb{R}^{1\times n}$. Let $\mathbf{p}(\mathbf{t}) \in \mathbb{R}^{1\times n}$ denote the distribution of the state after t steps. What is $\mathbf{p}(\mathbf{t})$ in terms of $\mathbf{P}$ and $\mathbf{p}(\mathbf{0})$? *(No justification needed.)*

The relation is $\mathbf{p}(\mathbf{t}) = \mathbf{p}(\mathbf{0})\mathbf{P}^t$.

(b) [6 pts] Define the running average by $\mathbf{a}(\mathbf{t}) = \frac{1}{t}(\mathbf{p}(\mathbf{0}) + \ldots + \mathbf{p}(\mathbf{t-1}))$. Show

$$\|\mathbf{a}(\mathbf{t})\mathbf{P} - \mathbf{a}(\mathbf{t})\|_1 \leq \frac{2}{t}.$$

(Remark: In particular, this is a key step in seeing that the Markov chain has a unique stationary distribution π, i.e., one such that $\pi = \pi\mathbf{P}$.)

From part (a) note $\mathbf{p}(\mathbf{k}) = \mathbf{p}(\mathbf{k-1})\mathbf{P}$. Therefore we get a telescoping sum:

$$\begin{aligned}
\mathbf{a}(\mathbf{t})\mathbf{P} - \mathbf{a}(\mathbf{t}) &= \tfrac{1}{t}(\mathbf{p}(\mathbf{0}) + \ldots + \mathbf{p}(\mathbf{t-1}))\mathbf{P} - \tfrac{1}{t}(\mathbf{p}(\mathbf{0}) + \ldots + \mathbf{p}(\mathbf{t-1})) \\
&= \tfrac{1}{t}(\mathbf{p}(\mathbf{1}) + \mathbf{p}(\mathbf{2}) + \ldots + \mathbf{p}(\mathbf{t})) - \tfrac{1}{t}(\mathbf{p}(\mathbf{0}) + \ldots + \mathbf{p}(\mathbf{t-1})) \\
&= \tfrac{1}{t}(\mathbf{p}(\mathbf{t}) - \mathbf{p}(\mathbf{0})).
\end{aligned}$$

The triangle inequality implies

$$\|\mathbf{a(t)P} - \mathbf{a(t)}\|_1 = \|\tfrac{1}{t}(\mathbf{p(t)} - \mathbf{p(0)})\|_1 \leq \tfrac{1}{t}(\|\mathbf{p(t)}\|_1 + \|\mathbf{p(0)}\|_1) = \frac{2}{t},$$

where in the last equality we used that $\mathbf{p(t)}$ *and* $\mathbf{p(0)}$ *are probability mass functions and hence have* ℓ_1 *norms of* 1.

(c) [**4 pts**] Next suppose a given distribution $\mathbf{p} \in \mathbb{R}^{1\times n}$ obeys

$$\mathbf{p}_i \mathbf{P}_{ij} = \mathbf{p}_j \mathbf{P}_{ji}$$

for all i, j. Explain why $\mathbf{p}$ must be the unique stationary distribution.

Taking the remark in part (b) for granted, there exists a unique stationary distribution for a strongly connected Markov chain. Therefore we just need to check that $\mathbf{p}$ *satisfies stationarity. Summing both sides of* $\mathbf{p}_i\mathbf{P}_{ij} = \mathbf{p}_j\mathbf{P}_{ji}$ *over* j,

$$\sum_{j=1}^{n} \mathbf{p}_i \mathbf{P}_{ij} = \sum_{j=1}^{n} \mathbf{p}_j \mathbf{P}_{ji}.$$

The left-hand side of the above equals $\mathbf{p}_i \sum_{j=1}^{n} \mathbf{P}_{ij} = \mathbf{p}_i$, *since the row sums of* $\mathbf{P}$ *are all* 1 *(as* $\mathbf{P}_{ij}$ *is the probability of walking from* i *to* j*). Meanwhile the right-hand side equals* $(\mathbf{pP})_i$, *i.e., the* i*th entry of the row vector matrix product. So*

$$\mathbf{p}_i = (\mathbf{pP})_i.$$

As this holds for all i, *conclude* $\mathbf{p} = \mathbf{pP}$.

(d) [**6 pts**] Turning things around, suppose we start just with n states and a target distribution (possibly unnormalized) $\mathbf{p} \in \mathbb{R}^{1\times n}$, from which we would like to sample. We choose a connected undirected graph G on the states. Metropolis-Hastings constructs a random walk on the states as follows.

- Let r be the maximum vertex degree of G.
- At state i, select neighbor j with probability $\frac{1}{r}$.
- If state j is selected, we walk to it with probability 1 if $\mathbf{p}_j \geq \mathbf{p}_i$ and with probability $\mathbf{p}_j/\mathbf{p}_i$ if $\mathbf{p}_j < \mathbf{p}_i$.
- Else we stay at state i for the next time step.

This describes a Markov chain on the n states. What is a formula for the (i, j) entry of the transition matrix $\mathbf{P}$? Can you show that $\mathbf{p}$ is the unique stationary distribution?

For simplicity of the description below, we treat G as loopless. We also assume that $\mathbf{p}$ is nonzero on all states.

As a formula, the transition probabilities for Metropolis-Hastings are

$$\mathbf{P}_{ij} = \begin{cases} \frac{1}{r}\min(1, \frac{\mathbf{p}_j}{\mathbf{p}_i}) & \text{if } (i,j) \in E(G) \\ 1 - \sum_{k:(i,k)\in E(G)} \frac{1}{r}\min(1, \frac{\mathbf{p}_k}{\mathbf{p}_i}) & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

To see $\mathbf{p}$ is the unique stationary distribution, it suffices to show that it satisfies the detailed balance equations from part (c) since the Markov chain is strongly connected as G is connected. Thus we may check

$$\mathbf{p}_i\mathbf{P}_{ij} = \mathbf{p}_j\mathbf{P}_{ji}$$

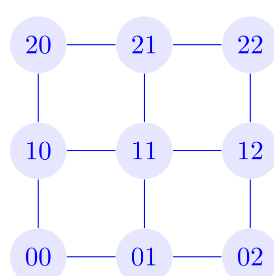
for all i, j. If $i = j$, the equation holds trivially. If $i \neq j$ and $(i, j) \notin E(G)$, the equation also holds because both sides are 0. Finally if i and j are neighbors,

$$\mathbf{p}_i\mathbf{P}_{ij} = \mathbf{p}_i\frac{1}{r}\min(1, \tfrac{\mathbf{p}_j}{\mathbf{p}_i}) = \frac{1}{r}\min(\mathbf{p}_i, \mathbf{p}_j) = \mathbf{p}_j\frac{1}{r}\min(\frac{\mathbf{p}_i}{\mathbf{p}_j}, 1) = \mathbf{p}_j\mathbf{P}_{ji}.$$

(e) [2 pts] As an example, consider using Metropolis-Hastings to create a Markov chain whose stationary probability is that given in the following table. Use the 3×3 lattice for the underlying graph. Please write down two rows of the transition matrix.

x_1x_2	00	01	02	10	11	12	20	21	22
Prob	1/16	1/8	1/16	1/8	1/4	1/8	1/16	1/8	1/16

The graph looks like this:



The transition matrix is 9×9. *Ordering its rows and columns in the order 00, 01, 02, 10, 11, 12, 20, 21, 22, the first two rows are*

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{8} & \frac{1}{2} & \frac{1}{8} & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Here we just plugged the given target distribution $\mathbf{p}$ *into the formula from part (d), e.g., the* $(01, 02)$ *transition probability is* $\mathbf{P}_{01,02} = \frac{1}{4}\min(1, \frac{1/16}{1/8}) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$.

END OF EXAM

CSEM Area B Preliminary Exam

Part 2: Numerical Analysis: Differential Equations (CSE 383L)

May 17, 2024, 100 points total for part 2, a page of notes is allowed

1. [30 pts.] Consider the ODE $u' = f(u)$. Let $\Delta t > 0$, $t^n = n\Delta t$, and $u^n \approx u(t^n)$ in the Runge-Kutta method

$$u^{n+1} = u^n + \frac{\Delta t}{4}\Big[f^n + 3f\big(u^n + \tfrac{2}{3}\Delta t f(u^n + \tfrac{1}{3}\Delta t f^n)\big)\Big],$$

where $f^n = f(u^n)$.

(a) Give the multiplication factor that arises in a linear stability analysis.

(b) Show that the local truncation error is $\mathcal{O}(\Delta t^4)$ (so the method is $\mathcal{O}(\Delta t^3)$ accurate).

2. [45 pts.] Let $\Omega \subset \mathbb{R}^2$ be a nice domain, $f \in L^2(\Omega)$, and consider the variational problem: Find $u \in H_0^1(\Omega)$ such that

$$(\nabla u, \nabla v) + (u, v) = (f, v) \quad \forall v \in H_0^1(\Omega).$$

Let Ω be decomposed into triangles and $V_h \subset H_0^1(\Omega)$ be a piecewise continuous finite element space of linear polynomials. We approximate the solution u by: Find $u_h \in V_h$ such that

$$(\nabla u_h, \nabla v) + (u_h, v) = (f, v) \quad \forall v \in V_h.$$

Recall that $\inf_{v_h \in V_h} \|u - v_h\|_{H^j} \leq Ch^{2-j}\|u\|_{H^2}$, $j = 0, 1$.

(a) Prove that the solution is stable, i.e., that there is some constant $C > 0$ such that $\|u_h\|_{H^1} \leq C\|f\|_{L^2}$.

(b) Prove that for some constant $C > 0$, $\|u - u_h\|_{H^1} \leq Ch\|u\|_{H^2}$.

(c) Prove that for some constant $C > 0$, $\|u - u_h\|_{L^2} \leq Ch^2\|u\|_{H^2}$ [Hint: Recall that in the proof we need to use the solution of the dual problem: $(\nabla\psi, \nabla v) + (\psi, v) = (u - u_h, v) \quad \forall v \in H_0^1(\Omega)$, and the elliptic regularity theorem.]

3. [25 pts.] Let $\Delta t > 0$ and $h > 0$ and consider the scheme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} + \frac{u_{j+1}^n - u_{j-1}^n}{h} = 0,$$

(a) Perform a von Neumann linear stability analysis to determine the multiplication factor $Q(\theta)$. [Hint: show that $|Q(\theta)|^2 = q(\xi)$, there $\xi = \cos\theta \in [-1, 1]$ and q is quadratic.]

(b) Show that the condition $\Delta t \leq h^2/2$ is a *necessary* condition (but it may not be *sufficient*) to obtain stability of the scheme.

1. $u^{n+1} = u^n + \frac{\Delta t}{4}\left[f^n + 3f\left(u^n + \frac{2}{3}\Delta t f\left(u^n + \frac{1}{3}\Delta t f^n\right)\right)\right]$

(a) $f(u) = au, \quad a < 0 \quad (z = a\Delta t < 0)$

$$u^{n+1} = u^n\left\{1 + \frac{1}{4}z\left[1 + 3\left(1 + \frac{2}{3}z\left(1 + \frac{1}{3}z\right)\right)\right]\right\}$$

$$Q(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3$$

(b) $f\left(u^n + \frac{1}{3}\Delta t f^n\right) = f^n + f'(u^n)\left[\frac{1}{3}\Delta t f^n\right] + O(\Delta t^2)$

$f\left(u^n + \frac{2}{3}\Delta t f\left(u^n + \frac{1}{3}\Delta t f^n\right)\right)$

$$= f^n + f'(u^n)\left[\frac{2}{3}\Delta t f\left(u^n + \frac{1}{3}\Delta t f^n\right)\right] + \frac{1}{2}f''(u^n)\left[\frac{2}{3}\Delta t f\left(u^n + \frac{1}{3}\Delta t f^n\right)\right]^2 + O(\Delta t^3)$$

$$= f^n + f'(u^n)\frac{2}{3}\Delta t\left(f^n + f'(u^n)\frac{1}{3}\Delta t f^n\right) + O(\Delta t^3) + \frac{1}{2}f''(u^n)\left(\frac{2}{3}\Delta t\right)^2\left(f^n + O(\Delta t)\right)^2$$

$$= f^n + \frac{2}{3}f'(u^n)f^n\Delta t + \left[(f'(u^n))^2\frac{2}{9}f^n + \frac{2}{9}f''(u^n)f^n\right]\Delta t^2 + O(\Delta t^3)$$

$$\Rightarrow u^{n+1} = u^n + \Delta t\left[f^n + \frac{1}{2}f'(u^n)f^n\Delta t + \frac{1}{6}\left((f')^2 f^n + f'' f\right)\right] + O(\Delta t^4)$$

Now

$$u(t^{n+1}) = u^n + \underbrace{u'(t_n)}_{f^n}\Delta t + \frac{1}{2}\underbrace{u''}_{f'f}\Delta t^2 + \frac{1}{6}\underbrace{u'''}_{f''f^2 + (f')^2 f}\Delta t^3 + \cdots$$

Thus

$$u(t^{n+1}) - u^{n+1} = O(\Delta t^4).$$

2. Find $u \in H_0^1(\Omega)$ s.t. $(\nabla u, \nabla v) + (u, v) = (f, v) \; \forall v \in H_0^1$

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^j} \leq C h^{2-j} \|u\|_{H^2}, \quad j = 0, 1$$

(a) Let $v = u_h \in V_h$ to see

$$\|\nabla u_h\|^2 + \|u_h\|^2 = (f, u_h) \leq \tfrac{1}{2}\|f\|^2 + \tfrac{1}{2}\|u_h\|^2$$

$$\Rightarrow \tfrac{1}{2}\|u_h\|_{H^1}^2 \leq \|\nabla u_h\|^2 + \tfrac{1}{2}\|u_h\| \leq \tfrac{1}{2}\|f\|^2$$

$$\Rightarrow \|u_h\|_{H^1} \leq \|f\|_{L^2}$$

(b) Error:

$$(\nabla(u - u_h), \nabla v_h) + (u - u_h, v_h) = 0$$

$$v_h = \tilde{u}_h - u_h \in V_h$$

$$= (u - u_h) + (\tilde{u}_h - u)$$

$$\Rightarrow \|\nabla(u - u_h)\|^2 + \|u - u_h\|^2$$

$$= (\nabla(u - u_h), \nabla(u - \tilde{u}_h)) + (u - u_h, u - \tilde{u}_h)$$

$$\leq \tfrac{1}{2}(\|\nabla(u - u_h)\|^2 + \|u - u_h\|^2)$$

$$+ \tfrac{1}{2}(\|\nabla(u - \tilde{u}_h)\|^2 + \|u - \tilde{u}_h\|^2)$$

$$\Rightarrow \|u - u_h\|_{H^1} \leq \inf_{\tilde{u}_h \in V_h} \|u - \tilde{u}_h\|_{H^1} \leq C h \|u\|_{H^2}$$

(c) Solve $(\nabla \psi, \nabla v) + (\psi, v) = (u - u_h, v)$

$$\Rightarrow \|u - u_h\|^2 = (\nabla \psi, \nabla(u - u_h)) + (\psi, u - u_h)$$

$$= (\nabla(\psi - \tilde{\psi}_h), \nabla(u - u_h)) + (\psi - \tilde{\psi}_h, u - u_h)$$

$$\leq \|\nabla(\psi - \tilde{\psi}_h)\| \, \|\nabla(u - u_h)\| + \|\psi - \tilde{\psi}_h\| \, \|u - u_h\|$$

$$\leq C h \|\psi\|_{H^2} \, \|u - u_h\|_{H^1}$$

$$\leq C h \|u - u_h\| \, \|u - u_h\|_{H^1}$$

$$\Rightarrow \|u - u_h\| \leq C h^2 \|u\|_{H^2}$$

3. $$\frac{u_j^{n+1}-u_j^n}{\Delta t} - \frac{u_{j+1}-2u_j^n+u_{j-1}^n}{h^2} + \frac{u_{j+1}^n-u_{j-1}^n}{h} = 0$$

(a)

Multiply by $e^{-ij\theta}$ and sum on j

$$\frac{1}{\Delta t}\left(\hat{u}^{n+1}-\hat{u}^n\right)+\frac{1}{h^2}\left(e^{i\theta}-2+e^{-i\theta}\right)\hat{u}+\frac{1}{h}\left(e^{i\theta}-e^{-i\theta}\right)\hat{u}=0$$

$$\Rightarrow \hat{u}^{n+1} = Q(\theta)\,\hat{u}^n$$

$$Q(\theta) = 1+\frac{\Delta t}{h^2}\Big(\underbrace{e^{i\theta}+e^{-i\theta}}_{2\cos\theta}-2\Big)-\frac{\Delta t}{h}\Big(\underbrace{e^{i\theta}-e^{-i\theta}}_{2i\sin\theta}\Big)$$

$$= 1+\frac{2\Delta t}{h^2}(\cos\theta-1)-\frac{2\Delta t}{h}\,i\sin\theta$$

(b)

$$|Q(\theta)|^2 = \left(1+\frac{2\Delta t}{h^2}(\cos\theta-1)\right)^2+\left(\frac{\Delta t}{h}\sin\theta\right)^2$$

$$= \left(1-\frac{2\Delta t}{h^2}\right)^2+\left(\frac{2\Delta t}{h}\right)^2\cos^2\theta+2\left(1-\frac{2\Delta t}{h^2}\right)\frac{2\Delta t}{h^2}\cos\theta$$
$$+\left(\frac{\Delta t}{h}\right)^2\underbrace{\sin^2\theta}_{1-\cos^2\theta}$$

$$= g(\xi),\quad \xi=\cos^2\theta.$$

At the end points:

$$g(\pm 1) = \left(1+\frac{2\Delta t}{h^2}(0 \text{ or } -2)\right)^2$$

$$\Rightarrow \quad \text{need} \quad \left(1-\frac{4\Delta t}{h^2}\right)^2 \le 1$$

$$\Rightarrow \quad -1 \le 1-\frac{4\Delta t}{h^2} \overset{\checkmark}{\le} 1$$

$$\Rightarrow \quad 2 \ge \frac{4\Delta t}{h^2}$$

$$\Rightarrow \quad \Delta t \le \frac{h^2}{2}$$