בס"ד

כלים לאנליזה נומרית 83-214, פרופ' ג'. שיף

מועד ב', סמסטר א', תשס"ד

זמן המבחן: שעתיים וחצי.

מותר להשתמש בכל חומר עזר ובמחשב כיס.

בחלק א' (50% של הציון) יש לענות על כל השאלות (ניקוד כל השאלות בחלק שווה)

בחלק ב' (50% של הציון) יש לענות על 2 מהשאלות (ניקוד כל השאתות בחלק שווה)

חלק א'

1. חשב, בנורמה 1, את הטעות הייחסית במטריצה

$$A = \begin{pmatrix} 34 & 762 & 204 \\ 121 & 24 & 650 \end{pmatrix}$$

כקירוב למטריצה

$$B = \begin{pmatrix} 45 & 788 & 213 \\ 109 & 20 & 640 \end{pmatrix}$$

האם זה מה שמקבלים כאשר כותבים `norm(B-A)/norm(B)` ב- Matlab ?

תשובה:

$$A - B = \begin{pmatrix} -11 & -26 & -9 \\ 12 & 4 & 10 \end{pmatrix}$$

ולכן $||A-B||_1 = 30$. היות ו- $||B||_1 = 854$ הטעות היחסית בנורמה אחד היא $0.35 \approx \frac{30}{854}$. ברירת המחדל לחישוב נורמה ב- Matlab הוא נורמה 2, ולכן לא נקבל בדיוק אותו דבר כאשר כותבים פקודות אלה ב- Matlab.

2. איך מקבלים את הערכים העצמיים של מטריצה ב- Matlab ואיך מקבלים את הערכים הסינגולריים ? איזה מהם בהכרח נותן מספרים ממשיים לא-שליליים ? איזה מהם מוגבל למטריצות ריבועיות ? אם, כאשר אני מחשב את הערכים הסינגולריים של מטריצה גדולה, אני מוצא רק מספר קטן של ערכים שאינם אפס, מה היא המשמעות של זה?

תשובה: לערכים העצמיים של `A` כותבים `eig(A)`. לערכים הסינגולריים כותבים `svd(A)`. הערכים הסינגולריים בהכרח ממשיים לא-שליליים. חישוב ערכים עצמיים מוגבל למטריצות ריבועיות. מספר הערכים הסינגולריים שאינם אפס הוא הדרגה של המטריצה, ולכן אם יש הרבה אפסים אזי הדרגה נמוכה.

3. מה יתנו לי הפקודות הבאות ב-Matlab:

```
x = [0 1 2 3 4 5 6 7 8 9 10];
y = [0 1 4 6 9 9 9 8 3 1 0];
xx = [0:0.1:10];
plot( xx , spline(x,y,xx) )
```

תשובה: הפקודות הראשונה והשנייה מגדירות ווקטורים בעלי אורך 11. ביחד הם מגדירים 11 נקודות במישור x, y. הפקודה השלישית מגדירה ווקטור בעל אורך 101 עם רכיבים בין 0 ל- 10 בקפיצות של 0.1. הפקודה `spline(x,y,xx)` מחשב גם הוא ווקטור בעל אורך 101 כך ש- 101 הנקודות המוגדרות על ידי הווקטורים `xx` ו- `spline(x,y,xx)` נמצאות על עקומה חלקה העוברת דרך 11 הנקודות הנתונות על ידי `x,y`. העקומה הזאת נקראת ה- cubic spline, והיא גזירה פעמיים. ולבסוף: הפקודה `plot` מצייר גרף המחבר בין 101 הנקודות ה"חדשות".

4. מצא פולינום מדרגה 4 או פחות שעובר דרך הנקודות הבאות:

x_i	0	1	2	3	4
y_i	0	3	3	4	4

תשובה: טבלת הפרשים מחולקים:

x_i	y_i				
0	0				
1	3	3			
2	3	0	$-\frac{3}{2}$		
3	4	1	$\frac{1}{2}$	$\frac{2}{3}$	
4	4	0	$-\frac{1}{2}$	$-\frac{1}{3}$	$-\frac{1}{4}$

ולכן הפולינום הוא

$$3x - \frac{3}{2}x(x-1) + \frac{2}{3}x(x-1)(x-2) - \frac{1}{4}x(x-1)(x-2)(x-3)$$

5. כאשר מחשבים את האינטגרל

$$\int_0^1 \sin(\sin x)\, dx$$

על ידי כלל הטרפז עם 2, 4 ו-8 צעדים, מקבלים את התוצאות

$$I_2 \approx 0.4170408129\ , \quad I_4 \approx 0.4272589632\ , \quad I_8 \approx 0.4297719534$$

העזר בשיטת רומברג למצוא קירוב משופר לאינטגרל מהתוצאות האלה.

תשובה: טבלת רומברג:

0.4170408129		
0.4272589632	0.4306650133	
0.4297719534	0.4306096168	0.4306059237

קירוב משופר לאינטגרל: 0.4306059.

6. הפונקציה $y(t)$ פותרת את הבעיה

$$y'' = y^2 + t^2\ , \qquad y(0) = 0\ , y'(0) = -1$$

איזה פקודות היית כותב ב-Matlab כדי למצוא קירוב ל-$y(1)$. (יש רק להסביר את השיטה ode45.)

תשובה: ניתן לכתוב את המשוואה כמערכת מסדר ראשון בצורה

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} y_2 \\ y_1^2 + t^2 \end{pmatrix}$$

ולכן: מכינים קובץ בשם `de.m` הכולל את הפקודות:

```
function z=de(t,y)
z=[ y(2 ) ; y(1)^2 + t^2 ] ;
```

ואז כותבים:

```
[t y] = ode45( @de, [0,1], [0;-1] );
```

המערך y נותן את הערכים של y_1 ו- y_2 בזמנים בין 0 ל-1 הנתונים בווקטור t. כדי לקבל את $y(1)$ ניתן לכתוב `y(size(t,1))`

חלק ב'

1. הפונקציה

$$f(t) = \int_0^1 \sin\left(x^2 + t^2\right) dx$$

היא חיובית כאשר $t = 1.6$ ושלילית כאשר $t = 1.7$. העזר בשלושה סיבובים של חציית קטעים למצוא קירוב לשורש של $f(t) = 0$. ניתן לחשב את האינטגרלים שצריך על ידי כלל הטרפז עם 4 קטעים. איך היית מפעיל שיטת ניוטון לבעייה זו ?

תשובה: עם משתמשים בכלל הטרפז עם 4 צעדים לעשות קירוב לאינטגרל אזי יש לנו

$$f(t) \approx \tilde{f}(t) = \frac{1}{8}\left(\sin\left(t^2\right) + 2\sin\left(\frac{1}{16} + t^2\right) + 2\sin\left(\frac{1}{4} + t^2\right) + 2\sin\left(\frac{9}{16} + t^2\right) + \sin\left(1 + t^2\right)\right)$$

נתון לנו ש- $\tilde{f}(1.6) > 0$ ו- $\tilde{f}(1.7) < 0$ בו'דקים ומוצאים

$$\tilde{f}(1.65) \approx 0.07588 > 0\ , \quad \tilde{f}(1.675) \approx -0.002993 < 0\ , \quad \tilde{f}(1.6625) \approx 0.03662 > 0$$

ולכן השורש בין 1.6625 ובין 1.675. שיטת ניוטון למציאת הפתשון של $f(t) = 0$ הוא

$$t_{n+1} = t_n - \frac{f(t_n)}{f'(t_n)}$$

בשבילנו

$$f'(t) = 2t\int_0^1 \cos\left(x^2 + t^2\right) dx$$

ולכן שיטת ניוטון היא

$$t_{n+1} = t_n - \frac{\int_0^1 \sin\left(x^2 + t_n^2\right) dx}{2t_n \int_0^1 \cos\left(x^2 + t_n^2\right) dx}\ .$$

ניתן לעשות קירובים לשני האינטגרלים, רק שהדיוק של טרפז עם 4 צעדים לא בדיוק מתאים לשיטת ניוטון שמגיעה מהר לדיוק גבוה.

2. (א) הסבר את חיפוש יחס הזהב למציאת מינימום של פונקציה של משתנה אחד. למה היא יעילה, ובכמה ניתן לקצר קטע שבו נמצא מינימום על ידי n סיבובים של השיטה ? יש להדגים על ידי (לא פחות מ-3 סיבובים) לפונקציה

$$\frac{x(2+x)}{2+\sin x}$$

שיש לו מינימום בין $x=-2$ ובין $x=0$.

(ב) יודעים שהפונקציה $f(x)$ מתאפסת כאשר $x=0$ וכאשר $x=1$, ושהיא אונימודלית, בעלת מינימום יחיד, בין 0 ל-1. רוצים לצמצם את הקטע בו נמצא המינימום, ומותר לנו לחשב את הפונקציה $f(x)$ רק עבור שני ערכים של x. למה הכי טוב לחשב את x בשתי נקודות קרובות ל- $x=\frac{1}{2}$? עם מותר ב-3 נקודות, למה הכי טוב ב- $x=\frac{1}{3}$ וב- $x=\frac{2}{3}$ ואח"כ עוד פעם קרוב או ל- $x=\frac{1}{3}$ או ל- $x=\frac{2}{3}$, תלוי באיזה מהם f יותר נמוך ? מה היית עושה אם מותר לחשב את $f(x)$ 4 פעמים ?

תשובה: (א) (לא אכתוב הסבר על השיטה.) אם

$$f(x)=\frac{x(2+x)}{2+\sin x}$$

ברור ש-

$$f(-2)=f(0)=0$$

להתיל חיפוש יחס הזהב נחשב

$$f(-1.236)\approx=-0.8942\ , \qquad f(-0.764)\approx-0.7219$$

ולכן המינימום בין -2 ובין -0.764. יש לנו כבר ערך של f ב- $x=-1.236$ בקטע זה. מוסיפים את הערך

$$f(-1.528)\approx-0.725$$

לקבל שמינימום בין -1.528 ובין -0.764. יש לנו כבר ערך ב- -1.236 בקטע זה, מוסיפים

$$f(-1.056)\approx-0.8872$$

ולכן המינימום בין -1.528 ובין -1.056

(ב) מצב ראשון: אני יודע ש- $f(0)=f(1)=0$. מותר לי לחשב את f רק בשתי נקודות $a<b$. אם יצא $f(a)<f(b)$ אזי המינימום בין 0 ובין b, כלומר אני מצמצם את הקטע על ידי אורך $1-b$. אם יצא $f(a)>f(b)$ אזי המינימום בין a ובין 1, כלומר אני מצמצם את הקטע על ידי אורך $1-a$. כלומר אני בטוח אצמצם את הקטע על ידי אורך

$$\min(1-a,1-b)$$

היות ו- $0<a<b<1$, הצימצום הכי גדול יהיה אם ניקח $a=\frac{1}{2}-\epsilon$, $b=\frac{1}{2}+\epsilon$. אם ניקח ככה שתי נקודות קרובות לאמצע, נוכל לצמצם את הקטע על ידי אורך כמעט חצי.

מצב שני: מותר לחשב f 3 פעמים. קודם כל מתשבים את f ב- $x=\frac{1}{3}$ ו- $x=\frac{2}{3}$. על סמך התוצאות זורקים או את השליש הראשון או את השליש האחרון של הקטע. נשאר לנו קטע בעל אורך $\frac{2}{3}$ שבו אנחנו יודעים את f בקצוות וגם באמצע. על ידי עוד חישוב של f קרוב לאמצע יהיה אפשרי לזרוק (כמעט) תצי ממה שנשאר.

מצב שלישי: מותר לחשב 4 פעמים. מתחילים ב- $x=\frac{2}{5}$ ו- $x=\frac{3}{5}$. זורקים אורך של 0.4. ונשאר קטע שבו אנחנו יודעים את f באחד הנקודות שהוא שליש מאורך הקטע מהקצה (ולכן אפשר להמשיך כמו במצב הקודם).

3. (א) מה הוא ההבדל בין פירוק LU ופירוק Choleski ?

(ב) מצא את פירוק LU למטריצה

$$A = \begin{pmatrix} 25 & -20 & 20 \\ -20 & 25 & -10 \\ 20 & -10 & 29 \end{pmatrix}$$

(אין צורך ב- pivoting).

(ג) למטריצה בסעיף הקודם יש פירוק Choleski $A = LL^T$, כאשר

$$L = \begin{pmatrix} 5 & 0 & 0 \\ -4 & 3 & 0 \\ 4 & 2 & 3 \end{pmatrix}$$

על ידי השוואה עם הפירוק שמצאת בסעיף הקודם, הסבר איך, באופן כללי, ניתן לחשב פירוק LU רגיל עם יודעים פירוק Choleski וגם הפוך.

תשובה: (א) פירוק LU הוא פירוק של מטריצה A לשני גורמים $A = LU$, כאשר L היא משולשת תחתון עם 1'ים על האלכסון, ו-U היא משולשת עליון. הפירוק קיים לכל מטריצה A, רק שלפעמים יש צורך להכפיל את A על ידי מטריצת תמורה P לפני הפירוק. פירוק Choleski הוא גם פירוק לשני גורמים, אחד משולש עליון, אחד משולש תחתון. אלה שהגורם המשולש עליון הוא המשוחלף של המשולש תחתון (ואין בהכרח 1'ים על שום אלכסון). פירוק זה רק קיים למטריצות סימטריות חיוביות בהחלט בלבד.

(ב) אם

$$A = \begin{pmatrix} 25 & -20 & 20 \\ -20 & 25 & -10 \\ 20 & -10 & 29 \end{pmatrix}$$

ניקח

$$l_1 = \begin{pmatrix} 1 \\ -\frac{4}{5} \\ \frac{4}{5} \end{pmatrix}, \qquad u_1 = \begin{pmatrix} 25 & -20 & 20 \end{pmatrix}$$

ואז

$$l_1 u_1 = \begin{pmatrix} 25 & -20 & 20 \\ -20 & 16 & -16 \\ 20 & -16 & 16 \end{pmatrix}, \qquad A - l_1 u_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 9 & 6 \\ 0 & 6 & 13 \end{pmatrix}$$

ניקח

$$l_2 = \begin{pmatrix} 0 \\ 1 \\ \frac{2}{3} \end{pmatrix}, \qquad u_2 = \begin{pmatrix} 0 & 9 & 6 \end{pmatrix}$$

ואז

$$l_2 u_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 9 & 6 \\ 0 & 6 & 4 \end{pmatrix}, \qquad A - l_1 u_1 - l_2 u_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

ולכן פירוק LU של A הוא

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{4}{5} & 1 & 0 \\ \frac{4}{5} & \frac{2}{3} & 1 \end{pmatrix} \begin{pmatrix} 25 & -20 & 20 \\ 0 & 9 & 6 \\ 0 & 0 & 9 \end{pmatrix}$$

(ג) לסכם: פירוק LU:

$$A=\begin{pmatrix}1&0&0\\-\frac{4}{5}&1&0\\\frac{4}{5}&\frac{2}{3}&1\end{pmatrix}\begin{pmatrix}25&-20&20\\0&9&6\\0&0&9\end{pmatrix}$$

פירוק Choleski:

$$A=\begin{pmatrix}5&0&0\\-4&3&0\\4&2&3\end{pmatrix}\begin{pmatrix}5&-4&4\\0&3&2\\0&0&3\end{pmatrix}$$

לעבור מ- Choleski ל- LU: צריך לדאוג שהרכיבים על האלכסון של המטריצה המשולשת תחתון יהיו 1. לעשות את זה נכניס בין שני הגורמים שני מטריצות אלכסוניות (שהכפל שלהם שווה מטריצת היחידה) ככה:

$$\begin{aligned}A&=\begin{pmatrix}5&0&0\\-4&3&0\\4&2&3\end{pmatrix}\begin{pmatrix}5&-4&4\\0&3&2\\0&0&3\end{pmatrix}\\&=\begin{pmatrix}5&0&0\\-4&3&0\\4&2&3\end{pmatrix}\left[\begin{pmatrix}\frac{1}{5}&0&0\\0&\frac{1}{3}&0\\0&0&\frac{1}{3}\end{pmatrix}\begin{pmatrix}5&0&0\\0&3&0\\0&0&3\end{pmatrix}\right]\begin{pmatrix}5&-4&4\\0&3&2\\0&0&3\end{pmatrix}\\&=\left[\begin{pmatrix}5&0&0\\-4&3&0\\4&2&3\end{pmatrix}\begin{pmatrix}\frac{1}{5}&0&0\\0&\frac{1}{3}&0\\0&0&\frac{1}{3}\end{pmatrix}\right]\left[\begin{pmatrix}5&0&0\\0&3&0\\0&0&3\end{pmatrix}\begin{pmatrix}5&-4&4\\0&3&2\\0&0&3\end{pmatrix}\right]\\&=\begin{pmatrix}1&0&0\\-\frac{4}{5}&1&0\\\frac{4}{5}&\frac{2}{3}&1\end{pmatrix}\begin{pmatrix}25&-20&20\\0&9&6\\0&0&9\end{pmatrix}\end{aligned}$$

באופן יותר כללי אם נתון פירוק Choleski

$$A_{ij}=\sum_k L_{ik}L_{jk}\ ,\qquad (L_{jk}=0,\ j<k)$$

אזי ברור ש-

$$A_{ij}=\sum_k\frac{L_{ik}}{L_{kk}}\left(L_{kk}L_{jk}\right)$$

ולכן אם נגדיר מטריצות $\mathcal{L}$, $\mathcal{U}$ ע"י

$$\mathcal{L}_{ik}=\frac{L_{ik}}{L_{kk}}\ ,\qquad \mathcal{U}_{kj}=L_{kk}L_{jk}$$

אזי $\mathcal{L}$ משולש תחתון, $\mathcal{U}$ משולש עליון ו- $A=\mathcal{L}\mathcal{U}$. בכיוון שני - מ- LU ל- Choleski: אין דרך לעשות במקרה הכללי, יש הרבה מקרים שיש פירוק LU ואין Choleski.

4. (א) רוצים למצוא נומרית את $y(1)$ כאשר $y(t)$ פותר את הבעייה

$$y' = y \ , \qquad y(0) = 1$$

(פתרון מדויק $y(t) = e^t$.) איזה קירוב לתשובה מקבלים על ידי שיטת אויילר עם $h = \frac{1}{4}$, ואיזה על ידי שיטת רונגה קוטה עם $h = 1$? באיזה h היה צריך להשתמש בשיטת אויילר כדי לקבל את הדיוק של שיטת רונגה קוטה עם $h = 1$?

(ב) רוצים למצוא נומרית את $y(1)$ כאשר $y(t)$ פותר את הבעייה

$$y' = -2y \ , \qquad y(0) = 1$$

(פתרון מדויק $y(t) = e^{-2t}$.) איזה קירוב לתשובה מקבלים על ידי שיטת אויילר עם $h = \frac{1}{4}$, ואיזה על ידי שיטת רונגה קוטה עם $h = 1$? באיזה h היה צריך להשתמש בשיטת רונגה קואה כדי לקבל את הדיוק של שיטת אויילר עם $h = \frac{1}{4}$?

תשובה: (א) שיטת אויילר למשוואה $y' = y$ היא $y_{n+1} = y_n + hy_n = (1+h)y_n$ ולכן, אם $y(0) = 1$ ולוקחים $h = \frac{1}{4}$

$$y(1) \approx \left(1 + \frac{1}{4}\right)^4 \approx 2.441$$

טעות של $|e - 2.441| \approx 0.28$. עבור Runge-Kutta עם $h = 1$ נקבל

$$\begin{aligned} k_1 &= f(y_0) = 1 \\ k_2 &= f\left(y_0 + \frac{h}{2}k_1\right) = \frac{3}{2} \\ k_3 &= f\left(y_0 + \frac{h}{2}k_2\right) = \frac{7}{4} \\ k_4 &= f\left(y_0 + hk_3\right) = \frac{11}{4} \\ y(1) &\approx y(0) + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = \frac{65}{24} \approx 2.708 \end{aligned}$$

והטעות היא $|e - 2.708| \approx 0.01$.

הטעות באויילר היא בערך פי 28 הטעות ב- RK. כדי להוריד את הטעות בשיטת אויילר על ידי גורם של 28 צריך להקטין את h על ידי גורם של 28 (סדר ראשון). ולכן הייתי משתמש ב- $h = \frac{1}{112}$ בשיטת אויילר כדי לקבל את הדיוק של RK שקבלנו עכשיו. זה עובד! מקבלים קירוב $\left(1 + \frac{1}{112}\right)^{112} \approx 2.706$.

(ב) שיטת אויילר למשוואה $y' = -2y$ היא $y_{n+1} = y_n - 2hy_n = (1-2h)y_n$ ולכן, אם $y(0) = 1$ ולוקחים $h = \frac{1}{4}$

$$y(1) \approx \left(1 - \frac{1}{2}\right)^4 \approx 0.0625$$

טעות של $|e^{-2} - 0.0625| \approx 0.074$. עבור Runge-Kutta עם $h = 1$ נקבל

$$\begin{aligned} k_1 &= f(y_0) = -2 \\ k_2 &= f\left(y_0 + \frac{h}{2}k_1\right) = 0 \end{aligned}$$

$$
\begin{aligned}
k_3 &= f\left(y_0 + \frac{h}{2}k_2\right) = -2 \\
k_4 &= f\left(y_0 + hk_3\right) = 2 \\
y(1) &\approx y(0) + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = \frac{1}{3} \approx 0.333
\end{aligned}
$$

והטעות היא $|e^{-2} - 0.333| \approx 0.20$.

הטעות ב- RK היא בערך פי 3 הטעות באויילר. כדי להוריד את הטעות בשיטת RK על ידי גורם של 3 צריך להקטין את h על ידי גורם של $3^{1/4}$ (סדר רביעי). בפועל ה- h "הבא" שאנחנו יכולים לקחת אחרי $h = 1$ הוא $h = \frac{1}{2}$. זה וודאי יספיק להקטין את הטעות ב- RK לרמה של זה של אויילר. זה גם עובד! הטעות עם $h = \frac{1}{2}$ היא בערך 0.005. (זה דווקא יותר קטן ממה שהייתי מצפה, התנהגות הטעות כ- h^4 היא רק נכון עבור h מספיק קטן.)