שיטות מתמטיות ביישומי מחשב (234299)

תוכן עניינים

פרק 1 - מערכות משוואות ה- LU

פרק 2 – פירוקי ה- LDV וה-Cholesky

פרק 3 – ריבועים פחותים

פרק 4 – אורתוגונאליות

פרק 5 - ערכים עצמיים וסינגולאריים

שיטות מתמטיות ביישומי מחשב (234299)
פרק 1 - מערכות משוואות ה- LU

נושאים שייסקרו:

1. מערכת משוואות כ- Ax=b
2. אלימינציה גאוסית ככלי לפתרון משוואות
3. פעולות בסיסיות על שורות כמכפלה במטריצה בסיסית
4. פירוק A=LU למטריצות רגילות
5. שימוש ב- LU לפתרון מערכות משוואות ע"י החלפה אחורית וקדמית
6. שימוש ב- pivot ופעולת החלפת שורות כהכפלה במטריצה
7. הכללת ה- LU ע"י שימוש במטריצת פרמוטציה – PA=LU
8. אי-סינגולאריות כהבטחה לפתרון והקשר לאברי הציר ב-LU
9. הכללה לתבניות לא ריבועיות
10. שיטות לבחירת Pivot מוצלח
11. כמות חישובים בפירוק LU ובפתרון מערכת משוואות בעזרת פירוק זה

---

1. מערכת משוואות כ- Ax=b

מערכות של משוואות ליניאריות הינן "סוס-העבודה" של המתמטיקה וההנדסה. אם נדע לטפל נכון במערכות משוואות, נוכל לעשות הרבה מאוד, כגון פתרון בעיות אופטימיזציה, פתרון משוואות לא ליניאריות, פתרון משוואות דיפרנציאליות, ועוד.

אנו כיום יודעים הרבה על מערכות משוואות ליניאריות ודרכים מעשיות – ישירות ושאינן ישירות – לפתרונן. הבסיס לידע זה מצוי בתורת המטריצות שהינו פרק באלגברה ליניארית. עם זאת, חשוב לדעת שיש פער בין תיאוריה למעשה, וגישות שנראות נכונות תיאורטית (כגון היפוך מטריצה) הינן למעשה חולניות למדי בבואן אל המעשה. אנו נתמקד בבניית גישות מעשיות לפתרון מערכות של משוואות.

נתונה לנו מערכת משוואות ליניארית. נתן לרושמה בצורה מפורשת

$$\begin{aligned} x+2y+z&=2\\ 2x+6y+z&=7,\\ x+y+4z&=3 \end{aligned}$$

או ע"י רישום מטריצי מהצורה

$$\begin{bmatrix}1&2&1\\2&6&1\\1&1&4\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix}=\begin{bmatrix}2\\7\\3\end{bmatrix}.$$

במקרה הכללי נרשום את המערכת כ- $\mathbf{A}\underline{x}=\underline{b}$, כאשר **A** הינה מטריצה בגודל m שורות על n עמודות (אין כל חובה שהמערכת תהיה ריבועית), וקטור הנעלמים $\underline{x}$ הינו וקטור עמודה באורך n, ווקטור התוצאה $\underline{b}$ הינו וקטור עמודה באורך m. אנו מעוניינים

למצוא את $\underline{x}$. גישה מעניינת המוכרת לנו הינה מניפולציות מפשטות על המשוואות, בגישה הידועה בשם Gaussian Elimination.

2. אלימינציה גאוסית ככלי לפתרון משוואות

למערכת משוואות כללית,

$$\mathbf{A}\underline{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \underline{b}\ .$$

שיטת האלימינציה הגאוסית (שיטה בת כ-150 שנה) מציעה מניפולציות על המשוואות כך שהמטריצה **A** תהפוך למטריצה מדורגת, כך שאיבריה השונים מאפס מרוכזים באלכסון הראשי ומעליו. נמחיש זאת ע"י דוגמה ריבועית שאותה הזכרנו כבר, ואיתה נעבוד לזמן מה. בהינתן המערכת

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 3 \end{bmatrix},$$

נבנה מטריצה מורחבת של נתוני הבעיה

$$\left[\mathbf{A} \mid \underline{b}\right] = \left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 2 & 6 & 1 & 7 \\ 1 & 1 & 4 & 3 \end{array}\right].$$

המשוואה הראשונה תישאר כפי שהיא. באשר למשוואות השנייה והשלישית, ננטרל את האיבר הראשון בהן ע"י פעולת שורות בסיסית בה אנו מצרפים את המשוואה הראשונה כשהיא מוכפלת במקדם מתאים לשורה הנדונה. עבור השורה השנייה ניקח את השורה הראשונה כשהיא מוכפלת ב- (2-) ונוסיפה לשורה השנייה. באופן דומה, לטיפול בשורה השלישית נחסיר את הראשונה ממנה. זה ייתן את המערכת החדשה

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & -1 & 3 & 1 \end{array}\right].$$

הנקודה המרכזית בתהליך הנ"ל היא שפעולות אלו לא שינו את פתרון המערכת ולכן זו שקולה למקורית. יופייה של המערכת החדשה בכך ששתי המשוואות התחתונות מגדירות 2 משוואות ב-2 נעלמים, ולכן פתרונן קל יותר. נשים לב לכך שהאיבר a_{11} במטריצה שימש כאיבר ציר להורדת הערכים מתחתיו, ולו היה 0 היינו תקועים. אנו נכנה איבר זה בשם "איבר ציר" או Pivot.

התהליך יכול להמשיך, כשהמטרה היא הבאה לאפס את האיבר ה-(3,2) שערכו כרגע (-1). אם נצליח בכך, ראה כי פתרון המערכת הופך להיות קל. לצורך זה נכפיל השורה השנייה ב-1/2 ונוסיפה לשלישית. זה ייתן

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & 0 & 2.5 & 2.5 \end{array}\right].$$

הפעם איבר הציר היה האיבר a_{22}. כעת פתרון המערכת קל, כיוון שהמטריצה החדשה **A** הינה משולשת עליונה. פירוש הדבר שאת x_3 נמצא ישירות מהמשוואה השלישית - הוא פשוט $x_3 = 1$. בידיעת משתנה זה, המשוואה השנייה אומרת לנו כי $2x_2 - x_3 = 3$, ולכן ברור כי $x_2 = 2$. המשוואה הראשונה באופן דומה תישען על ידיעת שני המשתנים הקודמים ותיתן כי $x_1 = -3$. תהליך חישובי זה נקרא "החלפה אחורית" (back substitution) והצורך בו נובע ישירות מהמבנה של **A** כמטריצה משולשת עליונה.

3. פעולות בסיסיות על שורות כמכפלה במטריצה בסיסית

אם נחזור על תהליך האלימינציה שתואר קודם, אך ניתן לו פרשנות מעט אחרת, נקבל תבנית מעניינת. הפעולות הבסיסיות על השורות שתוארו קודם ניתנות לתיאור כהכפלה של המטריצה $[\mathbf{A}|\underline{b}]$ במטריצות מיוחדות, אותן נכנה מטריצות בסיסיות. הפעולה הראשונה ניתנת לייצוג כהכפלה מהצורה

$$\mathbf{E}_1 \cdot \left[\mathbf{A} \mid \underline{b}\right] = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 2 & 6 & 1 & 7 \\ 1 & 1 & 4 & 3 \end{array}\right] = \left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 1 & 1 & 4 & 3 \end{array}\right].$$

באופן דומה, הפעולה שנייה והשלישית ניתנות לתיאור כהכפלה במטריצות

$$\mathbf{E}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \text{ ו- } \mathbf{E}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 1 \end{bmatrix}.$$

לכן, התוצאה שהתקבלה ניתנת לרישום כמכפלה

$$\mathbf{E}_3 \cdot \mathbf{E}_2 \cdot \mathbf{E}_1 \cdot \left[\mathbf{A} \mid \underline{b}\right] =$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 2 & 6 & 1 & 7 \\ 1 & 1 & 4 & 3 \end{array}\right] = \left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & 0 & 2.5 & 2.5 \end{array}\right].$$

המטריצות הבסיסיות הנ"ל מאופיינות בכך שאלכסונן הראשי כולל ערכי יחידה. זאת כיוון שבכל פעולת שורות בסיסית אנו מותירים את המשוואה הנוכחית ללא שינוי, ומחברים אליה שורות אחרות המוכפלות בקבוע. אפיון נוסף הוא שמטריצות אלה

משולשות תחתונות. זאת כיוון שבפעולות השורה שאנו עושים אנו ממזגים כל שורה עם אלה שמעליה ולא אלה שמתחתיה.

מעניין לציין שלכל מטריצה בסיסית כזו נוכל להציע את היפוכה בקלות. אם $\mathbf{E}_1$ מבצעת חיבור בין השורה השנייה ובין (2-) השורה הראשונה, הרי שההיפוך יהיה חיבור של השורה השנייה חזרה עם פעמיים הראשונה. כלומר,

$$\mathbf{E}_1=\begin{bmatrix}1&0&0\\-2&1&0\\0&0&1\end{bmatrix}\ \Rightarrow\ \mathbf{E}_1^{-1}=\mathbf{L}_1=\begin{bmatrix}1&0&0\\2&1&0\\0&0&1\end{bmatrix}.$$

באופן דומה נקבל כי

$$\mathbf{E}_2=\begin{bmatrix}1&0&0\\0&1&0\\-1&0&1\end{bmatrix}\ \Rightarrow\ \mathbf{E}_2^{-1}=\mathbf{L}_2=\begin{bmatrix}1&0&0\\0&1&0\\1&0&1\end{bmatrix}$$
$$\mathbf{E}_3=\begin{bmatrix}1&0&0\\0&1&0\\0&0.5&1\end{bmatrix}\ \Rightarrow\ \mathbf{E}_3^{-1}=\mathbf{L}_3=\begin{bmatrix}1&0&0\\0&1&0\\0&-0.5&1\end{bmatrix}.$$

אנו רואים כי ההיפוך של המטריצות הבסיסיות גם הוא מהווה מטריצות בסיסיות, כשכל ההבדל הוא היפוך הסימן של האיברים מתחת לאלכסון (שכל ערכיו נותרים 1). מעניין לציין שכשכופלים את המטריצות הנ"ל מתקבלת התוצאה הבאה

$$\mathbf{L}_1\mathbf{L}_2\mathbf{L}_3=\begin{bmatrix}1&0&0\\2&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&1&0\\1&0&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&1&0\\0&-0.5&1\end{bmatrix}=$$
$$=\begin{bmatrix}1&0&0\\2&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&1&0\\1&-0.5&1\end{bmatrix}=\begin{bmatrix}1&0&0\\2&1&0\\1&-0.5&1\end{bmatrix}.$$

המעניין בתוצאה זו הוא שבשל סדר ההכפלה, האיברים מתחת לאלכסון פשוט חברו יחדיו ללא השפעה הדדית. מעניין למה? גם זו, אגב, מטריצה בסיסית כמקודם.

קודם טענו שהפעלת פעולות בסיסיות על שורות אינה משפיעה על הפתרון של המערכת. כעת אנו יכולים גם להסביר למה. פעולת שורה בסיסית הינה הכפלה במטריצה בסיסית מהסוג שראינו קודם. מטריצה זו מכפילה את שני אגפי המשוואה $\mathbf{A}\underline{x}=\underline{b}$, ומכיוון שמטריצה זו לא סינגולארית (הרי היא הפיכה!), הפתרון נותר על-כנו.

תכונה חשובה בה נוכל להשתמש כאן היא העובדה שמכפלה של מטריצות משולשות תחתונות מניבה מטריצה משולשת תחתונה. באופן דומה, אם כל אלה הינן בסיסיות, אז

מכפלתן נותרת כזו אך היא. לכן המטריצות $\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1$ ו- $\mathbf{L}_1\mathbf{L}_2\mathbf{L}_3$ הינן משולשות תחתונות ובסיסיות.

4. פירוק A=LU למטריצות "רגילות"

מה שקיבלנו עד כה היא המשוואה הבא $\mathbf{L}_1\mathbf{L}_2\mathbf{L}_3 \cdot \mathbf{E}_3\mathbf{E}_2\mathbf{E}_1 = \mathbf{I}$ שפירושה ש- $\mathbf{L} = \mathbf{L}_1\mathbf{L}_2\mathbf{L}_3 = (\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1)^{-1}$, ומאידך קיבלנו גם כי $\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{U}$, כש-**U** הינה המטריצה המשולשת העליונה שיצרנו בסדרת הפעולות הבסיסיות. לכן, קיים הקשר

$$\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{U} \;\Rightarrow\; \mathbf{A} = (\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1)^{-1}\mathbf{U} = \mathbf{LU}.$$

כלומר, המטריצה **A** פורקה כתוצאה מתהליך ה-Gauss Elimination למכפלה של שתי מטריצות – **L** שהינה מטריצה משולשת תחתונה בסיסית (עם 1-ים על האלכסון), ו-**U** שהינה משולשת עליונה. ננסח זאת כמשפט:

משפט 1: מטריצה "רגילה" **A** ניתנת לפירוק יחיד מהצורה **A=LU**, כש-**L** מטריצה משולשת תחתונה בסיסית, ו-**U** מטריצה משולשת עליונה כלשהי.

המושג "מטריצה רגילה" בא לומר כי זו מטריצה שהתהליך שתואר פועל עליה בהצלחה מתחילה ועד הסוף. מהתיאור עד כה עלול להיווצר הרושם שתהליך זה צפוי להצליח תמיד ואין זה כך כללי! למשל, אם במהלך האלימינציה הגאוסית מתקבל איבר ציר שערכו 0 לא נוכל להשתמש בו לביטול איברים מתחתיו. כמקרה פרטי באנאלי של בעיה זו, מטריצה שבה $a_{11} = 0$ נתקעת כבר בהתחלה. אנו נראה בהמשך מה קורה במקרים אלו ואחרים בהם המטריצה אינה "רגילה". לעת עתה נניח כי היא כזו.

5. שימוש ב- LU לפתרון מערכות משוואות ע"י החלפה אחורית וקדמית

למה חשוב לנו הפירוק **A=LU**? כבר ראינו כי פתרון מערכת משוואות ליניאריות מתקבל בקלות בעזרת Gauss Elimination. מה הקשר ל-LU? נניח כי לפנינו מערכת משוואות ריבועית עם n משוואות ב-n נעלמים, המשוואה $\mathbf{A}\underline{x}=\underline{b}$, ורצוננו בפתרון. נניח כי **A** רגילה, וכיוון שכך ניתן לפרקה למכפלה $\mathbf{A} = \mathbf{LU}$. לכן עלינו לפתור את המשוואה

$$\underline{b} = \mathbf{A}\underline{x} = \mathbf{LU}\underline{x} = \mathbf{L}(\mathbf{U}\underline{x}) = \mathbf{L}\underline{y}.$$

לפתור את המערכת $\underline{b} = \mathbf{L}\underline{y}$ די קל – זוהי מערכת מהצורה

$$
\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & 0 & \cdots & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \ell_{n4} & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ \vdots \\ b_n \end{bmatrix}.
$$

פתרונה דורש החלפה קדמית, בה פותרים עבור y_1, שימוש בערך שנמצא לחילוץ y_2, וכך הלאה, עד להגעה לפתרון המלא. עם סיום שלב זה יש לנו מערכת משוואות חדשה $\mathbf{U}\underline{x} = \underline{y}$, מהצורה

$$
\begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & u_{24} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & u_{34} & \cdots & u_{3n} \\ 0 & 0 & 0 & u_{44} & \cdots & u_{4n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{bmatrix}.
$$

פתרונה קל במידה זהה, כשהפעם יש לבצע החלפה אחורית לקבלת הפתרון מ- x_n ועד x_1. כלומר, קיבלנו את התכונה הבאה: אם המטריצה **A** "רגילה", לא רק נוכל לפרקה ל-**LU** כמתואר, אלא שמובטח שלכל $\underline{b}$ שהוא יהיה פתרון אחד ויחיד למערכת $\mathbf{A}\underline{x}=\underline{b}$.

6. שימוש ב- (pivot) ופעולת החלפת שורות כהכפלה במטריצה

מה קורה כשאיבר ציר (pivot) מתאפס? האם השיטה קורסת? לא בהכרח! כל שעלינו לעשות הוא להחליף סדר השורות. האם פתרון זה יבטיח שהפירוק LU יצליח לכל מטריצה? לצערנו התשובה היא לא! תהיינה מטריצות שבשלב מסוים יתנו כי כל המועמדים לשמש כאברי ציר מאופסים, ואז התהליך נתקע. מסתבר כי אנו מכירים היטב מטריצות אלה – הם המטריצות הסינגולאריות. לעת עתה נעזוב את אלה ונתרכז במקרים בהם ניתן לסיים בהצלחה את האלימינציה הגאוסית בכפוף להחלפת סדר שורות.

ניתן דוגמה להמחשת העניין. נטפל במערכת המשוואות הבאה:

$$
\begin{bmatrix} 0 & 3 & 1 \\ 2 & 6 & 1 \\ 1 & 0 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 3 \end{bmatrix}.
$$

ארגון המטריצה המורחבת תיתן

$$[\mathbf{A} \mid \underline{b}] = \left[\begin{array}{ccc|c} 0 & 3 & 1 & 2 \\ 2 & 6 & 1 & 7 \\ 1 & 0 & 4 & 3 \end{array}\right].$$

ברור כי לא ניתן להתחיל בתהליך תוך שימוש ב- a_{11}. מאידך, החלפת סדר השורות בין הראשונה לשנייה תפתור את הבעיה לחלוטין. כלומר, נתייחס למערכת מעט אחרת אך שקולה שנראית כך:

$$[\mathbf{A}_1 \mid \underline{b}_1] = \left[\begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 3 & 1 & 2 \\ 1 & 0 & 4 & 3 \end{array}\right].$$

למעשה, מה שהיווה בעיה קודם הפך ליתרון – כעת אנו פתורים מלטפל במשוואה השנייה כיוון שהיא כבר כוללת 0 בתחילתה. קל לראות כי הכפלת מטריצה זו במטריצה הבסיסית הבאה תנטרל את העמודה הראשונה:

$$\mathbf{E}_1 \cdot [\mathbf{A}_1 \mid \underline{b}_1] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -0.5 & 0 & 1 \end{bmatrix} \left[\begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 3 & 1 & 2 \\ 1 & 0 & 4 & 3 \end{array}\right] = \left[\begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 3 & 1 & 2 \\ 0 & -3 & 3.5 & -0.5 \end{array}\right].$$

כעת איבר הציר למשען הוא האיבר $a_{22} = 3$ ושימוש בו נותן כי יש לכפול במטריצה הבסיסית

$$\mathbf{E}_2\mathbf{E}_1 \cdot [\mathbf{A}_1 \mid \underline{b}_1] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \left[\begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 3 & 1 & 2 \\ 0 & -3 & 3.5 & -0.5 \end{array}\right] = \left[\begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 3 & 1 & 2 \\ 0 & 0 & 4.5 & 1.5 \end{array}\right],$$

ובזאת סיימנו, ובהצלחה. מכאן בהחלפה לאחור אנו יכולים לקבל את הפתרון של מערכת משוואות זו.

7. הכללת ה- LU ע"י שימוש במטריצת פרמוטציה – PA=LU

למעשה, המעבר שעשינו עם החלפת השורות אינו אלא הכפלה של מערכת המשוואות במטריצה ריבועית מהצורה

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

לכן, מה שקיבלנו בפועל הוא הקשר

$$\mathbf{PA} = \mathbf{LU},$$

שפירושו – את שורות **A** יש לסדר אחרת, ואז המטריצה הופכת לרגילה, אז ידוע כי יש לה פירוק **LU**. מסתבר כי תכונה זו כללית כש-**P** אמורה להיות מטריצת פרמוטציה המסדרת מחדש את משוואות המערכת. מטריצה תהיה מטריצת פרמוטציה אם בכל

שורה ובכל עמודה יש רק איבר יחד השונה מ-0 וערכו 1. למטריצות בגודל n-על-n יש !n מטריצות פרמוטציה אפשריות. מעניין שמכפלת מטריצות פרמוטציה אף היא מטריצת פרמוטציה. למרות זאת, מכפלת מטריצות אלה אינה קומוטטיבית ויש להיזהר בזה.

ניתן כעת דוגמה אחרת ממנה נבין כי לא כל מטריצה צפויה להוביל להצלחת הפירוק הנ"ל. נניח כי עלינו לפתור את המערכת

$$\left[\mathbf{A} \mid \underline{b}\right]=\left[\begin{array}{cccc|c} 0 & 2 & 3 & 4 & 9 \\ 2 & 1 & 1 & 1 & 5 \\ 4 & 4 & 5 & 6 & 19 \\ 2 & 2 & 2 & 2 & 8 \end{array}\right].$$

צעד ראשון מחייב החלפת השורה הראשונה באחרת (בשנייה בדוגמה זו, אך כבר כאן ברור כי אין יחידות כי יכולנו לבחור החלפה אחרת). זה יתקבל ע"י ההכפלה ב- $\mathbf{P}_1$. לאחר מכן נבצע שתי פעולות בסיסיות $\mathbf{E}_1$ ו-$\mathbf{E}_2$ לנטרול האיברים הראשונים בשורות השלישית והרביעית ונקבל:

$$\mathbf{E}_2\mathbf{E}_1\mathbf{P}_1\left[\mathbf{A} \mid \underline{b}\right]=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\left[\begin{array}{cccc|c} 2 & 1 & 1 & 1 & 5 \\ 0 & 2 & 3 & 4 & 9 \\ 4 & 4 & 5 & 6 & 19 \\ 2 & 2 & 2 & 2 & 8 \end{array}\right]=\left[\begin{array}{cccc|c} 2 & 1 & 1 & 1 & 5 \\ 0 & 2 & 3 & 4 & 9 \\ 0 & 2 & 3 & 4 & 9 \\ 0 & 1 & 1 & 1 & 3 \end{array}\right].$$

ניכר כי המשוואות השנייה והשלישית זהות, עובדה שפירושה שיש לנו בפועל 3 משוואות ב-4 נעלמים. נעביר את השורה השלישית להיות הרביעית, אך נעשה זאת כשלב מקדים לכל הפעולות הקודמות. לכן נתחיל במערכת

$$\mathbf{P}_2\mathbf{P}_1\left[\mathbf{A} \mid \underline{b}\right]=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\left[\begin{array}{cccc|c} 0 & 2 & 3 & 4 & 9 \\ 2 & 1 & 1 & 1 & 5 \\ 4 & 4 & 5 & 6 & 19 \\ 2 & 2 & 2 & 2 & 8 \end{array}\right]=$$

$$=\left[\begin{array}{cccc|c} 2 & 1 & 1 & 1 & 5 \\ 0 & 2 & 3 & 4 & 9 \\ 2 & 2 & 2 & 2 & 8 \\ 4 & 4 & 5 & 6 & 19 \end{array}\right]=\left[\mathbf{A}_1 \mid \underline{b}_1\right],$$

וכעת נתחיל מהתחלה עם פעולות בסיסיות – אלה תהיינה

$$\mathbf{E}_2\mathbf{E}_1\left[\mathbf{A}_1 \mid \underline{b}_1\right]=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -2 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\left[\begin{array}{cccc|c} 2 & 1 & 1 & 1 & 5 \\ 0 & 2 & 3 & 4 & 9 \\ 2 & 2 & 2 & 2 & 8 \\ 4 & 4 & 5 & 6 & 19 \end{array}\right]=\left[\begin{array}{cccc|c} 2 & 1 & 1 & 1 & 5 \\ 0 & 2 & 3 & 4 & 9 \\ 0 & 1 & 1 & 1 & 3 \\ 0 & 2 & 3 & 4 & 9 \end{array}\right].$$

שתי פעולות בסיסיות נוספות תבואנה עם שימוש באיבר הציר a_{22}, ותיתנה

$$\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\left[\mathbf{A}_1 \mid \underline{b}_1\right] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -0.5 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \left[\begin{array}{cccc|c} 2 & 1 & 1 & 1 & 5 \\ 0 & 2 & 3 & 4 & 9 \\ 0 & 1 & 1 & 1 & 3 \\ 0 & 2 & 3 & 4 & 9 \end{array}\right] =$$

$$= \left[\begin{array}{cccc|c} 2 & 1 & 1 & 1 & 5 \\ 0 & 2 & 3 & 4 & 9 \\ 0 & 0 & -0.5 & -1 & -1.5 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

כאן למעשה נעצרים, למרות שבמקרה הכללי היה נדרש שלב נוסף שמשתמש ב-a_{33} כאיבר ציר. במקרה זה אין צורך כי מתחתיו האיבר מאופס כבר. יתרה מזו – גם האיבר a_{44} מאופס, עובדה שפירושה "סינגולאריות". לסיכום מלוא תהליך זה, קיבלנו כי מתקיימת המשוואה

$$\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{P}_2\mathbf{P}_1\mathbf{A} = \mathbf{U},$$

ולכן

$$\mathbf{P}_2\mathbf{P}_1\mathbf{A} = \mathbf{PA} = \left(\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\right)^{-1}\mathbf{U} = \mathbf{LU},$$

כשהמטריצה **L** כמקודם היא היפוך של מכפלת מטריצות בסיסיות ואף היא כזו. אלא שהמטריצה **U** הפעם הינה משולשת עליונה אך עם הערכים לאורך אלכסונה כוללים גם אפס. תכונה זו מעידה על היות המטריצה המקורית סינגולארית.

נסכם את הנאמר עד כה ונביא את המשפט הבא:

משפט 2: כל מטריצה **A** ניתנת לפירוק מהצורה **PA=LU**, כש-**L** מטריצה משולשת תחתונה בסיסית, ו-**U** מטריצה משולשת עליונה כלשהי. אם איברי האלכסון הראשי של **U** כולם שונים מאפס, **A** אינה סינגולארית.

נשים לב לשתי נקודות מעניינות שעלו במהלך הדוגמה האחרונה: (1) אין יחידות בפירוק – יכולנו להוביל לפירוק אחר אם היינו בוחרים החלפות שונות, (2) את זהות המטריצה **P** אנו קובעים מראש ומשנים את הסדר בתחילת התהליך ולא בעיצומו. פירוש הדבר שעלינו לבצע הפירוק, להיתקל בבעיות, לקבוע את סדר המשוואות הראוי, ואז לחזור ולבצע התהליך מחדש.

8. אי-סינגולאריות כהבטחה לפתרון והקשר לאברי הציר ב-LU

במקרה הכללי, אם במהלך תהליך האלימינציה אנו מגיעים למבנה בו איבר ציר מתאפס על כל מה שמתחתיו, כמתואר כאן באיבר (5,5)

$$
\begin{bmatrix}
u_{11} & u_{12} & u_{13} & u_{14} & u_{15} & u_{16} & \cdots & u_{1n} \\
0 & u_{22} & u_{23} & u_{24} & u_{25} & u_{26} & \cdots & u_{2n} \\
0 & 0 & u_{33} & u_{34} & u_{35} & u_{36} & \cdots & u_{3n} \\
0 & 0 & 0 & u_{44} & u_{45} & u_{46} & \cdots & u_{4n} \\
0 & 0 & 0 & 0 & 0 & u_{56} & \cdots & u_{5n} \\
0 & 0 & 0 & 0 & 0 & u_{66} & \cdots & u_{6n} \\
0 & 0 & 0 & 0 & 0 & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & 0 & u_{n6} & \cdots & u_{nn}
\end{bmatrix}.
$$

הדבר גוזר בבירור סינגולאריות. דרך פשוטה להשתכנע שזה אמנם כך היא הדרך הבאה – במטריצה הנ"ל השורות 5 עד n (כלומר n-4 שורות) כוללות כולן איברים שונים מאפס במקומות 6 עד n (כלומר n-5 איברים שונים מאפס). קבוצה של שורות אלה יכולה לפרוס מימד מרבי n-5, ולכן מימד אחד לפחות אבד לנו.

ככלל, קיימת התכונה הבאה: בהינתן מטריצה **A** כלשהי, נוכל לקבוע את דרגתה (rank) ע"י תהליך אלימינציה גאוסית. נבצע את התהליך שהוביל למטריצה U, ונספור את כמות איברי הציר (pivots). כמותם היא דרגת המטריצה. לכן, אם במטריצה ריבועית אחד מאיברי האלכסון התאפס לנו, איבדנו איבר ציר ואיתו נפלה הדרגה.

דרך אחרת לראות את בעיית הסינגולאריות היא הבאה: נזכור כי דרך אחת לאפיין מטריצות סינגולאריות היא ערך הדטרמיננטה. כזכור, לדטרמיננטה יש משמעות גיאומטרית – הדטרמיננטה של מטריצה קובע את נפח המקבילית המתקבלת מעמודות (או שורות) המטריצה. אם השורות תלויות ליניארית, המקבילית פחוסה ונפחה הוא 0, עובדה שמעידה על סינגולאריות. איך זה קשור אלינו?

נשתמש בתכונה הבאה של דטרמיננטות של מטריצות - $\det\{\mathbf{AB}\} = \det\{\mathbf{A}\}\det\{\mathbf{B}\}$. לכן, לפירוק ה- LU שקיבלנו נוכל לומר כי קיים

$$\det\{\mathbf{PA}\} = \det\{\mathbf{P}\}\det\{\mathbf{A}\} = \det\{\mathbf{L}\}\det\{\mathbf{U}\}.$$

תכונה אחרת של דטרמיננטות היא שניתן לחשבה ע"י סכום (עם סימן שנקבע לפי היות הסידור זוגי או אי-זוגי, כלומר מספר ההחלפות הנדרש להבאת הסדר למקור) של מכפלות על-פני כל הסידורים האפשריים של איברי המטריצה כך שמכל שורה ומכל עמודה נבחר איבר יחיד. לכן,

$$\det\{\mathbf{P}\} = \pm 1.$$

בהכרח, ובאופן דומה,

$$\det\{\mathbf{L}\} = 1 \quad \text{ו-} \quad \det\{\mathbf{U}\} = \prod_{k=1}^{n} u_{kk}.$$

לכן, די באיבר יחיד באלכסון **U** מאופס על מנת לגרום לדטרמיננטה של **A** להיות אפס גם כן.

9. הכללה לתבניות לא ריבועיות

שום דבר ממה שתואר כאן אינו מגביל את התהליך למערכות ריבועיות. התוצאה בד"כ במערכות מלבניות תהיה בעלת מבנה מדורג שבו אנו רואים ברכה כיוון שהוא חושף את אופייה של המערכת ככזו עם פתרון יחיד, אינסוף פתרונות, או אף פתרון. נמחיש זאת דרך הדוגמה הבאה – נתונה המערכת הליניארית:

$$\begin{bmatrix} 1 & 3 & 2 & -1 & 0 \\ 2 & 6 & 1 & 4 & 3 \\ -1 & -3 & -3 & 3 & 1 \\ 3 & 9 & 8 & -7 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}.$$

נמחיש את פירוקה ל-LU ע"י הכפלה במטריצות בסיסיות. טיפול בעמודה הראשונה תוך שימוש ב- a_{11} כאיבר ציר ייתן

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \left[\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 2 & 6 & 1 & 4 & 3 & b \\ -1 & -3 & -3 & 3 & 1 & c \\ 3 & 9 & 8 & -7 & 2 & d \end{array}\right] = $$
$$= \left[\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b-2a \\ 0 & 0 & -1 & 2 & 1 & c+a \\ 0 & 0 & 2 & -4 & 2 & d-3a \end{array}\right].$$

אנו רואים כי איבר הציר השני a_{22} מנוטרל ואין צורך לטפל בו, ולכן נעבור לאיבר a_{23}. שוב סדרת פעולות בסיסיות תיתן

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 2/3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1/3 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \left[\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b-2a \\ 0 & 0 & -1 & 2 & 1 & c+a \\ 0 & 0 & 2 & -4 & 2 & d-3a \end{array}\right] = $$
$$= \left[\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b-2a \\ 0 & 0 & 0 & 0 & 0 & c-b/3+5a/3 \\ 0 & 0 & 0 & 0 & 4 & d+2b/3-13a/3 \end{array}\right].$$

לכן, הפירוק שבפועל היה כאן ניתן לתיאור ע"י

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&-1/3&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&2/3&1&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\1&0&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\-3&0&1&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\-2&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}.$$

$$.\qquad .\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&1\\0&0&1&0\end{bmatrix}\left[\mathbf{A}\,\middle|\,\underline{b}\right]=\mathbf{E}_5\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{P}\left[\mathbf{A}\,\middle|\,\underline{b}\right]$$

כשבתיאור זה אנו קודם מבצעים החלפת שורות שלישית ורביעית ואז חוזרים על תהליך האלימינציה עם התייחסות נאותה לשינוי בסדר המשוואות. תוצאת כל פעולות אלה היא

$$.\left[\begin{array}{ccccc|c}1&3&2&-1&0&a\\0&0&-3&6&3&b-2a\\0&0&0&0&4&d+2b/3-13a/3\\0&0&0&0&0&c-b/3+5a/3\end{array}\right]=\left[\mathbf{U}\,\middle|\,\hat{\underline{b}}\right]$$

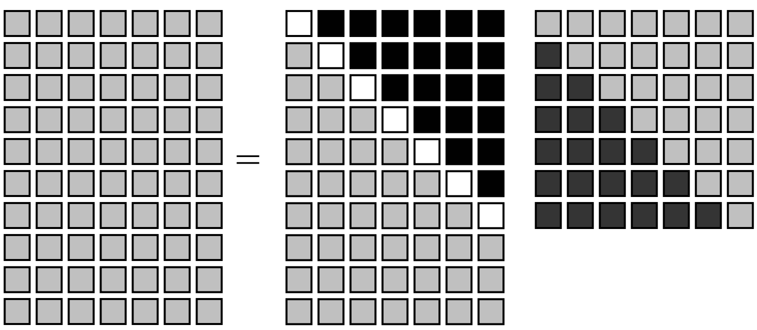
לכן, קיבלנו כאן את הפירוק הצפוי הבא

$$\mathbf{E}_5\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{PA}=\mathbf{U}$$

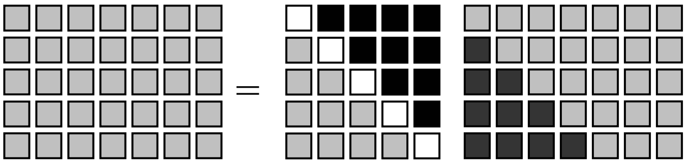
$$.\Rightarrow\ \mathbf{PA}=\left(\mathbf{E}_5\mathbf{E}_4\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\right)^{-1}\mathbf{U}=\mathbf{LU}=\begin{bmatrix}1&0&0&0\\2&1&0&0\\3&-2/3&1&0\\-1&1/3&0&1\end{bmatrix}\begin{bmatrix}1&3&2&-1&0\\0&0&-3&6&3\\0&0&0&0&4\\0&0&0&0&0\end{bmatrix}$$

וכמו כן קיבלנו מערכת משוואות שממנה נובע כי אם $c-b/3+5a/3=0$ יש למערכת זו אינסוף פתרונות, כיוון שניוותר עם 3 משוואות ו-5 נעלמים – כלומר שתי דרגות חופש. אם לעומת זאת $c-b/3+5a/3\neq 0$, למערכת זו אין פתרון.

במקרה הכללי, בהינתן מטריצה בגודל n-על-m כאשר n (מספר השורות) גדול ממספר העמודות, לאחר פרמוטציה ראויה, נצפה לפירוק מהצורה (קובייה לבנה מייצגת '1', שחורה '0', ואפורה ערך כלשהו)



באופן דומה, כאשר מספר השורות קטן ממספר העמודות נקבל מבנה מהצורה



10. שיטות לבחירת Pivot מוצלח

ראינו כי בתהליך הפירוק יש לנו החופש לבצע החלפת סדר שורות. עד כה ניסינו להימנע מהחלפת הסדר וניצלנו אפשרות זאת רק במקרים של התאפסות איבר הציר – כלומר במקרה של מטריצה שאינה "רגילה". האם יש רווח בשינוי סדר שורות גם במקרים אחרים? התשובה היא כן – להחלפת סדר שורות כוח בהתמודדות עם שגיאות נומריות. נמחיש זאת דרך הדוגמה הפשוטה הבאה – נניח כי אנו מטפלים במערכת המשוואות

$$\begin{bmatrix} 0.01 & 1.6 \\ 1 & 0.6 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 32.1 \\ 22 \end{bmatrix}.$$

פתרונה האמיתי של מערכת זו הוא $x = 10; y = 20$. טיפול במערכת זו ללא החלפת סדר פירושו הכפלה במטריצה הבסיסית

$$\begin{bmatrix} 1 & 0 \\ -100 & 1 \end{bmatrix}.$$

והתוצאה תהיה

$$\begin{bmatrix} 1 & 0 \\ -100 & 1 \end{bmatrix}\begin{bmatrix} 0.01 & 1.6 \\ 1 & 0.6 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -100 & 1 \end{bmatrix}\begin{bmatrix} 32.1 \\ 22 \end{bmatrix}$$
$$\Rightarrow \quad \begin{bmatrix} 0.01 & 1.6 \\ 0 & -159.4 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 32.1 \\ -3188 \end{bmatrix}.$$

נניח (על מנת להקצין את המצב) כי בידינו מחשב בעל דיוק של שלוש ספרות עשרוניות. לכן מספר כמו 3188 יהיה בפועל 3190, ובאופן דומה, 159.4 יהפוך ל-159. המערכת לפתרון הינה, אם כך,

$$\begin{bmatrix} 0.01 & 1.6 \\ 0 & -159 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 32.1 \\ -3190 \end{bmatrix}.$$

בשיטת ההחלפה לאחור נקבל כי $y = 3190/159 = 20.1$ (אחרי העגלה נחוצה בשל מגבלת הדיוק שלנו), ומכאן נגיע למסקנה ש-$0.01x + 1.6 \cdot 20.1 = 32.1 \Leftarrow x = -6$. אנו רואים כי בפתרון x ישנה שגיאה קשה.

לו החלפנו את סדר השורות היינו מקבלים כי הפעולה הבסיסית הנחוצה היא

$$\begin{bmatrix} 1 & 0 \\ -0.01 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0.6 \\ 0.01 & 1.6 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -0.01 & 1 \end{bmatrix}\begin{bmatrix} 22 \\ 32.1 \end{bmatrix},$$

$$\Rightarrow \quad \begin{bmatrix} 1 & 0.6 \\ 0 & 1.59 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 22 \\ 31.9 \end{bmatrix}$$

כשכאן כבר הכנסנו את שגיאות העגלה שנובעות מדיוק מוגבל. בהחלפה אחורית היינו מקבלים y כמקודם, $y = 20.1$. שימוש במשוואה הראשונה היה נותן

$$x = (22 - 0.6 \cdot 20.1)/1 = 9.94,$$

וזו תוצאה הרבה יותר מדויקת.

מדוגמה זו נובע כי רצוי לבחור כאיבר ציר את האיבר הגדול ביותר בעמודה הנדונה. בדוגמה שלנו התלבטנו בין 0.01 ו-1 והיה עלינו לבחור את 1. הרציונל ברור – בחירת איבר ציר גדול פירושה שכל יתר השורות מוכפלות במקדם קטן מ-1 (המנה בה מכפילים את השורה המתחברת היא $-a_{jk}/a_{kk}$ לשורה ה-j). גישה זו מוכרת בשם Partial Pivoting.

עם זאת, עדיין יכולות להתקבל בעיות. נניח שמערכת המשוואות שלנו הינה

$$\begin{bmatrix} 10 & 1600 \\ 1 & 0.6 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 32100 \\ 22 \end{bmatrix}.$$

(המשוואה הראשונה נותרה כשהייתה אך הוחלפה ב- 1000). האם כעת נכון לבחור ב-10 כאיבר ציר? לא! הבעיה שנחשפה קודם הייתה מתרחשת שוב באותה צורה בדיוק. מהות הבעיה כאן היא שהיחס בין איברי השורה יגרום לסטייה הגדולה בערכו של x.

הגישה לפתרון צריכה להיות החלפת סדר עמודות!! בהחלפת סדר עמודות אנו בפועל מארגנים מחדש את סדר המשתנים, ולכך אין כל השפעה על המערכת. עד כה לא עשינו שימוש ברעיון זה, אך כאן הוא חיוני. לכן, בבואנו לבחור את איבר הציר בשלב ה-k, לא נסתכל רק על האיברים $a_{jk},\ j = k, k+1, \dots, n$, אלא על כל תת המטריצה הריבועית $a_{j_1 j_2},\ j_1, j_2 = k, k+1, \dots, n$ ונחפש את האיבר הגדול ביותר ובו נבחר כאיבר ציר. עם מציאתו, נחליף את סדר המשתנים כדי לשקף את העובדה ש-$j_2 \neq k$, ונחליף את סדר השורות כדי לשקף את העובדה ש- $j_1 \neq k$. גישה זו קרויה Full Pivoting.

נקודה חשובה שמחייבת הסבר היא שהחלפות סדר אלה לא חייבות להתבצע בפועל, וכל משמעותן אינה אלא החזקה של שני וקטורי מצביעים (Pointers) לרשימת

המשוואות ורשימת המשתנים. תרגיל זה אינו מחייב הסבר עמוק לתלמידי מדעי המחשב הבקיאים בכאלה תרגילים.

11. כמות חישובים בפירוק LU ובפתרון מערכת משוואות בעזרת פירוק זה

זה אולי נראה קטנוני, אבל ספירת כמות החישובים הנחוצה לפירוק LU ולשימוש בו חשובה מאוד. כשעוסקים במערכות של אלפי משוואות באלפי נעלמים, ניואנסים עדינים וחסכונות הופכים מהותיים. לכן, ננסה לכמת את כמות החישובים בפעולות שתיארנו, וזה ייתן לנו מושג על יעילותם. בניתוח שנעשה נתייחס לחיבורים (חיסורים) ומכפלות (חלוקות) כשתי קטגוריות נפרדות, ונכמת את שתיהן.

לשם ייחוס, מכפלה של מטריצה בגודל n-על-n בווקטור באורך n דורשת n מכפלות ו-(n-1) סיכומים לכל איבר, ולכן סך של $n^2{}_{[mul]} + n(n-1)_{[add]}$. עבור $n=10^6$ (מליון משוואות ומליון נעלמים – זה לא גודל מופרך – בעיות רבות בעיבוד תמונות ובגרפיקה הן בגודל זה) נידרש לסדר גודל של 10^{12} מכפלות וסיכומים, וגם מחשב חזק יימצא כמות זו כאתגר.

כמה חישובים נחוצים בפירוק LU או באלימינציה גאוסית? לשם טיפול בשאלה זו נניח כי המטריצה "רגילה" ואין כל התחכמויות בארגון סדר השורות והעמודות ל-pivoting. מסתבר כי בכל מקרה, טיפול בפרמוטציות ע"י פוינטרים לא מעלה את כמות המכפלות והסיכומים ולכן השפעתו זניחה. עבור מטריצה בגודל n-על-n, ה-pivot הראשון יהיה a_{11}. איפוס האיבר הראשון בשורה ה-j ייעשה ע"י חישוב המנה a_{j1}/a_{11}, ו- (n-1) הכפלות וסיכומים לשינוי שורה זו. כיוון שיש (n-1) שורות, סך הפעולות יהיה n(n-1) מכפלות ו- $(n-1)^2$ סיכומים. כל זאת בהתייחס למטריצה **A**. אם אנו עוסקים באלימינציה גאוסית, יש גם לטפל בווקטור b - יידרשו (n-1) מכפלות וככמות הזאת חיבורים לעדכון וקטור זה. כל זאת מול ה-pivot הראשון

באופן דומה, בשלב ה-k יידרשו k(k-1) מכפלות ו-$(k-1)^2$ חיבורים לעדכון **A**, ו-(k-1) חיבורים ומכפלות לעדכון b. לכן, סך הכול, יידרשו

$$\sum_{k=1}^{n} k(k-1)_{[mul]} + \sum_{k=1}^{n}(k-1)^2{}_{[add]} = \frac{n^3-n}{3}{}_{[mul]} + \frac{2n^3-3n^2+n}{6}{}_{[add]}$$

לבניית **L** ו-**U**. לעדכון b יידרשו באופן דומה

$$.\sum_{k=1}^{n}(k-1)_{[mul]} + \sum_{k=1}^{n}(k-1)_{[add]} = \frac{n^2-n}{2}{}_{[mul]} + \frac{n^2-n}{2}{}_{[add]}$$

(ניתן להוכיח נוסחאות סכומים אלה באינדוקציה).

לאחר סיום תהליך זה, אנו משתמשים בהחלפה אחורית וקדמית לפתרון המערכת $\mathbf{A}\underline{x}=\underline{b}$ ע"י ניצול הפירוק ל-LU. שני שלבים אלה בעלי כמות חישובים זהה, אם נתעלם מהעובדה שהאלכסון הראשי של **L** כולל רק 1-ים ולכן קל לטפל בו. בתהליך ההחלפה האחורית, חישוב האיבר ה-k דורש חישוב של

$$x_k = \frac{1}{u_{kk}}\left[b_k - \sum_{j=k+1}^{n} u_{jk} x_j\right].$$

לכן נדרשים לשלב זה (n-k) מכפלות וחיבורים ומנה אחת. בסיכום על פני כל המערכת יידרשו

$$\left[n + \sum_{k=1}^{n}(n-k)\right]_{[mul]} + \sum_{k=1}^{n}(n-k)_{[add]} = \frac{n^2+2n}{3}\,[mul] + \frac{n^2-n}{3}\,[add].$$

לסיכום, עבור n גדול דיו, פירוק LU דומיננטי והוא דורש סדר גודל של $n^3/3$ פעולות לצורך קבלת המרכיבים **L** ו-**U**. שימוש בפירוק זה לפתרון המערכת $\mathbf{A}\underline{x}=\underline{b}$ דורש סדר גודל של $2n^2/3$ פעולות (החלפה קדמית ואחורית) נוספות.

אם אנו מנהלים את האלימינציה הגאוסית, תהליך הכנת הצמד $[\mathbf{A},\underline{b}]$ וביצוע החלפה אחורית לפתרון המערכת עולה מעט יותר – אותה כמות חישובים תביא את **A** ליעדה (משולשת עליונה), וכמות של $5n^2/6$ נדרשת לשם עדכון b וביצוע ההחלפה האחורית לפתרון.

אנו נראה בהמשך כי שימוש בהיפוך של מטריצה מחייב כמות גדולה (פי 5 בקירוב) של פעולות, ולכן היא פחות יעילה משמעותית.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 2 – פירוקי ה- LDV וה-Cholesky

נושאים שייסקרו:

1. היפוך מטריצה – יסודות
2. אלימינציית גאוס-ג'ורדן – דמיון לתהליך ה- LU
3. ייצוג היפוך מטריצה לא סינגולארית כמכפלת מטריצות בסיסיות
4. פירוק ה- LDV
5. כמות החישובים בפירוק ה- LDV והשוואה להיפוך ישיר
6. מטריצות סימטריות ופירוק LDV עבורן
7. מטריצות חיוביות מוגדרות
8. מטריצות Gram
9. פירוק LDV של מטריצות חיוביות (חצי) מוגדרות
10. פירוק Cholesky

---

1. היפוך מטריצה – יסודות

בהינתן מטריצה ריבועית **A**, היפוכה תהיה המטריצה **X** אם היא מקיימת

$$\mathbf{AX} = \mathbf{XA} = \mathbf{I},$$

כאשר **I** הינה מטריצת היחידה – מטריצה אלכסונית עם 1-ים על אלכסונה הראשי. בהגדרה זו יש יתירות ודי בצד אחד שלה, אך משיקולי סימטריה, ובפזילה למטריצות מלבניות אנו מציגים הגדרה דו-צדדית זו.

אנו כבר יודעים כי למטריצה קיים היפוך רק אם היא אינה סינגולארית. כמו כן אנו יודעים כי אם המטריצה אינה סינגולארית, קיים היפוך והוא יחיד. אגב – אם **X** הוא ההיפוך של **A**, אזי ההיפוך של **X** הוא **A**, כלומר,

$$\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}.$$

קיימת התכונה הבאה:

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1},$$

וקל לוודא נכונות תכונה זו, כיוון ש:

$$\mathbf{I} = (\mathbf{AB})^{-1}(\mathbf{AB}) = \mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{AB} = \mathbf{B}^{-1}\left(\mathbf{A}^{-1}\mathbf{A}\right)\mathbf{B} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$$
$$\mathbf{I} = (\mathbf{AB})(\mathbf{AB})^{-1} = \mathbf{ABB}^{-1}\mathbf{A}^{-1} = \mathbf{A}\left(\mathbf{BB}^{-1}\right)\mathbf{A}^{-1} = \mathbf{AA}^{-1} = \mathbf{I}$$

חשוב להזהיר כי $(\mathbf{A}+\mathbf{B})^{-1} \neq \mathbf{A}^{-1} + \mathbf{B}^{-1}$ - למעשה, זה לא נכון אפילו לסקלרים.

2. אלימינציית גאוס-ג'ורדן – דמיון לתהליך ה- LU

שיטה מוכרת לבניית מטריצת ההיפוך היא שיטת האלימינציה של גאוס-ג'ורדן, אשר דומה מאוד לשיטת האלימינציה שממנה נבע פירוק LU. אנו מעוניינים למצוא את **X**

שתקיים את השוויון $\mathbf{AX} = \mathbf{I}$. ברישום מפורש, דרישה זו כמוה כמו רישום של n מערכות של משוואות מהצורה

$$k = 1,2,\ \dots\ ,n \qquad \mathbf{A}\underline{x}_k = \underline{e}_k$$

כאשר

$$,\underline{e}_1 = \begin{bmatrix}1\\0\\0\\0\\\vdots\\0\end{bmatrix},\quad \underline{e}_2 = \begin{bmatrix}0\\1\\0\\0\\\vdots\\0\end{bmatrix},\quad \underline{e}_3 = \begin{bmatrix}0\\0\\1\\0\\\vdots\\0\end{bmatrix},\quad \underline{e}_4 = \begin{bmatrix}0\\0\\0\\1\\\vdots\\0\end{bmatrix},\quad \cdots \quad \underline{e}_n = \begin{bmatrix}0\\0\\0\\0\\\vdots\\1\end{bmatrix}$$

ו- $\underline{x}_k$ הן העמודות של המטריצה **X**. לכן, נוכל לחזור על התהליך אותו עשינו קודם בו מוגדרת מטריצה מורחבת, אלא שבמקום עמודת $\underline{b}$ אחת יוצבו n העמודות הנ"ל, ופעולות האלימינציה יופעלו על כולן יחדיו. נמחיש זאת דרך הדוגמה הבאה – ברצוננו להפוך את המטריצה

$$.\begin{bmatrix}0 & 2 & 1\\1 & 6 & 1\\1 & 1 & 4\end{bmatrix}$$

לצורך כך נגדיר מטריצה מורחבת

$$,\left[\begin{array}{ccc|ccc}0 & 2 & 1 & 1 & 0 & 0\\2 & 6 & 1 & 0 & 1 & 0\\1 & 1 & 4 & 0 & 0 & 1\end{array}\right]$$

ועליה נבצע פעולות בסיסיות כגון החלפת שורות, או צירוף ליניארי שלהן. צעד ראשון יהיה החלפת סדר השורות על מנת לאפשר pivoting,

$$.\begin{bmatrix}0 & 1 & 0\\1 & 0 & 0\\0 & 0 & 1\end{bmatrix}\left[\begin{array}{ccc|ccc}0 & 2 & 1 & 1 & 0 & 0\\2 & 6 & 1 & 0 & 1 & 0\\1 & 1 & 4 & 0 & 0 & 1\end{array}\right] = \left[\begin{array}{ccc|ccc}2 & 6 & 1 & 0 & 1 & 0\\0 & 2 & 1 & 1 & 0 & 0\\1 & 1 & 4 & 0 & 0 & 1\end{array}\right]$$

כעת נאפס את האיבר a_{31},

$$.\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\-0.5 & 0 & 1\end{bmatrix}\left[\begin{array}{ccc|ccc}2 & 6 & 1 & 0 & 1 & 0\\0 & 2 & 1 & 1 & 0 & 0\\1 & 1 & 4 & 0 & 0 & 1\end{array}\right] = \left[\begin{array}{ccc|ccc}2 & 6 & 1 & 0 & 1 & 0\\0 & 2 & 1 & 1 & 0 & 0\\0 & -2 & 3.5 & 0 & -0.5 & 1\end{array}\right]$$

וכצעד אחרון לפני קבלת מטריצה משולשת עליונה, נאפס את a_{32}:

$$.\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 1 & 1\end{bmatrix}\left[\begin{array}{ccc|ccc}2 & 6 & 1 & 0 & 1 & 0\\0 & 2 & 1 & 1 & 0 & 0\\0 & -2 & 3.5 & 0 & -0.5 & 1\end{array}\right] = \left[\begin{array}{ccc|ccc}2 & 6 & 1 & 0 & 1 & 0\\0 & 2 & 1 & 1 & 0 & 0\\0 & 0 & 4.5 & 1 & -0.5 & 1\end{array}\right]$$

בעקרון ניתן לעצור כאן – קיבלנו מטריצה משולשת עליונה, ולכן בשלושה תהליכי החלפה לאחור נפרדים נוכל לחלץ את עמודות **X**. כלומר עלינו לפתור בפועל את המערכות הבאות

$$\begin{bmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 4.5 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 4.5 \end{bmatrix}\begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -0.5 \end{bmatrix}.$$

$$\begin{bmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 4.5 \end{bmatrix}\begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

בשיטת האלימינציה של גאוס-ג'ורדן מוצעת גישה מעט שונה ויותר יעילה – במקום עצירה כאן והפעלת n תהליכי החלפה לאחור דומים כל כך, ניתן להמשיך את תהליך האלימינציה, כשהפעם המטרה היא איפוס האיברים מעל האלכסון בחלק השמאלי של המטריצה המורחבת. למעשה, אין זאת אלא גישה שיטתית לביצוע ההחלפה לאחור.

המטרה שלנו היא הגעה למבנה $[\mathbf{I}\,|\,\mathbf{X}]$. עולה כאן נקודה מעניינת – אם נגיע על –ידי פעולות שורה בסיסיות למצב בו החלק השמאלי הוא **I**, אזי שלושת מערכות המשוואות שתוארו קודם הופכות למשוואות טריוויאליות שבונות את עמודות **X**, ולכן יוצא שהחלק הימני אינו אלא ההיפוך בכבודו ובעצמו.

עד כה ראינו שתי פעולות בסיסיות הפועלות על שורות במטריצה – פרמוטציה המחליפה סדר, וצירוף ליניארי של משוואות. כעת נציג פעולה שלישית מאותה משפחה – הכפלת משוואה בקבוע שאינו אפס, המיוצגת ע"י מטריצה אלכסונית. שימוש בכלי זה לכל שורה (והסדר כאן לא משנה) ייתן בדוגמה שלנו

$$\begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 2/9 \end{bmatrix}\left[\begin{array}{ccc|ccc} 2 & 6 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 4.5 & 1 & -0.5 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 1 & 3 & 0.5 & 0 & 0.5 & 0 \\ 0 & 1 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 2/9 & -1/9 & 2/9 \end{array}\right].$$

כיוון שאנו חותרים להתקרב למטריצת יחידה בצד שמאל, פעולה זו ברוכה.

ההמשך דומה מאוד ל-pivoting שנעשה כבר, כשהפעם אנו עובדים מהעמודה האחרונה פנימה. שימוש ב- a_{33} כאיבר ציר יאפשר את איפוס האיברים מעליו ונקבל

$$\begin{bmatrix}1&0&0\\0&1&-0.5\\0&0&1\end{bmatrix}\left[\begin{array}{ccc|ccc}1&3&0.5&0&0.5&0\\0&1&0.5&0.5&0&0\\0&0&1&2/9&-1/9&2/9\end{array}\right]=\left[\begin{array}{ccc|ccc}1&3&0.5&0&0.5&0\\0&1&0&7/18&1/18&-1/9\\0&0&1&2/9&-1/9&2/9\end{array}\right],$$

ו-

$$\begin{bmatrix}1&0&-0.5\\0&1&0\\0&0&1\end{bmatrix}\left[\begin{array}{ccc|ccc}1&3&0.5&0&0.5&0\\0&1&0&7/18&1/18&-1/9\\0&0&1&2/9&-1/9&2/9\end{array}\right]=\left[\begin{array}{ccc|ccc}1&3&0&-1/9&5/9&-1/9\\0&1&0&7/18&1/18&-1/9\\0&0&1&2/9&-1/9&2/9\end{array}\right].$$

באופן דומה, שימוש ב- a_{22} כציר תיתן

$$\begin{bmatrix}1&-3&0\\0&1&0\\0&0&1\end{bmatrix}\left[\begin{array}{ccc|ccc}1&3&0&-1/9&5/9&-1/9\\0&1&0&7/18&1/18&-1/9\\0&0&1&2/9&-1/9&2/9\end{array}\right]=\left[\begin{array}{ccc|ccc}1&0&0&-23/18&7/18&2/9\\0&1&0&7/18&1/18&-1/9\\0&0&1&2/9&-1/9&2/9\end{array}\right].$$

וכעת בידינו ההיפוך,

$$\mathbf{A}^{-1}=\begin{bmatrix}-23/18&7/18&2/9\\7/18&1/18&-1/9\\2/9&-1/9&2/9\end{bmatrix}.$$

תכונה מעניינת שסייעה לנו בתהליך אחרון זה היא שכיוון שהתחלנו ממטריצה משולשת עליונה, פעולות השורה שנעשו לא "משחיתות" את תוכן המשולש התחתון, והוא נותר מלא באפסים, ובאופן דומה האלכסון הראשי נותר עם '1' בשל שימוש בפעולות בסיסיות.

3. ייצוג היפוך מטריצה לא סינגולארית כמכפלת מטריצות בסיסיות

התהליך הנ"ל כלל הכפלה של מטריצת המקור **A** בפעולות בסיסיות ממספר סוגים:

- מטריצות פרמוטציה
- מטריצות משולשות תחתונות בסיסיות,
- מטריצה אלכסונית
- מטריצות משולשות עליונות בסיסיות,

כשבסיום הפעלתם הגענו למטריצת היחידה. כלומר, מתקיים הקשר

$$\mathbf{E}_N\mathbf{E}_{N-1}\cdots\mathbf{E}_2\mathbf{E}_1\mathbf{A}=\mathbf{I},$$

ולכן ההיפוך נתון ע"י

$$\mathbf{A}^{-1}=(\mathbf{E}_N\mathbf{E}_{N-1}\cdots\mathbf{E}_2\mathbf{E}_1)^{-1}=\mathbf{E}_1^{-1}\mathbf{E}_2^{-1}\cdots\mathbf{E}_{N-1}^{-1}\mathbf{E}_N^{-1}.$$

המסקנה מכאן היא שמטריצה שאינה סינגולארית ניתנת לרישום כסדרה של פעולות בסיסיות מהסוגים הנ"ל.

4. פירוק ה- LDV

נניח כי את התהליך הנ"ל ביצענו על מטריצה "רגילה" – כזכור – זוהי מטריצה שאין בה צורך בהחלפת סדר שורות במהלך תהליך האלימינציה הנ"ל. סידרת פעולות ראשונה אחראית להבאת המטריצה $\left[\mathbf{A} \mid \mathbf{I}\right]$ למבנה בו החלק השמאלי הינו מטריצה משולשת עליונה,

$$\mathbf{E}_{N_1}\cdots\mathbf{E}_2\mathbf{E}_1\left[\mathbf{A} \mid \mathbf{I}\right]=\left[\begin{array}{ccccc|ccc} * & * & * & \cdots & * & & & \\ 0 & * & * & \cdots & * & \ddots & \vdots & \iddots \\ 0 & 0 & * & \cdots & * & \cdots & \mathbf{Z}_1 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \iddots & \vdots & \ddots \\ 0 & 0 & 0 & \cdots & * & & & \end{array}\right].$$

סידרת פעולות אלה כולה בנויה על מטריצות בסיסיות בהן האלכסון הראשי כולל 1-ים והן משולשות תחתונות, ולכן

$$\mathbf{L} = \left(\mathbf{E}_{N_1}\cdots\mathbf{E}_2\mathbf{E}_1\right)^{-1},$$

גם היא מטריצה משולשת תחתונה ובסיסית (ואף ראינו כי בנייתה טריוויאלית וכרוכה באיסוף איברים מתחת לאלכסון).

סידרת פעולות שנייה מכפילה כל שורה בגורם מתקן שאינו אפס על מנת להביא את איבר האלכסון במטריצה המשולשת העליונה ל-1-ים,

$$\mathbf{D}^{-1}\mathbf{E}_{N_1}\cdots\mathbf{E}_2\mathbf{E}_1\left[\mathbf{A} \mid \mathbf{I}\right]=\left[\begin{array}{ccccc|ccc} 1 & * & * & \cdots & * & & & \\ 0 & 1 & * & \cdots & * & \ddots & \vdots & \iddots \\ 0 & 0 & 1 & \cdots & * & \cdots & \mathbf{Z}_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \iddots & \vdots & \ddots \\ 0 & 0 & 0 & \cdots & 1 & & & \end{array}\right].$$

המטריצה **D** היא מטריצה אלכסונית בה איברי האלכסון אינם אלא איברי המטריצה המשולשת העליונה לפני התיקון. לכן היפוכם נותן נרמול לאברי האלכסון במבנה החדש.

קבוצת פעולות שלישית ואחרונה דומה לראשונה ובה מבוצעות פעולות בסיסיות ע"י מטריצות משולשות עליונות, ומושגת מטריצת יחידה,

$$\mathbf{F}_{N_2}\cdots\mathbf{F}_2\mathbf{F}_1\mathbf{D}^{-1}\mathbf{E}_{N_1}\cdots\mathbf{E}_2\mathbf{E}_1\left[\mathbf{A} \mid \mathbf{I}\right]=\left[\begin{array}{ccccc|ccc} 1 & 0 & 0 & \cdots & 0 & & & \\ 0 & 1 & 0 & \cdots & 0 & \ddots & \vdots & \iddots \\ 0 & 0 & 1 & \cdots & 0 & \cdots & \mathbf{X} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \iddots & \vdots & \ddots \\ 0 & 0 & 0 & \cdots & 1 & & & \end{array}\right].$$

סידרת פעולות אלו עושה שימוש במטריצות בסיסיות משולשות עליונות ולכן גם מכפלתן כזו וגם היפוכן,

$$.\left(\mathbf{F}_{N_2}\cdots\mathbf{F}_2\mathbf{F}_1\right)^{-1}=\mathbf{V}$$

מעניין לציין כי אוסף הפעולות האחרון נתן בפועל הכפלה של המטריצה המורחבת בהיפוך של המטריצה המשולשת העליונה הבסיסית שהייתה בתחילת שלב זה.

מכל האמור לעיל נובע כי המטריצה A ניתנת לתיאור כפירוק

$$.\left(\mathbf{F}_{N_2}\cdots\mathbf{F}_2\mathbf{F}_1\right)\mathbf{D}^{-1}\left(\mathbf{E}_{N_1}\cdots\mathbf{E}_2\mathbf{E}_1\right)\mathbf{A}=\mathbf{I}\;\Rightarrow\;\mathbf{A}=\mathbf{LDV}$$

עבור הדוגמה ממקודם, בה רצינו לפרק את המטריצה

$$,\mathbf{A}=\begin{bmatrix}2&6&1\\0&2&1\\1&1&4\end{bmatrix}$$

אוסף הפעולות הראשון היה

$$,\begin{bmatrix}1&0&0\\0&1&0\\0&1&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&1&0\\-0.5&0&1\end{bmatrix}\left[\begin{array}{ccc|ccc}2&6&1&0&1&0\\0&2&1&1&0&0\\1&1&4&0&0&1\end{array}\right]=\left[\begin{array}{ccc|ccc}2&6&1&0&1&0\\0&2&1&1&0&0\\0&0&4.5&1&-0.5&1\end{array}\right]$$

ולכן

$$.\mathbf{L}=\left(\begin{bmatrix}1&0&0\\0&1&0\\0&1&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&1&0\\-0.5&0&1\end{bmatrix}\right)^{-1}=\begin{bmatrix}1&0&0\\0&1&0\\0.5&0&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&1&0\\0&-1&1\end{bmatrix}=\begin{bmatrix}1&0&0\\0&1&0\\0.5&-1&1\end{bmatrix}$$

המטריצה **D** מהדוגמה הנ"ל הינה

$$,\mathbf{D}=\begin{bmatrix}2&0&0\\0&2&0\\0&0&4.5\end{bmatrix}$$

והמטריצה **V** נתונה ע"י החלק בצד השמאלי לאחר הנרמול,

$$.\mathbf{V}=\begin{bmatrix}1&3&0.5\\0&1&0.5\\0&0&1\end{bmatrix}$$

לא נדרש כלל לאסוף את הפעולות הבסיסיות שבונות את ההיפוך של **V** ומהן לחשב את **V** – זו נתונה מלכתחילה. הסיבה לאותן פעולות בסיסיות היא הרצון לקבל באופן מפורש את היפוך המטריצה

ננסח תוצאת הפירוק המתקבל כמשפט.

משפט 1: מטריצה "רגילה" **A** ניתנת לפירוק יחיד מהצורה **A**=**LDV**, כש-**L** מטריצה משולשת תחתונה בסיסית, **V** מטריצה משולשת עליונה בסיסית, ו-**D** מטריצה אלכסונית כלשהי עם איברים שונים מאפס באלכסון הראשי.

ישנו כמובן קשר הדוק בין פירוק ה- LU לפירוק ה-LDV – אם למטריצה **A** ידוע לנו פירוקה ל-**L** ו-**U**, הרי שפירוק ה- LDV שלה קל לקבלה ע"י בניית **D** מאברי האלכסון של **U** ונרמול שורות **U** בהתאם לקבלת מרכיב ה-**V**. חשוב להבהיר כי הכפלת מטריצה כלשהי משמאל ע"י מטריצה אלכסונית יוצרת הכפלה של כל שורה במקדם זהה, כפי שקרה כאן (בניגוד להכפלה מימין שתיתן אפקט דומה לעמודות). לכן גם ברור מה יקרה במטריצות שאינן "רגילות". זה מביא אותנו למשפט הבא:

משפט 2: כל מטריצה **A** ניתנת לפירוק מהצורה **PA**=**LDV**, כש-**L** מטריצה משולשת תחתונה בסיסית, **U** מטריצה משולשת עליונה בסיסית, ו-**D** מטריצה אלכסונית. אם איברי האלכסון הראשי של **D** כולם שונים מאפס, **A** אינה סינגולארית.

טענו קודם כי הדטרמיננטה של מטריצה **A** ניכר בקלות מאלכסונה של המטריצה **U** בפירוק LU. ובכן, כעת זה אפילו קל יותר, כיוון ש:

$$\det\{\mathbf{A}\} = \det\{\mathbf{LDV}\} = \det\{\mathbf{L}\}\det\{\mathbf{D}\}\det\{\mathbf{V}\} = \det\{\mathbf{D}\} = \prod_{k=1}^{n} d_{kk}.$$

5. כמות החישובים בפירוק ה- LDV והשוואה להיפוך ישיר

כמה חישובים דרושים לקבלת פירוק ה- LDV? כמעט כמו בפירוק ה-LU! נניח כי המדובר במטריצה "רגילה". התחלת התהליך זהה לפירוק ה-LU ודורשת

$$\sum_{k=1}^{n} k(k-1)\,[\text{mul}] + \sum_{k=1}^{n} (k-1)^2\,[\text{add}] = \frac{n^3-n}{3}\,[\text{mul}] + \frac{2n^3-3n^2+n}{6}\,[\text{add}]$$

לבניית **L** ו-**U**. לאחר סיום תהליך זה, אנו מחלקים כל שורה באיבר האלכסון הראשי ב-**U**, פעולה הדורשת

$$\sum_{k=1}^{n} k\,[\text{mul}] = \frac{n^2-n}{2}\,[\text{mul}]$$

והיא זו שנותנת לנו את **D** ואת **V**. לכן, עדיין יידרשו סדר גודל של $n^3/3$ פעולות לבניית פירוק זה. כזכור, בהינתן פירוק זה, שימוש בו לפתרון מערכת המשוואות $\mathbf{A}\underline{x}=\underline{b}$ דורשת עוד $2n^2/3$ פעולות.

כמה חישובים נחוצים לשם בנייה מלאה של מטריצת ההיפוך בגישת האלימינציה של גאוס-ג'ורדן? בחלק הראשון אנו בונים את סידרת הפעולות שתביא את המערכת למבנה משולש עליון, אך הפעם החלק הנלווה במטריצה המורחבת אשר עובר את כל התהליך כבר אינו עמודה יחידה אלא בגודל של n-על-n. לכן, בנוסף ל-

$$\sum_{k=1}^{n} k(k-1)\,_{[mul]} + \sum_{k=1}^{n} (k-1)^2\,_{[add]} = \frac{n^3-n}{3}\,_{[mul]} + \frac{2n^3-3n^2+n}{6}\,_{[add]}$$

פעולות שמתייחסות לחלק השמאלי במטריצה [**A** | **I**], יידרשו

$$n\sum_{k=1}^{n} (k-1)\,_{[mul]} + n\sum_{k=1}^{n} (k-1)\,_{[add]} = \frac{n^3-n^2}{2}\,_{[mul]} + \frac{n^3-n^2}{2}\,_{[add]}$$

פעולות (השתמשנו באומדן מפרק קודם לעדכון העמודה b כשהוא מוכפל ב-n העמודות הדורשות עדכון). אגב, בביטוי זה הזנחנו את היות המטריצה ההתחלתית מטריצת היחידה – עובדה שמקילה במידת מה את כמות החישובים.

החלק הבא בתהליך בו מנורמלות השורות זניח כי כמות החישובים בו מסדר גודל של n^2. החלק האחרון דומה לראשון ודורש $n^3/3$ פעולות להבאת החלק השמאלי במטריצה המורחבת למבנה אלכסוני. גם הפעם יידרשו סדר גודל של $n^3/2$ פעולות לטיפול דומה בחלק הימני.

לסיכום, עבור n גדול דיו, היפוך המטריצה המפורש דורש סדר גודל של $5n^3/3$ פעולות. לכן, כפי שכבר הוזכר קודם, שימוש ב- LU יעיל פי 5 יותר. מעבר לכך, שימוש במטריצה ההופכית לחישוב הפתרון למערכת $\mathbf{A}\underline{x}=\underline{b}$ ידרוש n2 מכפלות וסיכומים – מעט יותר מהכמות הנדרשת בשיטות ההחלפה בשימוש ב-LU. המסקנה המתבקשת היא ששיטת פירוק LU כגישה לפתרון מערכות של משוואות יעילה יותר מהיפוך ישיר ושימוש בו.

6. מטריצות סימטריות ופירוק LDV עבורן

בעבודה עם מטריצות ריבועיות יש לנו עניין רב במטריצות שהינן סימטריות או אנטי-סימטריות. מסתבר כי במדעים מטריצות כאלה עולות לעיתים קרובות בחלק מהגדרת הבעיה. בהינתן מטריצה **A** ריבועית בגודל n-על-n, מטריצה זו סימטרית אם

$$\mathbf{A} = \mathbf{A}^T,$$

כלומר,

$$\forall 1 \le k, j \le n \quad a_{kj} = a_{jk}.$$

נשים לב כי אנו דנים במטריצות מעל הממשיים כאן – כשעוסקים במספרים מרוכבים, יש להתלבט בין פעולת Transpose לבין פעולת Conjugate-Transpose, אך נניח לעניין זה כעת. באופן דומה, אנטי-סימטריות מוגדרת כמטריצה המקיימת

$$\mathbf{A} = -\mathbf{A}^T \quad \Rightarrow \quad \forall 1 \le k, j \le n \ \ a_{kj} = -a_{jk} .$$

מהגדרה זו ברור כי אברי האלכסון הראשי חייבים להיות מאופסים כדי לתת אנטי-סימטריות.

כל מטריצה ניתנת לפירוק (ובקלות) לנתח סימטרי ונתח שני שהינו אנטי סימטרי. זאת מקבלים ע"י הפעולה הטריוויאלית

$$\mathbf{A} = \frac{\mathbf{A}+\mathbf{A}^T}{2} + \frac{\mathbf{A}-\mathbf{A}^T}{2} .$$

החלק הראשון סימטרי כיוון ש:

$$\left(\frac{\mathbf{A}+\mathbf{A}^T}{2}\right)^T = 0.5\left(\mathbf{A}+\mathbf{A}^T\right)^T = 0.5\left(\mathbf{A}^T+\mathbf{A}\right) = \frac{\mathbf{A}+\mathbf{A}^T}{2} .$$

באופן דומה, החלק השני הוא אנטי-סימטרי כיוון ש:

$$\left(\frac{\mathbf{A}-\mathbf{A}^T}{2}\right)^T = 0.5\left(\mathbf{A}-\mathbf{A}^T\right)^T = 0.5\left(\mathbf{A}^T-\mathbf{A}\right) = -\frac{\mathbf{A}-\mathbf{A}^T}{2} .$$

איך מתקבלים הפירוקים LU ו-LDV עבור מטריצות מיוחדות אלו?

משפט 3: כל מטריצה סימטרית **A** ניתנת לפירוק מהצורה $\mathbf{PAP}^T=\mathbf{LDL}^T$, כאשר **L** מטריצה משולשת תחתונה בסיסית, **D** מטריצה אלכסונית, ו-**P** מטריצת פרמוטציה. אם המטריצה סימטרית ו-"רגילה", פירוקים אלה הם יחידים עבור **P=I**.

נשים לב לכך שעל מנת לקבל פירוק סימטרי היה עלינו לסדר לא רק את השורות אלא גם את העמודות על מנת לשמר מבנה סימטרי. באשר למבנה אנטי-סימטרי, לא ניתן לטעון דבר מה בעל ערך.

7. מטריצות חיוביות מוגדרות

תת קבוצה מעניינת מתוך אוסף המטריצות הסימטריות היא קבוצת המטריצות החיוביות מוגדרות. מטריצה **K** סימטרית בגודל n-על-n מוגדרת כמטריצה חיובית מוגדרת (PD - Positive Definite) אם לכל וקטור $\underline{x}$ יתקיים כי

$$\forall \underline{x} \ne 0, \quad \underline{x}^T \mathbf{K} \underline{x} > 0 .$$

אנו נסמן תכונה זו כ- $\mathbf{K} > 0$. באופן זה הגדרת החיוביות המוגדרת אינה אלא הכללה למטריצות של רעיון החיוביות של משתנה סקלארי.

חשוב להבהיר כי אין פירוש הדבר שכל כניסות המטריצה חיוביים. נמחיש זאת דרך דוגמה ראשונה בא נראה מטריצה סימטרית שכל כניסותיה חיוביות ולמרות זאת היא אינה חיובית מוגדרת. נתייחס למטריצה

$$\mathbf{K} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

אם נמצא וקטור $\underline{x}$ שונה מאפס שמאפס את התבנית הריבועית $\underline{x}^T\mathbf{K}\underline{x}$, הרי שהראינו כי היא אינה חיובית מוגדרת. נציע את הווקטור $\underline{x}^T = \begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix}$. ברור כי $\mathbf{K}\underline{x} = 0$ ולכן גם נובע מכך ישירות כי $\underline{x}^T\mathbf{K}\underline{x} = 0$. נשים לב לתכונה מעניינת – אם למטריצה **K** קיים Null-Space שאינו ריק (תת מרחב של וקטורים שמוטלים לאפס בעקבות הכפלה ב-**K**), הרי שהמטריצה בהכרח לא חיובית מוגדרת. לכן ברור כי מטריצה סינגולארית לא יכולה להיות מטריצה חיובית מוגדרת. ננסח זאת כמשפט:

משפט 4: אם **K** הינה PD אזי בהכרח היא אינה סינגולארית.

נניח כי בידינו המטריצה הסימטרית

$$K = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix},$$

ורצוננו לקבוע את טווח ערכי הפרמטר a שיבטיחו כי המטריצה חיובית מוגדרת. דרך דוגמה זו אנו נצליח להראות כי ייתכנו ערכים שליליים במטריצה ועדיין זו תהיה חיובית מוגדרת. נדרוש

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + 2ax_1x_2 + x_2^2 > 0 \qquad \forall\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \neq 0,$$

ובניסוח שונה מעט זה יהיה

$$x_1^2 + 2ax_1x_2 + x_2^2 = (x_1 + ax_2)^2 + \left(1 - a^2\right)x_2^2 > 0.$$

אם $|a| < 1$, כל צמד שונה מאפס (די באחד מהם!) של $x_{1,2}$ ייתן כי דרישה זו מתקיימת, ולכן זוהי הדרישה לחיוביות מוגדרת. אם למשל $a = 1$ אז בחירה שתיתן $x_1 + ax_2 = 0$, כלומר $x_1 = -x_2$, תאפס התבנית. עבור $a > 1$ אותה בחירה אף תיתן ערך שלילי. כך גם יקרה ל- $a \le -1$. התהליך הנ"ל לבחינת PD מעניין ואנו עוד נשוב אליו.

הרחבת מה של המושג "חיוביות מוגדרת" הוא המושג "חיוביות חצי מוגדרת" בה מותר שוויון לאפס. לכן, **K** הינה Positive Semi-Definite (PSD) אם הי מקיימת

$$\underline{x}^T \mathbf{K} \underline{x} \geq 0 \qquad \forall \underline{x} \neq 0,\ .$$

אנו נסמן תכונה זו כ- $\mathbf{K} \geq 0$. ברוח הדוגמאות קודם נוכל לומר כי המטריצה

$$\mathbf{K} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

אמנם אינה חיובית מוגדרת אך היא עדיין יכולה להיות PSD. נראה שזה אמנם כך! לא קשה להשתכנע בנכונות השוויון הבא:

$$\underline{x}^T \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \underline{x} = \underline{x}^T \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \underline{x} = \left\| \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \underline{x} \right\|_2^2 \geq 0\ .$$

באשר לדוגמה עם המטריצה תלוית ה-a, זו תהיה חיובית חצי מוגדרת עבור $|a| \leq 1$, כלומר, כל ההבדל הוא בכלילת קצוות האינטרוול [-1,1].

8. מטריצות Gram

למה מטריצות חיוביות (חצי) מוגדרות כה פופולאריות? משום שיש להן קשר הדוק לאפיון של בעיות אופטימיזציה (ואת זה עוד נראה בהמשך), מכפלות פנימיות, ועוד. מקור לא אכזב של מטריצות חיוביות (חצי) מוגדרות הוא המבנה

$$\mathbf{K} = \mathbf{A}^T \mathbf{A},$$

כש-**A** מטריצה כלשהי ולאו דווקא ריבועית! מבנים כאלה קרויים מטריצות Gram (ע"ש מתמטיקאי דני בשם יורגן גרם מהמאה ה-19). כל מטריצת Gram היא בהכרח חיובית חצי מוגדרת כיוון ש:

$$\underline{x}^T \mathbf{K} \underline{x} = \underline{x}^T \mathbf{A}^T \mathbf{A} \underline{x} = (\mathbf{A}\underline{x})^T (\mathbf{A}\underline{x}) = \|\mathbf{A}\underline{x}\|_2^2 \geq 0 \qquad \forall \underline{x},\ .$$

יתרה מזו, על מנת שתבנית ה-Gram תוביל למטריצה חיובית מוגדרת, עלינו לדרוש שלמערכת המשוואות ההומוגנית $\mathbf{A}\underline{x} = 0$ יש רק פתרון טריוויאלי $\underline{x} = 0$, וזה שקול לאמירה שעמודות **A** בלתי תלויות ליניארית. ננסח זאת כמשפט:

משפט 5: כל מטריצת Gram, $\mathbf{K} = \mathbf{A}^T \mathbf{A}$, חיובית חצי מוגדרת מעצם הגדרתה. אם עמודות המטריצה **A** בלתי-תלויות-ליניארית, מטריצת ה-Gram חיובית מוגדרת.

אם נסתכל על האיבר k_{ij} במטריצה $\mathbf{K} = \mathbf{A}^T\mathbf{A}$, ערכו שווה

$$k_{ij} = \sum_{m=1}^{n} a_{im} a_{jm} = \underline{a}_i^T \underline{a}_j ,$$

כאשר סימנו ב-$\underline{a}_i$ את העמודה ה-i-ית במטריצה **A**. כלומר, האיבר ה-ij אינו אלא תוצאת המכפלה הפנימית הרגילה בין העמודה ה-i לעמודה ה-j. לכן, אם ברשותנו מקבץ של n וקטורים $\underline{v}_j$ ($j = 1,2,\ldots,n$) ממימד N ונבנה מטריצה מהצורה

$$\mathbf{K} = \begin{bmatrix} \langle \underline{v}_1, \underline{v}_1 \rangle & \langle \underline{v}_1, \underline{v}_2 \rangle & \cdots & \langle \underline{v}_1, \underline{v}_n \rangle \\ \langle \underline{v}_2, \underline{v}_1 \rangle & \langle \underline{v}_2, \underline{v}_2 \rangle & \cdots & \langle \underline{v}_2, \underline{v}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{v}_n, \underline{v}_1 \rangle & \langle \underline{v}_n, \underline{v}_2 \rangle & \cdots & \langle \underline{v}_n, \underline{v}_n \rangle \end{bmatrix},$$

אזי מטריצה זו חיובית חצי מוגדרת לכל מכפלה פנימית חוקית. לתזכורת, מכפלה פנימית מקיימת את האקסיומות והתכונות הבאות (בהנחה שההגדרה היא מעל הממשיים):

- $\langle a\underline{v}_1 + b\underline{v}_2, \underline{u} \rangle = a\langle \underline{v}_1, \underline{u} \rangle + b\langle \underline{v}_2, \underline{u} \rangle$ ו- $\langle \underline{u}, a\underline{v}_1 + b\underline{v}_2 \rangle = a\langle \underline{u}, \underline{v}_1 \rangle + b\langle \underline{u}, \underline{v}_2 \rangle$ - תכונת הבי-ליניאריות.
- $\langle \underline{v}, \underline{u} \rangle = \langle \underline{u}, \underline{v} \rangle$ - סימטריות.
- לכל $\underline{u}$ שונה מאפס, $\langle \underline{u}, \underline{u} \rangle = \|\underline{u}\|^2 > 0$.
- אי-שוויון קושי-שוורץ נובע מהנ"ל: $|\langle \underline{v}, \underline{u} \rangle| \le \|\underline{v}\| \cdot \|\underline{u}\|$.
- גם אי-שוויון המשולש נובע מהנ"ל: $\|\underline{v} + \underline{u}\| \le \|\underline{v}\| + \|\underline{u}\|$

נחזור לענייננו, אם אוסף n הוקטורים בלתי תלויים ליניארית, זו תהיה מטריצה חיובית מוגדרת. לצורך כך יהיה עלינו לדרוש ש-$n \le N$, שאם לא כן, אוסף הוקטורים חייב להיות תלוי ליניארית.

משפט 6: כל מטריצה **K** הנבנית ע"י אוסף מכפלות פנימיות בין n וקטורים לעצמם הינה בהכרח חיובית חצי מוגדרת. אם הוקטורים בלתי תלויים ליניארית, תהיה זו מטריצה חיובית מוגדרת.

להוכחת תכונה זו, די לשים לב לשוויון הבא:

$$\underline{x}^T\mathbf{K}\underline{x} = \underline{x}^T\begin{bmatrix}\langle\underline{v}_1,\underline{v}_1\rangle & \langle\underline{v}_1,\underline{v}_2\rangle & \cdots & \langle\underline{v}_1,\underline{v}_n\rangle\\ \langle\underline{v}_2,\underline{v}_1\rangle & \langle\underline{v}_2,\underline{v}_2\rangle & \cdots & \langle\underline{v}_2,\underline{v}_n\rangle\\ \vdots & \vdots & \ddots & \vdots\\ \langle\underline{v}_n,\underline{v}_1\rangle & \langle\underline{v}_n,\underline{v}_2\rangle & \cdots & \langle\underline{v}_n,\underline{v}_n\rangle\end{bmatrix}\underline{x} =$$

$$= \left\langle \sum_{j=1}^{n} x_j\underline{v}_j, \sum_{j=1}^{n} x_j\underline{v}_j \right\rangle = \left\| \sum_{j=1}^{n} x_j\underline{v}_j \right\|_2^2 \geq 0 \ .$$

לכן, ככלל תמיד ביטוי זה יהיה אי-שלילי ומכאן נובעת החיוביות החצי מוגדרת. אם n הוקטורים בקבוצה בלתי-תלויים-ליניארית פירוש הדבר שלא קיימים מקדמים $\underline{x}$ שייתנו כי הווקטור הנוצר מתאפס. כיוון שנורמה של וקטור שונה מאפס הינה בהכרח גדולה מאפס, תתקבל החיוביות המוגדרת.

דוגמה מעניינת להמחשת התוצאה הנ"ל היא המטריצה

$$\mathbf{K} = \begin{bmatrix} 1/1 & 1/2 & \cdots & 1/n\\ 1/2 & 1/3 & \cdots & 1/(n+1)\\ \vdots & \vdots & \ddots & \vdots\\ 1/n & 1/(n+1) & & 1/(2n-1)\end{bmatrix} .$$

מטריצה זו ידועה בשם מטריצת הילברט. האם היא חיובית מוגדרת? אם נמצא מקור למטריצה זו כאוסף של מכפלות פנימיות בין וקטורים בלתי-תלויים-ליניארית, נוכל לקבוע שאמנם היא חיובית מוגדרת, וממילא לא סינגולארית. אינטואיטיבית נוח לנו לחשוב על וקטורים כ-n-יות של סקלרים, אך מרחבים וקטוריים יכולים לכלול פונקציות ומרכיבים אבסטרקטיים אחרים. נסתכל על אוסף המונומיאלים (פולינומים עם מרכיב חזקה יחיד)

$$\left\{x^j\right\}_{j=0}^{n-1}$$

המוגדרים מעל האינטרוול [0,1]. משפחת פונקציות אלו יוצרת בקומבינציות ליניאריות את כל הפולינומים מסדר n-1 ומטה, והיא אינה תלויה ליניארית (לא ניתן לייצר פונקציה אחת מאוסף זה ע"י שימוש בצירוף ליניארי של האחרות). מכפלה פנימית תוגדר ע"י

$$\forall i,j \geq 0,\ \ \left\langle x^i, x^j \right\rangle = \int_0^1 x^{i+j}dx = \frac{1}{i+j+1}\left[x^{i+j+1}\right]_0^1 = \frac{1}{i+j+1} \ .$$

אנו רואים כי מכפלות פנימיות אלו בונים את מטריצת הילברט בדיוק, ולכן נקיש כי זו מטריצה חיובית מוגדרת (וככזו, בהכרח לא סינגולארית).

למעשה, ניתן להכליל את עניין מטריצות ה-Gram ולומר כי אם **C** מטריצה חיובית חצי מוגדרת אז גם $\mathbf{K}=\mathbf{A}^T\mathbf{C}\mathbf{A}$ כזו. אם **C** מטריצה חיובית מוגדרת ו-**A** בעלת עמודות בלתי תלויות ליניארית, אזי **K** חיובית מוגדרת. זאת כיוון שניתן לרשום

$$\underline{x}^T\mathbf{A}^T\mathbf{C}\mathbf{A}\underline{x} \underset{\underset{\underline{y}=\mathbf{A}\underline{x}}{\uparrow}}{=} \underline{y}^T\mathbf{C}\underline{y}>0 .$$

ברישום זה השתמשנו בעובדה שעמודות **A** בלתי תלויות ליניארית ולכן הווקטור $\underline{y}$ אינו מתאפס. כמו כן, המטריצה **C** ידועה כחיובית מוגדרת, ולכן החיוביות.

9. פירוק LDV של מטריצות חיוביות (חצי) מוגדרות

מה הקשר בין מטריצות חיוביות (חצי) מוגדרות ופירוק LDV? ראינו קודם כי למטריצות סימטריות רגילות הפירוק הזה ניתן לאפיון מיוחד ע"י $\mathbf{K}=\mathbf{L}\mathbf{D}\mathbf{L}^T$. נניח כי נתונה לנו מטריצה **K** סימטרית בגודל n-על-n והיא פורקה למבנה

$$\mathbf{K}=\mathbf{L}\mathbf{D}\mathbf{L}^T=\mathbf{U}^T\mathbf{D}\mathbf{U} .$$

הרי המטריצה **U** (ה-Transpose של **L**) הנדונה כאן מכילה 1-ים על אלכסונה הראשי, וככזו היא אינה סינגולארית. לכן, המטריצה **K** בעלת מבנה של מטריצת Gram. אם **D** חיובית מוגדרת, אזי גם **K** כזו, כי עמודות **U** בלתי-תלויות ליניארית. כלומר, ברור כי אם הצלחנו בפירוק LDV של **K** ללא פרמוטציה, וקיבלנו מבנה $\mathbf{K}=\mathbf{U}^T\mathbf{D}\mathbf{U}$, ואברי האלכסון של **D** חיוביים, אזי **K** בהכרח חיובית מוגדרת. אם איברי אלכסונה של **D** כוללים גם אפסים, תהיה זו מטריצה חיובית חצי מוגדרת.

השאלה שנותרת היא – האם כל מטריצה חיובית מוגדרת ניתנת לפירוק כזה? אם התשובה היא חיובית, נקבל שיטה לבחינת היותה של מטריצה חיובית מוגדרת. מסתבר כי אמנם פירוק ה-LDV מספק מענה מוצלח לעניין זה. נמחיש זאת דרך הדוגמה הבאה – ניקח את המטריצה ונבחן אם היא חיובית מוגדרת:

$$\mathbf{K}=\begin{bmatrix}1&2&-1\\2&6&0\\-1&0&9\end{bmatrix} .$$

את הבחינה נעשה לפי הגדרה – נפרק את התבנית $\underline{x}^T\mathbf{K}\underline{x}$ ונבדוק את חיוביותה. מתקבל

$$\begin{aligned}\underline{x}^T\mathbf{K}\underline{x}=\underline{x}^T\begin{bmatrix}1&2&-1\\2&6&0\\-1&0&9\end{bmatrix}\underline{x}&=\begin{bmatrix}x_1&x_2&x_3\end{bmatrix}\begin{bmatrix}x_1+2x_2-x_3\\2x_1+6x_2\\-x_1+9x_3\end{bmatrix}=\\&=x_1^2+4x_1x_2-2x_1x_3+6x_2^2+9x_3^2\end{aligned} .$$

את הביטוי שהתקבל נקבץ לגורמים ריבועיים נוחים ונקבל

$$\begin{aligned}\underline{x}^T\mathbf{K}\underline{x} &= x_1^2+4x_1x_2-2x_1x_3+6x_2^2+9x_3^2 = \\ &= (x_1+2x_2-x_3)^2+2x_2^2+8x_3^2+4x_2x_3 = \\ &= (x_1+2x_2-x_3)^2+2(x_2+x_3)^2+6x_3^2\end{aligned} \quad .$$

בפירוק זה הגישה הייתה הבאה: תפקידו של הגורם הראשון להעלים כל הופעה של x_1. בגורם ראשון זה אנו עשויים לשנות איברים העושים שימוש ב-x_2 ו-x_3 , אך אלה יטופלו בהמשך. הגורם השני מסיר את כל האיברים המשתמשים ב-x_2, וכך הלאה.

אנו רואים כי הגענו לסיכום חיובי של אוסף תבניות ריבועיות כשכל אחת מייצגת צירוף ליניארי של ערכי הווקטור $\underline{x}$. לכן מטריצה זו חיובית מוגדרת. בשל דרך איסוף הביטויים, קיבלנו כי צירופים ליניאריים אלו בעלי מבנה משולש. התהליך הנ"ל יוכל להיכתב מעט אחרת כ:

$$\begin{aligned}\underline{x}^T\mathbf{K}\underline{x} &= (x_1+2x_2-x_3)^2+2(x_2+x_3)^2+6x_3^2 = \\ &= \begin{bmatrix}x_1 & x_2 & x_3\end{bmatrix}\begin{bmatrix}1&0&0\\2&1&0\\-1&1&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&2&0\\0&0&6\end{bmatrix}\begin{bmatrix}1&2&-1\\0&1&1\\0&0&1\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}\end{aligned} \quad .$$

והרי זה אינו אלא פירוק LDV למטריצה! המסקנה ניתנת לרישום כללי יותר כמשפט הבא:

משפט 7: מטריצה סימטרית **K** הינה חיובית מוגדרת ואם ורק אם היא ניתנת לפירוק LDV ללא פרמוטציה בו האלכסון של **D** כולל ערכים חיוביים.

משפט זה אומר בעצם כי מטריצה חיובית מוגדרת חייבת להיות "רגילה" – כלומר ניתנת לפירוק LU ללא צורך לשינויי סדר בשורותיה.

10.פירוק Cholesky

ממש כמו ש-"נחמד" להוציא שורש למספר ממשי וחיובי, יהיה זה נחמד אם נצליח לפרק מטריצה חיובית מוגדרת למכפלה $\mathbf{K}=\mathbf{A}^T\mathbf{A}$ - כלומר למצוא את ה-"שורש" שלה **A**. העניין הוא ששורש זה אינו יחיד! זאת כיוון שאם מצאנו פירוק כזה עם **A**, אז כל מטריצה אחרת $\mathbf{B}=\mathbf{QA}$ עם **Q** מטריצה ריבועית אורתונורמלית גם היא שורש אפשרי, כיוון ש

$$\mathbf{K}=\mathbf{A}^T\mathbf{A}=\mathbf{A}^T\mathbf{Q}^T\mathbf{Q}\mathbf{A}=\mathbf{B}^T\mathbf{B}.$$

ממגוון כלל הפירוקים האפשריים יש לנו עניין בפירוק מהצורה $\mathbf{K} = \mathbf{A}^T\mathbf{A}$, בו המטריצה **A** משולשת עליונה (או תחתונה). פירוק זה קרוי פירוק Cholesky ויש לו חשיבות רבה במגוון יישומים. ברור כי אם בנינו פירוק LDV למטריצה **K**, הרי שפירוק Cholesky נובע בפשטות והוא יחיד. ניקח את המטריצה **D** ונפרק אותה למכפלה

$$D = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix} =$$

$$= \begin{bmatrix} \sqrt{d_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix} \begin{bmatrix} \sqrt{d_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix} = \mathbf{S}\mathbf{S}^T.$$

פעולת השחלוף הינה פעולת סרק כאן, אך היא נוחה לרישום בהמשך. לכן, פירוק Cholesky מתקבל ע"י

$$\mathbf{K} = \mathbf{L}\mathbf{D}\mathbf{L}^T = \mathbf{L}\mathbf{S}\mathbf{S}^T\mathbf{L}^T = \mathbf{M}\mathbf{M}^T,$$

כאשר $\mathbf{M} = \mathbf{L}\mathbf{S}$ היא מטריצה משולשת תחתונה.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 3 – ריבועים פחותים

נושאים שייסקרו:

1. מערכות של משוואות וריבועים פחותים
2. מינימיזציה של בעיות ריבועיות
3. גזירת ביטויים מטריציים-וקטוריים
4. התאמת עקומות לנתונים
5. ריבועים פחותים ממושקלים
6. רגולריזציה לבעיות LS

1. מערכות של משוואות וריבועים פחותים (LS)

נניח כי בידינו סידרה בת m משוואות ב-n נעלמים שאת פתרונן אנו מחפשים –

$$f_1(\underline{x})=0,\ \ f_2(\underline{x})=0,\ \ \dots\ \ f_m(\underline{x})=0\ .$$

נוכל להגדיר פונקציה שמשלבת את כל אלה ע"י

$$p(\underline{x})=[f_1(\underline{x})]^2+[f_2(\underline{x})]^2+\cdots+[f_m(\underline{x})]^2=\|\underline{f}(\underline{x})\|_2^2$$

ולחפש לה את הווקטור $\underline{x}*$ שייתן מינימום. ברור כי אם יש פתרון למערכת המקורית, ערך הפונקציה $p(\underline{x}*)$ במינימום יהיה זהותית אפס. לכן, מינימיזציה של p כמוה כמו פתרון מערכת המשוואות המקורית.

במקרה המיוחד של מערכת משוואות ליניארית, $\mathbf{A}\underline{x}=\underline{b}$, התהליך הנ"ל שקול להגדרת פונקצית המחיר

$$p(\underline{x})=(\mathbf{A}\underline{x}-\underline{b})^T(\mathbf{A}\underline{x}-\underline{b})=\|\mathbf{A}\underline{x}-\underline{b}\|_2^2\ .$$

ושוב, החלפנו משימת פתרון מערכת משוואות ליניארית במשימת מינימיזציה. אגב, הסימון $\|\underline{a}\|_2^2$ פירושו נורמת ℓ^2 (לכן ה-2 למטה) בריבוע (ולכן ה-2 למעלה), כלומר,

$$\|\underline{a}\|_2^2=\underline{a}^T\underline{a}=\sum_{j=1}^{n}a_j^2\ .$$

האם שינוי זה בעל חשיבות? מסתבר שהתשובה היא חיובית. נניח כי למערכת המשוואות אין פתרון. מקרה כזה יכול למשל לקרות כאשר לפנינו מערכת של 5 משוואות ב-4 נעלמים עם פתרון קיים. סטייה קלה בערכי המטריצה **A** יגרמו לאובדן של פתרון, ובמקרה כזה גישת טיפול מבוססת מערכות של משוואות פשוט תאמר "אין

פתרון". גישת המינימיזציה לא מרימה ידיים, ולפתרון המביא את p למינימום יש ערך כאותו פתרון שייתן בקירוב טוב את פתרון המערכת המקורית. אם נסמן את השגיאה

$$\underline{r} = \mathbf{A}\underline{x} - \underline{b},$$

הרי שבמינימיזציה של p נרצה למצוא את $\underline{x}*$ שיביא למינימום אנרגיה של שגיאה זו, כיוון ש

$$p(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T(\mathbf{A}\underline{x} - \underline{b}) = \|\underline{r}\|_2^2.$$

כדוגמה קלאסית לבעיה מהסוג שתואר קודם, נציג את בעיית מציאת הנקודה הקרובה ביותר מתת-מרחב. נתונה לנו נקודה $\underline{b}$ במרחב n מימדי. כמו כן, נתון לנו סט $\{\underline{v}_k\}$ הכולל $K_0 < n$ וקטורים, גם הם במימד n. סט הוקטורים פורסים תת מרחב על-ידי כל הקומבינציות הליניאריות האפשריות. כל קומבינציה כזו ניתן לכתוב כמכפלה

$$\begin{bmatrix} | & | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \underline{v}_3 & \cdots & \underline{v}_{K_0} \\ | & | & | & & | \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{K_0} \end{bmatrix} = \underline{v} = \mathbf{A}\underline{x}$$

אנו מעוניינים למצוא את הנקודה $\underline{v}*$ שנמצאת בתת-המרחב הנ"ל (ולכן ניתנת לרישום כ-$\mathbf{A}\underline{x}$) ואשר תהיה הכי קרובה ל-$\underline{b}$. פתרון בעיה זו אינו אלא מינימיזציה של הפונקציה

$$p(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T(\mathbf{A}\underline{x} - \underline{b}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2,$$

כשאנו בפועל מחפשים את מקדמי הצירוף x אשר יבנו את הווקטור ע"י $\mathbf{A}\underline{x}^* = \underline{v}*$.

2. מינימיזציה של בעיות ריבועיות

אז איך פותרים בפועל את בעיית האופטימיזציה הנדונה? נשים לב כי הפונקציה שבידינו ניתנת לרישום אחר ע"י פתיחת הסוגריים,

$$\begin{aligned} p(\underline{x}) &= \|\mathbf{A}\underline{x} - \underline{b}\|_2^2 = (\mathbf{A}\underline{x} - \underline{b})^T(\mathbf{A}\underline{x} - \underline{b}) = \\ &= \underline{x}^T\mathbf{A}^T\mathbf{A}\underline{x} - \underline{b}^T\mathbf{A}\underline{x} - \underline{x}^T\mathbf{A}^T\underline{b} + \underline{b}^T\underline{b} = \underline{x}^T\mathbf{A}^T\mathbf{A}\underline{x} - 2\underline{x}^T\mathbf{A}^T\underline{b} + \underline{b}^T\underline{b} \end{aligned}.$$

נסמן $\mathbf{K} = \mathbf{A}^T\mathbf{A}$, $\underline{f} = \mathbf{A}^T\underline{b}$, ו- $c = \underline{b}^T\underline{b}$. נשים לב כי הגדרת $\mathbf{K}$ מבטיחה את היותה סימטרית ואף חיובית חצי מוגדרת. הפונקציה שבידינו הינה

$$p(\underline{x}) = \underline{x}^T\mathbf{K}\underline{x} - 2\underline{x}^T\underline{f} + c = \sum_{i=1}^{n}\sum_{j=1}^{n} k_{ij}x_i x_j - 2\sum_{i=1}^{n} b_i x_i + c.$$

זוהי פונקציה של n נעלמים המקבלת ערך סקלר. יתרה מזו – ניכר כי זוהי פונקציה ריבועית במשתניה, המכלילה את המשוואה הריבועית המוכרת לנו מחד-מימד. כעת נראה מהו פתרונה של בעיה זו לסט $\{\mathbf{K},\underline{f},c\}$ כללי:

משפט 1: בהינתן הבעיה הריבועית

$$p(\underline{x}) = \underline{x}^T\mathbf{K}\underline{x} - 2\underline{x}^T\underline{f} + c$$

המאופיינת על-ידי $\{\mathbf{K},\underline{f},c\}$, אם **K** מטריצה חיובית מוגדרת, אזי לפונקציה p יש נקודת מינימום יחידה וגלובלית הנתונה ע"י

$$\underline{x}^* = \mathbf{K}^{-1}\underline{f},$$

וערך הפונקציה בנקודת מינימום גלובלית זו נתונה ע"י

$$p(\underline{x}^*) = c - (\underline{x}^*)^T\mathbf{K}(\underline{x}^*).$$

לפני שנוכיח משפט זה, חשוב להבהיר כי בנוסחת הפתרון מותר ההיפוך של **K** כיוון שבהיותה מטריצה חיובית מוגדרת היא אינה סינגולארית. על מנת להוכיח כי אמנם זה הפתרון, נציע רישום מחדש של בעיית האופטימיזציה הריבועית שבידינו:

$$\begin{aligned} p(\underline{x}) &= \underline{x}^T\mathbf{K}\underline{x} - 2\underline{x}^T\underline{f} + c = \\ &= \underline{x}^T\mathbf{K}\underline{x} - 2\underline{x}^T\mathbf{K}\underline{x}^* + c = \\ &= \underline{x}^T\mathbf{K}\underline{x} - \underline{x}^T\mathbf{K}\underline{x}^* - (\underline{x}^*)^T\mathbf{K}^T\underline{x} + c = \\ &= (\underline{x} - \underline{x}^*)^T\mathbf{K}(\underline{x} - \underline{x}^*) + \left[c - (\underline{x}^*)^T\mathbf{K}(\underline{x}^*)\right] \end{aligned}$$

נשים לב כי בפיתוח הנ"ל ניצלנו את הגדרת הפתרון ככזה המקיים $\mathbf{K}\underline{x}^* = \underline{f}$, וניצלנו את היות המטריצה **K** סימטרית.

הביטוי שקיבלנו כולל תבנית ריבועית מהצורה $\underline{y}^T\mathbf{K}\underline{y}$ (נשתמש בהגדרה $\underline{y} = \underline{x} - \underline{x}^*$) אשר לגביה ידוע לנו כי, בשל היות המטריצה **K** חיובית מוגדרת, ערך ביטוי זה חיובי תמיד ומתאפס רק עבור $\underline{y} = \underline{x} - \underline{x}^* = 0$. המרכיב השני הינו קבוע ולא משפיע. לכן, ברור כי המינימום של הפונקציה יושג עבור $\underline{x} = \underline{x}^*$ כפי שנטען, ופתרון זה יחיד ומוביל למינימום גלובלי (כלומר אין נקודה בה יש גובה נמוך יותר לפונקציה). כמו כן, בהצבת פתרון זה יתאפס הגורם הראשון וייוותר השני, כפי שהמשפט טוען.

פועל יוצא של המשפט הנ"ל הוא שבהבאת הביטוי המקורי

$$p(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T(\mathbf{A}\underline{x} - \underline{b}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2$$

למינימום, הפתרון יהיה

$$.\left(\mathbf{A}^T\mathbf{A}\right)\underline{x}^* = \mathbf{A}^T\underline{b} \quad \Rightarrow \quad \underline{x}^* = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\underline{b}$$

אבל פתרון זה תקף רק אם המטריצה **A** בעלת עמודות בלתי תלויות ליניארית, כיוון שאז המטריצה $\mathbf{K} = \mathbf{A}^T\mathbf{A}$ חיובית מוגדרת כפי שהמשפט דורש. המשוואה

$$\left(\mathbf{A}^T\mathbf{A}\right)\underline{x}^* = \mathbf{A}^T\underline{b}$$

מוכרת בשם "המשוואה הנורמאלית".

נמחיש תוצאה זו ע"י דוגמה – נתייחס לפונקציה הדו-ממדית הבאה:

$$.\, p(x_1, x_2) = 4x_1^2 - 2x_1x_2 + 3x_2^2 + 3x_1 - 2x_2 + 1$$

פונקציה זו ניתנת לרישום גם כ:

$$.\, p(\underline{x}) = \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 4 & -1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 2\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} -1.5 \\ 1 \end{bmatrix} + 1$$

נקודת המינימום של פונקציה זו תתקבל ב-

$$,\mathbf{K}\underline{x}^* = \begin{bmatrix} 4 & -1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} -1.5 \\ 1 \end{bmatrix}$$

אך פתרון זה יהיה נכון רק אם **K** חיובית מוגדרת. לכן נפעיל פירוק LDV על מטריצה זו ונבחן את היותה של המטריצה "רגילה" ובעלת ערכים חיוביים ב-**D**. פעולת שורה ראשונה תנטרל את האיבר a_{21},

$$.\begin{bmatrix} 1 & 0 \\ 0.25 & 1 \end{bmatrix}\begin{bmatrix} 4 & -1 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 4 & -1 \\ 0 & 2.75 \end{bmatrix}$$

כלומר **K** ניתנת לרישום כמכפלה של מטריצות משולשת תחתונה בסיסית **L** ומשולשת עליונה **U**,

$$.\mathbf{K} = \begin{bmatrix} 4 & -1 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.25 & 1 \end{bmatrix}^{-1}\begin{bmatrix} 4 & -1 \\ 0 & 2.75 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -0.25 & 1 \end{bmatrix}\begin{bmatrix} 4 & -1 \\ 0 & 2.75 \end{bmatrix}$$

נרמול השורות של המטריצה **U** תיתן

$$.\mathbf{U} = \begin{bmatrix} 4 & -1 \\ 0 & 2.75 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 2.75 \end{bmatrix}\begin{bmatrix} 1 & -0.25 \\ 0 & 1 \end{bmatrix}$$

לכן, השגנו פירוק LDV של המטריצה **K** כ -

$$,\begin{bmatrix} 4 & -1 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -0.25 & 1 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 2.75 \end{bmatrix}\begin{bmatrix} 1 & -0.25 \\ 0 & 1 \end{bmatrix} = \mathbf{LDL}^T$$

ובפירוק זה לא נדרשנו להחלפת סדר שורות, והאיברים ב-D (איברי ה-Pivot) כולם חיוביים, עובדה שמעידה על חיוביות מוגדרת. לכן הפתרון לבעיית המינימיזציה יהיה

$$.\,\underline{f} = \begin{bmatrix} -1.5 \\ 1 \end{bmatrix} = \mathbf{K}\underline{x}^* = \mathbf{LU}\underline{x}^* = \begin{bmatrix} 1 & 0 \\ -0.25 & 1 \end{bmatrix}\begin{bmatrix} 4 & -1 \\ 0 & 2.75 \end{bmatrix}\underline{x}^*$$

בשלב ראשון נפתור את המשוואה הבאה בהחלפה קדמית

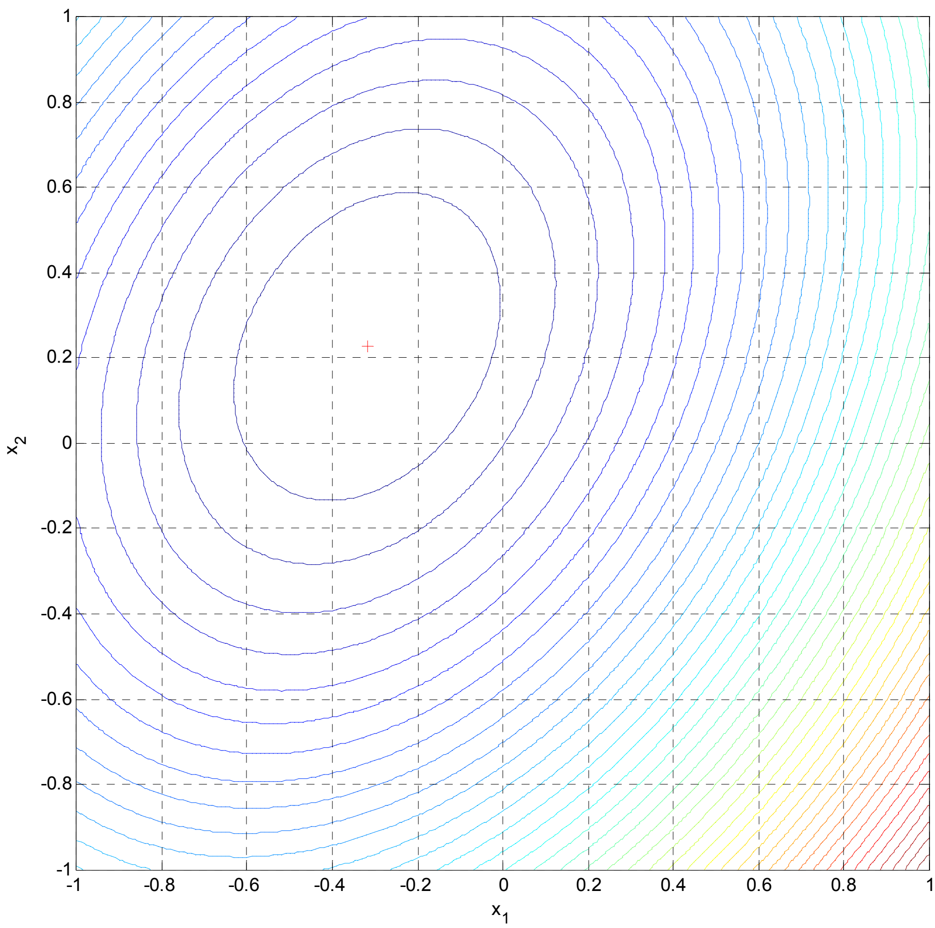
$$\begin{bmatrix} -1.5 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -0.25 & 1 \end{bmatrix} \underline{y}^* \quad \Rightarrow \quad \underline{y}^* = \begin{bmatrix} -1.5 \\ 5/8 \end{bmatrix}.$$

ולאחריה בהחלפה אחורית נקבל

$$\begin{bmatrix} -1.5 \\ 5/8 \end{bmatrix} = \begin{bmatrix} 4 & -1 \\ 0 & 2.75 \end{bmatrix} \underline{x}^* \quad \Rightarrow \quad \underline{x}^* = \begin{bmatrix} -7/22 \\ 5/22 \end{bmatrix},$$

וזהו הפתרון לבעיה שלנו. להלן תוכנית Matlab הממחישה פונקציה זו, ונקודת המינימום שלה.

```
[x1,x2]=meshgrid(-1:0.01:1,-1:0.01:1);
p=4*x1.^2-2*x1.*x2+3*x2.^2+3*x1-2*x2+1;
contour(x1,x2,p,40);
axis image;
xlabel('x_1'); ylabel('x_2');
grid on; hold on;
Xsol=[-7/22; 5/22]; plot(Xsol(1),Xsol(2),'+r');
```



מעניין לציין שיכולנו לגשת לכל עניין הבאת הפונקציה

$$p(x_1,x_2)=4x_1^2-2x_1x_2+3x_2^2+3x_1-2x_2+1$$

למינימום בגישה קונוונציונלית, לפי חישוב נגזרותיה ואיפוסם. גישה זו תביא למערכת המשוואות

$$\left\{\begin{aligned}\frac{\partial p(x_1,x_2)}{\partial x_1}&=8x_1-2x_2+3=0\\ \frac{\partial p(x_1,x_2)}{\partial x_2}&=6x_2-2x_1-2=0\end{aligned}\right\}\quad\Rightarrow\quad\begin{bmatrix}8&-2\\-2&6\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}-3\\2\end{bmatrix}$$

ואנו רואים כי הגענו לאותה מערכת משוואות בדיוק! מעניין לציין כי בגישה זו, על מנת שהפתרון המאפס את הנגזרות יהיה אמנם פתרון בעיית המינימום עלינו לדרוש שהנגזרת השנייה תקיים תכונה מסוימת – היות מטריצת ההסיין (Hessian) חיובית מוגדרת. אנו נכיר תכונה זו בהמשך כשנדון בבעיות אופטימיזציה באופן כללי. לעת עתה נאמר כי מטריצת הנגזרות השניות צריכה לקיים

$$\begin{bmatrix}\frac{\partial^2 p(x_1,x_2)}{\partial x_1^2}&\frac{\partial^2 p(x_1,x_2)}{\partial x_1\partial x_2}\\ \frac{\partial^2 p(x_1,x_2)}{\partial x_2\partial x_1}&\frac{\partial^2 p(x_1,x_2)}{\partial x_1^2}\end{bmatrix}=\begin{bmatrix}8&-2\\-2&6\end{bmatrix}>0,$$

ואנו רואים כי למעט מקדם (2), זוהי בדיוק הדרישה שעלתה גם בגישה שניתחנו.

מה קורה אם המטריצה **K** היא חיובית חצי מוגדרת? ברוח הניתוח שהוצע קודם, במקרה כזה כל וקטור שמקיים את המשוואה

$$\mathbf{K}\underline{x}^*=\underline{f}$$

יהווה נקודת מינימום של הפונקציה, כיוון שניתן לכותבה כ:

$$p(\underline{x})=\underline{x}^T\mathbf{K}\underline{x}-2\underline{x}^T\underline{f}+c=(\underline{x}-\underline{x}^*)^T\mathbf{K}(\underline{x}-\underline{x}^*)+\left(c-(\underline{x}^*)^T\mathbf{K}(\underline{x}^*)\right).$$

אלא שהפעם לא יהיה פתרון יחיד אלא סט אינסופי של פתרונות אפשריים. זאת כיוון שאם מצאנו וקטור ב-Null-Space של **K**, כלומר $\mathbf{K}\underline{z}=0$, אזי כל פתרון מהצורה $\underline{x}^*+\underline{z}$ יהיה גם הוא פתרון בעל אותה איכות (במובן של הבאת הפונקציה לגובהה המינימלי). נוכל לומר כי אוסף כל הפתרונות האפשריים מהווה קבוצה קמורה כיוון שאם $\underline{x}_A,\underline{x}_B$ הם שני פתרונות אפשריים, אזי כך גם כל פתרון המצוי על הקטע הישר שביניהם,

$$\underline{x}^*=\alpha\underline{x}_A+(1-\alpha)\underline{x}_B\qquad\forall\alpha\in[0,1].$$

אנו נחזור לתכונות אלו בהמשך.

כדוגמה, אם נחזור לרגע לבעיית מציאת הנקודה הקרובה ביותר מתוך תת-מרחב לנקודה נתונה (בעיה זו הוצגה למעלה), אם הסט הכולל K_0 וקטורים $\{\underline{v}_k\}$ הינו סט בלתי תלוי ליניארית, נקבל כי המטריצה **K** חיובית מוגדרת ואז יש פתרון יחיד לבעיה. אם לעומת זאת סט הווקטורים תלוי ליניארית, יהיו אינסוף פתרונות אפשריים.

כפי שכבר הזכרנו קודם, יתרון עצום של שיטת ה-LS בהשוואה לטיפול במערכות של משוואות היא היכולת להתמודד עם מערכות של משוואות נטולות פתרון. נמחיש זאת דרך דוגמה. נניח כ בידינו מערכת המשוואות הבאה:

$$\mathbf{A}\underline{x} = \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & -1 & -2 \\ 2 & 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 2 \\ 2 \end{bmatrix} = \underline{b}$$

נפעיל מספר פעולות שורה בסיסיות של אלימינציה גאוסית על מנת לבחון האם מדובר כאן במערכת משוואות המובילה לפתרון. תהליך זה נותן

$$\left[\begin{array}{ccc|c} 1 & 2 & 0 & 1 \\ 3 & -1 & 1 & 0 \\ -1 & 2 & 1 & -1 \\ 1 & -1 & -2 & 2 \\ 2 & 1 & -1 & 2 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 2 & 0 & 1 \\ 0 & -7 & 1 & -3 \\ 0 & 4 & 1 & 0 \\ 0 & -3 & -2 & 1 \\ 0 & -3 & -1 & 0 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 2 & 0 & 1 \\ 0 & -7 & 1 & -3 \\ 0 & 0 & 11/7 & -12/7 \\ 0 & 0 & -17/7 & 16/7 \\ 0 & 0 & -10/7 & 0 \end{array}\right]$$

וניכר כי המערכת מובילה למסקנה שאין פתרון – שלושת המשוואות האחרונות מובילות לערך x_3 שונה.

טיפול בבעיה זו ע"י LS יתנהל בדרך אחרת, בה נחפש את נקודת המינימום של הפונקציה

$$p(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T (\mathbf{A}\underline{x} - \underline{b}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2 .$$

כבר ראינו כי אם **A** לא סינגולארית, הפתרון יהיה

$$\underline{x}^* = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\underline{b} = \mathbf{K}^{-1}\underline{f} .$$

למעשה מתהליך האלימינציה הגאוסי ברור כי **A** אינה סינגולארית כיוון שכבר ראינו כי דרגת המטריצה שווה למספר איברי הציר, וכאלה יש שלושה (1, 7-, ו-11/7). לכן, הפתרון במקרה כזה יהיה

$$\underline{x}^* = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\underline{b} = \mathbf{K}^{-1}\underline{f} = \begin{bmatrix} 16 & -2 & -2 \\ -2 & 11 & 2 \\ -2 & 2 & 7 \end{bmatrix}^{-1} \begin{bmatrix} 8 \\ 0 \\ -7 \end{bmatrix} \cong \begin{bmatrix} 0.412 \\ 0.248 \\ -0.953 \end{bmatrix} .$$

אם נציב פתרון זה נקבל

$$\mathbf{A}\underline{x}^* = \begin{bmatrix} 0.9083 \\ 0.0342 \\ -0.8687 \\ 2.0701 \\ 2.0252 \end{bmatrix} \quad \underline{b} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 2 \\ 2 \end{bmatrix} \quad \Rightarrow \quad \underline{r} = \mathbf{A}\underline{x}^* - \underline{b} = \begin{bmatrix} -0.0917 \\ 0.0342 \\ 0.1313 \\ 0.0701 \\ 0.0252 \end{bmatrix}.$$

אנו רואים כי נמצא פתרון הנותן סטיות קטנות (הכי קטנות במדד סכום ריבועיהם) בין שני אגפי המשוואה. שאלה לגיטימית היא מדוע למדוד את השגיאה בפתרון ע"י סכום ריבועים? יכולנו למשל למדוד את המכסימלי שבשגיאות, או סכומם בערך מוחלט, ולבקש מינימיזציה של שגיאות אלו. מסתבר כי תהליכים אלו אפשריים אך מורכבים יותר לחישוב, והם בבירור יובילו לפתרון אחר. היתרון הגדול ב-LS הוא היכולת של כלי אלגברה ליניארית פשוטים לספק פתרון סגור.

3. גזירת ביטויים מטריציים-וקטוריים

בדוגמה קודמת הראנו כי הגישה מבוססת המטריצות והגישה הישירה למינימיזציה ע"י איפוס נגזרות מובילות לאותה תוצאה, כצפוי. כעת נראה כי הן אינן שונות, ומדובר בפניה השונות של אותה שיטה. שאלה ראשונה בה נתמקד היא השאלה הבאה: עבור הפונקציה הריבועית הבאה

$$p(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T(\mathbf{A}\underline{x} - \underline{b}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2,$$

מהי נגזרתה? כלומר, כעת בידינו פונקציה מרובת משתנים, ועלינו לחשב נגזרת ביחס לכל אחד מהם. אנו נניח כי נגזרות אלו מאורגנות בווקטור עמודה - הגרדיאנט,

$$\operatorname{grad}\{p(\underline{x})\} = \begin{bmatrix} \frac{\partial p(\underline{x})}{\partial x_1} \\ \frac{\partial p(\underline{x})}{\partial x_2} \\ \vdots \\ \frac{\partial p(\underline{x})}{\partial x_n} \end{bmatrix}.$$

המטרה היא קבלת ביטוי וקטורי מטריצי לחישוב קל ופשוט של הגרדיאנט. ברור, אגב, שאם הגרדיאנט חושב (בדרך כלשהי), איפוסו מגדיר סט של n משוואות ב-n נעלמים, ממש כמו שנעשה בדוגמה הקודמת.

על מנת לפתח נוסחא כללית, נתחיל עם פונקציה פשוטה יותר מהצורה

$$p(\underline{x}) = (\underline{a}\underline{x} - b)^2 = \|\underline{a}\underline{x} - b\|_2^2 = \left(\sum_{k=1}^{n} a_k x_k - b\right)^2.$$

בפונקציה זו הוחלפה המטריצה המלאה **A** בווקטור שורה יחיד $\underline{a}$, והווקטור $\underline{b}$ בסקלר b. עדיין הפונקציה תלויה ב- n נעלמים. חישוב מפורש של וקטור הגרדיאנט ייתן

$$\operatorname{grad}\{p(\underline{x})\} = \begin{bmatrix} \frac{\partial p(\underline{x})}{\partial x_1} \\ \frac{\partial p(\underline{x})}{\partial x_2} \\ \vdots \\ \frac{\partial p(\underline{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 2\left(\sum_{k=1}^{n} a_k x_k - b\right)a_1 \\ 2\left(\sum_{k=1}^{n} a_k x_k - b\right)a_2 \\ \vdots \\ 2\left(\sum_{k=1}^{n} a_k x_k - b\right)a_n \end{bmatrix} = 2\left(\sum_{k=1}^{n} a_k x_k - b\right)\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

במאמץ לא גדול, ניתן יהיה להשתכנע שביטוי זה אינו אלא

$$\operatorname{grad}\{p(\underline{x})\} = 2\underline{a}^T(\underline{a}\underline{x} - b),$$

וזו נוסחא נוחה לאין-ערוך, משום שברישומה אנו מתייחסים לאבני הבניין בפונקציה המקורית במלואם. באשר לסדר המכפלה, אין לו חשיבות כרגע כי גורם סקלר יכול לכפול וקטור משמאל ומימין עם אותה תוצאה. בהמשך, לסדר זה תינתן חשיבות רבה.

אם נעבור עתה לפונקציה מעט כללית יותר

$$p(\underline{x}) = (\underline{a}_1\underline{x} - b_1)^2 + (\underline{a}_2\underline{x} - b_2)^2,$$

מעניין לציין כי פונקציה חדשה זו שהצענו ניתנת לרישום אלטרנטיבי, כ:

$$p(\underline{x}) = (\underline{a}_1\underline{x} - b_1)^2 + (\underline{a}_1\underline{x} - b_2)^2 = \left\| \begin{bmatrix} - & \underline{a}_1 & - \\ - & \underline{a}_2 & - \end{bmatrix}\underline{x} - \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \right\|_2^2 = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2.$$

המטריצה **A** הפעם בעלת שתי שורות בלבד, וכמוה גם **b**. אם נתייחס לתיאור המקורי של הפונקציה כסכום של שני מרכיבים פשוטים מהסוג שכבר ניתחנו, הרי שברור כי בשל ליניאריות פעולת הנגזרת נקבל כי

$$\operatorname{grad}\{p(\underline{x})\} = 2\underline{a}_1^T(\underline{a}_1\underline{x} - b_1) + 2\underline{a}_2^T(\underline{a}_2\underline{x} - b_2).$$

אבל וקטור נגזרות זה ניתן לארגון מעט אחר, כ:

$$\begin{aligned} \operatorname{grad}\{p(\underline{x})\} &= 2\underline{a}_1^T(\underline{a}_1\underline{x} - b_1) + 2\underline{a}_2^T(\underline{a}_2\underline{x} - b_2) = \\ &= 2\begin{bmatrix} | & | \\ \underline{a}_1^T & \underline{a}_2^T \\ | & | \end{bmatrix}\left(\begin{bmatrix} - & \underline{a}_1 & - \\ - & \underline{a}_2 & - \end{bmatrix}\underline{x} - \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right) = 2\mathbf{A}^T(\mathbf{A}\underline{x} - \underline{b}) \end{aligned}.$$

המרכיב $(\mathbf{A}\underline{x} - \underline{b})$ הינו וקטור עמודה המכיל את השגיאות $(\underline{a}_1\underline{x} - b_1)$ ו- $(\underline{a}_2\underline{x} - b_2)$ בזה מתחת זה. המטריצה $\mathbf{A}^T$ מכילה את השורות $\underline{a}_1$ ו- $\underline{a}_2$ כעמודותיה (לאחר שחלוף).

הערה: ברישום זה וברבים אחרים שיבואו אנו מנצלים את העובדה שבהכפלת מטריצות ניתן לפרק את תוכן המטריצות לבלוקים בצורות שונות, וכל עוד הפירוק עקבי (במובן של התאמת מימדים בין גורמים מוכפלים), התוצאה זהה. לדוגמה, המכפלה

$$\mathbf{A}\,\underline{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{3} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

לרוב מטופלת כמכפלה של שורות המטריצה **A** בווקטור $\underline{x}$,

$$\mathbf{A}\underline{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{3} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} [a_{11} & a_{12} & a_{13}]\underline{x} \\ [a_{21} & a_{22} & a_{23}]\underline{x} \\ [a_{31} & a_{32} & a_{33}]\underline{x} \end{bmatrix}.$$

אבל, ניתן גם להתייחס למכפלה זו אחרת כשורת עמודות **A** הכופלות את עמודת $\underline{x}$,

$$\mathbf{A}\underline{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} x_1 + \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} x_2 + \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix} x_3.$$

במעבר שעשינו קודם בפישוט ביטוי הגרדיאנט עשינו שימוש בתובנה זו.

למעשה, מכאן לא רחוקה הדרך להבין שעבור הפונקציה

$$p(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T(\mathbf{A}\underline{x} - \underline{b}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2 = \sum_{k=1}^{m} (\underline{a}_k\,\underline{x} - b_k)^2$$

עם מטריצה כללית **A** בגודל m שורות על n עמודות, הגרדיאנט נתון ע"י

$$\mathrm{grad}\{p(\underline{x})\} = 2\mathbf{A}^T(\mathbf{A}\underline{x} - \underline{b}).$$

ניכר כי המכפלה שהתקבלה חוקית מבחינת מימדי המטריצות המעורבות, וגודל הווקטור נכון – n – כאורך וקטור הגרדיאנט (בדיקת שפיות).

לכן, אם רצוננו להביא למינימום את הפונקציה $p(\underline{x})$, נוכל להמיר זאת בדרישה

$$\mathrm{grad}\{p(\underline{x})\} = 2\mathbf{A}^T(\mathbf{A}\underline{x} - \underline{b}) = 0 \quad \Rightarrow \quad \mathbf{A}^T\mathbf{A}\underline{x} = \mathbf{A}^T\underline{b},$$

ומשוואה זו כבר פגשנו. אם המטריצה $\mathbf{A}^T\mathbf{A}$ לא סינגולארית, נוכל להופכה ולקבל את הפתרון האלגברי שכבר ראינו.

מה באשר לפונקציה הריבועית בגרסתה האחרת, $p(\underline{x}) = \underline{x}^T\mathbf{K}\underline{x} - 2\underline{x}^T\underline{f} + c$? איך יראה וקטור הגרדיאנט. נחזור על התהליך הנ"ל ונבנה את הנוסחא הכללית. עבור המקרה הדו-מימדי נקבל

$$
\begin{aligned}
p(\underline{x}) &= \underline{x}^T\mathbf{K}\underline{x} - 2\underline{x}^T\underline{f} + c = \\
&= \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 2\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + c = \\
&= k_{11}x_1^2 + k_{12}x_1x_2 + k_{21}x_1x_2 + k_{22}x_2^2 - 2f_1x_1 - 2f_2x_2 + c
\end{aligned}
.
$$

וקטור הנגזרות בחישוב ישיר נותן

$$
\begin{aligned}
\text{grad}\{p(\underline{x})\} &= \begin{bmatrix} 2k_{11}x_1 + k_{12}x_2 + k_{21}x_2 - 2f_1 \\ 2k_{22}x_2 + k_{12}x_1 + k_{21}x_1 - 2f_2 \end{bmatrix} = \\
&= 2\begin{bmatrix} k_{11} & (k_{12}+k_{21})/2 \\ (k_{12}+k_{21})/2 & k_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 2\begin{bmatrix} f_1 \\ f_2 \end{bmatrix}
\end{aligned}
.
$$

לכן, אם המטריצה K סימטרית נקבל כי ביטוי זה ניתן לרישום כ:

$$
\text{grad}\{p(\underline{x})\} = 2(\mathbf{K}\underline{x} - \underline{f}),
$$

ונוסחה זו נכונה גם למימדים גדולים יותר, וגם כאן ניכר כי אורך הווקטור המתקבל תואם את אורך וקטור הגרדיאנט. גם מכאן ניכר הקשר לתוצאות סעיף קודם, כיוון שבאיפוס הגרדיאנט נקבל את מערכת המשוואות $\mathbf{K}\underline{x} = \underline{f}$.

מסתבר כי נוסחאות אלו ודומות להן מאוד חשובות בבעיות אופטימיזציה, בחשבון וריאציות, וטיפול במשוואות דיפרנציאליות. ישנן נוסחאות רבות נוספות, אך לא נביא אותן כאן. לדוגמה, האתר הבא מביא רשימה ארוכה של נוסחאות כאלה:
http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html#deriv_linear

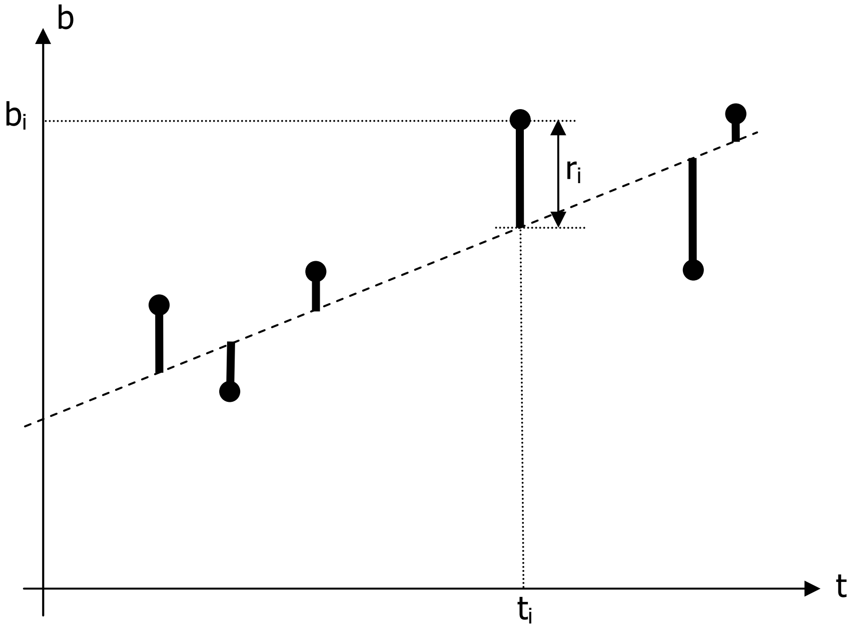
4. התאמת עקומות לנתונים

אחד השימושים החשוב ביותר ל-LS הוא התאמת עקומים. נניח כי נתון לנו אוסף צמדי נתונים $\{t_i, b_i\}_{i=1}^m$. צמדי נתונים אלו מתארים נקודות במישור, כפי שהציור הבא ממחיש. נניח כי אנו יודעים כי נקודות אלו היו אמורים ליפול על קו ישר, וכתוצאה מסטיות במדידות הם מפוזרים מעט. נרצה למצוא את הישר

$$
b = \alpha t + \beta
$$

אשר ייתן כי הוא הקרוב ביותר לתאר את כל הנקודות. לכן, נגדיר את מערכת המשוואות

$$
\underline{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix} \approx \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}\begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \mathbf{A}\underline{x}.
$$



לו נפלו כל הנקודות על הישר, היה מתקבל שוויון מדויק. בשל הסטיות, אין שוויון ואנו מחפש פרמטרים $\{\alpha,\beta\}$ מיטביים – כאלה שיתנו כי סכום ריבועי הסטיות בין שני האגפים מובא למינימום. אם נמצא את הפרמטרים הללו, הסטיות בפועל שאת סכומן הריבועי הבאנו למינימום הם הקווים המעובים בציור הקודם. זאת כיוון שלנקודה $[t_i,b_i]$ הגובה הרצוי הוא b_i בשעה שהגובה המתקבל הוא $\alpha t_i+\beta$. את ההפרש בין שני הגבהים הללו הגדרנו כשגיאת המשוואה ה-i-ית, r_i.

שימוש בשיטת הריבועים הפחותים (LS) יביא להגדרת בעיית מינימיזציה מהצורה

$$p(\underline{x})=(\mathbf{A}\underline{x}-\underline{b})^T(\mathbf{A}\underline{x}-\underline{b})=\|\mathbf{A}\underline{x}-\underline{b}\|_2^2$$

ופתרונה, כפי שכבר ראינו, נתון ע"י

$$\underline{x}^*=\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\underline{b}=\mathbf{K}^{-1}\underline{f}.$$

במקרה שלנו,

$$\mathbf{A}^T\mathbf{A}=\begin{bmatrix} m & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2\end{bmatrix} \quad \text{and} \quad \mathbf{A}^T\underline{b}=\begin{bmatrix}\sum_{i=1}^m b_i \\ \sum_{i=1}^m t_i b_i\end{bmatrix}.$$

האם המטריצה $\mathbf{A}^T\mathbf{A}=\mathbf{K}$ חיובית מוגדרת? ברור כי די בכך ששני ערכי t_i,t_j יהיו שונים זה מזה על מנת להבטיח כי ישנו פתרון. זאת כיוון שאז, אפילו אם נתעלם מכל יתר הנקודות, שתי השורות (ה-i וה-j) בלתי תלויות ליניארית ולכן הדרגה של **A** מלאה.

כל שנאמר קודם יכול לעבור הכללה למגוון רחב של פונקציות, כל עוד הנעלמים הם מקדמים ליניאריים בפונקציה. לדוגמה, אם ידוע שצמדי הנקודות צריכים ליפול על עקומה מהצורה

$$b = \alpha t + \beta + \gamma \cos(5t) + \delta \ln\left(t^2\right)$$

קל להכליל את הנאמר ולהתייחס הפעם למטריצה **A** ככזו הנבנית ע"י

$$\mathbf{A} = \begin{bmatrix} 1 & t_1 & \cos(t_1) & \ln\left(t_1^2\right) \\ 1 & t_2 & \cos(t_2) & \ln\left(t_2^2\right) \\ 1 & t_3 & \cos(t_3) & \ln\left(t_3^2\right) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_m & \cos(t_m) & \ln\left(t_m^2\right) \end{bmatrix}$$

מטריצה זו "סופגת" את כל הכיעור הקיים בפונקציה המבוקשת, ומקדמיה הליניאריים הנעלמים מטופלים יפה ע"י שיטת ה-LS.

מקרה פרטי חשוב ונפוץ לשימוש של השיטה הנ"ל מתייחס לפולינומים מסדר n. במקרה כזה נרצה שהנקודות תיפולנה על העקום המאופיין ע"י

$$b = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \cdots + \alpha_n t^n .$$

בעיית ה-LS שאותה נרצה לפתור במקרה זה תהיה

$$p(\alpha_0, \alpha_1, \ldots, \alpha_n) = \left\| \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ 1 & t_3 & t_3^2 & \cdots & t_3^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} - \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix} \right\|_2^2 .$$

מעניין לציין שהמטריצה A בבעיה שתוארה מוכרת בשם מטריצת Vandermonde והיא אופיינית בבעיות קירוב. ידוע לגבי מטריצה זו שאם מספר הערכים הסקלרים t_i השונים זה מזה הינו מעל n, מטריצה זו תהיה לא-סינגולארית, ולכן לבעיה פתרון יחיד אופטימאלי. אמירה זו לא צריכה להפתיע איש – בהינתן n+1 נקודות כלשהם שונות, ניתן להעביר פולינום מסדר n אחד ויחיד דרכן. אם יש יותר נקודות, יש לחפש את הפולינום שיתפשר עם כל הנקודות בצורה זהה להבאת סכום ריבועי הסטייה למינימום. אם יש פחות, לא קיים פתרון יחיד.

5. ריבועים פחותים ממושקלים

נניח כי אנו עוסקים בבעיית התאמת עקומות לנתונים ובידינו שלשות של מספרים:

$$\{t_i, b_i, w_i\}_{i=1}^{m} .$$

צמד המספרים הראשון הינו כמקודם – קואורדינאטה אופקית וזו האנכית התואמת לה. הערך השלישי משקף את מידת האמינות של צמד הנקודות – ערך גבוה עבור נקודה שבמדידתה יש דיוק טוב, וערך נמוך עבור נקודה עם דיוק ירוד. כיצד ניקח בחשבון "משקלות" אלו ב-LS? קודם הבאנו למינימום את השגיאה

$$p(\underline{x}) = \sum_{i=1}^{m} r_i^2(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T (\mathbf{A}\underline{x} - \underline{b}) .$$

כעת ישנן שגיאות שצריכות להשפיע יותר ואחרות שצריכות להשפיע פחות (בשל חוסר אמינותן). הרחבה טבעית של ה- LS היא הביטוי

$$p_W(\underline{x}) = \sum_{i=1}^{m} w_i \cdot r_i^2(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T \mathbf{W} (\mathbf{A}\underline{x} - \underline{b}) .$$

ברישום הנ"ל, כל שגיאה ריבועית מוכפלת במשקל מתאים – גבוה למדידה חשובה ונמוך למדידה באיכות ירודה. ברישום המטריצי השתמשנו במטריצה אלכסונית להשגת אותו אפקט בדיוק! מטריצה זו כוללת על אלכסונה את איברי המשקל w_i. פתיחת הסוגריים תגלה כי בידינו בעיה ריבועית חדשה מהצורה

$$p_W(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T \mathbf{W} (\mathbf{A}\underline{x} - \underline{b}) = \underline{x}^T \mathbf{A}^T \mathbf{W} \mathbf{A} \underline{x} - 2\underline{x}^T \mathbf{A}^T \mathbf{W} \underline{b} + \underline{b}^T \mathbf{W} \underline{b} .$$

לכן, עבור הגדרה חדשה

$$\mathbf{K} = \mathbf{A}^T \mathbf{W} \mathbf{A} \quad , \quad \underline{f} = \mathbf{A}^T \mathbf{W} \underline{b} \quad , \quad c = \underline{b}^T \mathbf{W} \underline{b}$$

זוהי בעיה ריבועית לכל דבר ובידינו לפותרה בקלות. מתכונות של מטריצות חיוביות מוגדרות אנו גם יודעים שאם **A** מטריצה בה עמודותיה בלתי תלויות ליניארית ואם **W** מטריצה בה כל איברי האלכסון חיוביים, **K** בהכרח מטריצה חיובית מוגדרת ולכן יש פתרון יחיד למינימיזציה של ה-LS במבנה זה, המוכר בשם Weighted Least Squares (WLS).

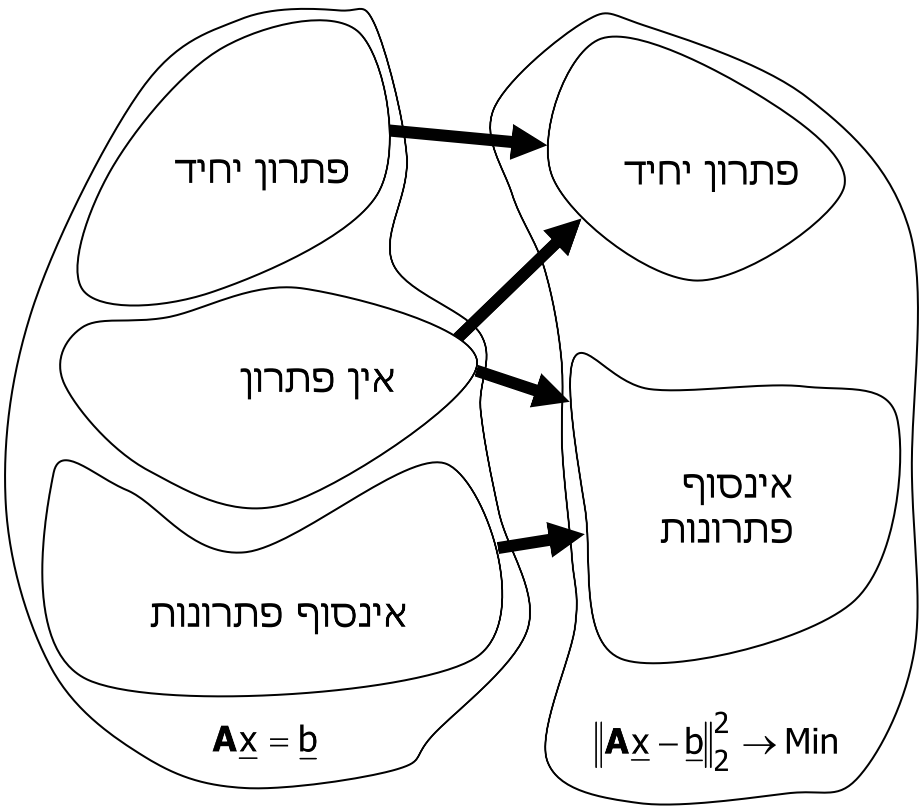
למעשה, אין כל הכרח שהמטריצה **W** תהיה אלכסונית על מנת להכליל את ה-LS. אם חוזרים על כל הנאמר קודם עם מטריצה **W** חיובית מוגדרת, יהיה פתרון יחיד למינימיזציה של בעיית ה-LS הממושקלת, ופתרון זה נתון ע"י

$$\underline{x}^* = \mathbf{K}^{-1} \underline{f} = \left(\mathbf{A}^T \mathbf{W} \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{W} \underline{b} .$$

6. רגולריזציה לבעיות LS

מדרך הצגת הדברים עד כה ברור כי אם לפנינו מערכת ליניארית $\mathbf{A}\underline{x}=\underline{b}$, אם למערכת זו קיים פתרון יחיד, שיטת ה-LS הינה אף היא בעלת פתרון יחיד – אותו הפתרון. אם לא קיים פתרון למערכת $\mathbf{A}\underline{x}=\underline{b}$, שיטת ה-LS תצליח להצמיח פתרון יחיד אם עמודות **A** בלתי-תלויות ליניארית, ולפתרון זה משמעות חשובה כפתרון המקרב באופן מיטבי את פתרון מערכת המשוואות המקורית.

אם עמודות A תלויות ליניארית, נקבל אינסוף פתרונות לבעיית ה-LS. כשלמערכת $\mathbf{A}\underline{x}=\underline{b}$ יש אינסוף פתרונות, כל אותם פתרונות יימצאו גם כפתרונות של ה-LS. הציור הבא מתאר את כמות הפתרונות בבעיית המקור $\mathbf{A}\underline{x}=\underline{b}$ וכמות הפתרונות בבעיית ה-LS החליפית.



במדעים ובהנדסה קורה לא אחת שנתקלים בבעיית LS בה יש אינסוף פתרונות (אם בשל חוסר פתרון במערכת ליניארית שתורגם לאינסוף פתרונות או בשל מצב של אינסוף פתרונות מלכתחילה), ורוצים לבודד פתרון מסוים ולבחור בו. כיצד נעשה זאת?

במהלך שנות השישים הציעו שני חוקרים רוסיים, טיכונוב וארסנין, טכניקה לייצב בעיות כאלו ולהביא לכך שפתרון יחיד ייבחר. טכניקה זו, הקרויה רגולריזציה, מציעה הוספת ביטוי ריבועי לבעיית המינימיזציה, כך שכעת היא תהיה

$$p_R(\underline{x}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2 + \lambda\|\underline{x}\|_2^2 .$$

משמעות שינוי זה היא הבאה – מכלל הפתרונות שמביאים את האיבר הראשון למינימום (או קרבתו), נרצה להתמקד באלה מהפתרונות הקצרים ביותר. אורך הפתרון נמדד בנורמת ℓ^2, ומסתבר כי תוספת זו מייצבת את הבעיה והופכת אותה לבעלת פתרון יחיד.

משפט 2: בהינתן הבעיה הריבועית $p_R(\underline{x}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2 + \lambda\|\underline{x}\|_2^2$, לפונקציה זו יש נקודת מינימום יחידה וגלובלית הנתונה ע"י $\underline{x}^* = \left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)^{-1}\mathbf{A}^T\underline{b}$.

עיון בביטוי של הפתרון מראה כי למעשה גישתם של טיכונוב וארסנין מובילה לתיקון מאוד אינטואיטיבי. כיוון ש-$\mathbf{A}^T\mathbf{A}$ חיובית חצי מוגדרת וככזו אינה הפיכה, מוסף גורם האמור לתקן הבעיה. אנו נראה בהמשך שגורם זה אמנם מרפה את הבעיה ובכך אותה ל-"רגולארית" בלשון המתמטיקאים, ומכאן השם רגולריזציה.

בדומה לנעשה בהוכחת משפט 1, נוכיח את נכונות היות פתרון זה הפתרון היחיד ואשר מוביל למינימום גלובלי ע"י הצגה אלטרנטיבית של פונקצית המחיר. נוכל לומר כי

$$\begin{aligned} p_R(\underline{x}) &= \|\mathbf{A}\underline{x} - \underline{b}\|_2^2 + \lambda\|\underline{x}\|_2^2 = \\ &= \underline{x}^T\mathbf{A}^T\mathbf{A}\underline{x} - 2\underline{x}^T\mathbf{A}^T\underline{b} + \underline{b}^T\underline{b} + \lambda\underline{x}^T\underline{x} = \\ &= \underline{x}^T\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)\underline{x} - 2\underline{x}^T\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)\underline{x}^* + \underline{b}^T\underline{b} = \\ &= (\underline{x} - \underline{x}^*)^T\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)(\underline{x} - \underline{x}^*) + \left(\underline{b}^T\underline{b} - (\underline{x}^*)^T\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)(\underline{x}^*)\right) \end{aligned}$$

ושוב קיבלנו ביטוי בו הגורם הראשון נראה כמו $\underline{y}^T\mathbf{K}\underline{y}$, והגורם השני מהווה קבוע. אם המטריצה $\mathbf{K}$ חיובית מוגדרת, הרי שהפתרון היחיד להבאת הפונקציה הנ"ל למינימום הוא $\underline{x} = \underline{x}^*$, כפי שאמנם נטען במשפט. מדוע אם כך $\mathbf{K}$ חיובית מוגדרת? כזכור, ההגדרה לחיוביות מוגדרת דורשת ש-

$$\forall\underline{y} \neq 0, \quad \underline{y}^T\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)\underline{y} > 0 .$$

פתיחת הסוגריים מגלה כי אמנם דרישה זו מתקיימת, כיוון שלכל $\underline{y}$ $\underline{y}^T\mathbf{A}^T\mathbf{A}\underline{y} \geq 0$ כיוון ש-$\mathbf{A}^T\mathbf{A}$ חיובית חצי מוגדרת. הגורם השני, לעומת זאת, הינו $\underline{y}^T\underline{y}$ והוא גדול

מאפס לכל $\underline{y} \neq 0$. למעשה מסתתר כאן כלל רחב יותר שאומר כי אם **K** מטריצה חיובית חצי מוגדרת, אזי **K+I** חיובית מוגדרת. לכן לבעיה החדשה פתרון יחיד.

למעשה, גישתם של טיכונוב וארסנין יכולה להיעשות עם ביטוי כללי יותר של רגולריזציה, מהצורה

$$p_R(\underline{x}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2 + \lambda\|\mathbf{C}\underline{x}\|_2^2 .$$

ואז ניתן להראות כי הפתרון יהיה

$$\underline{x}^* = \left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{A}^T\underline{b} ,$$

בהנחה שהמטריצה להיפוך בביטוי זה אמנם לא סינגולארית.

עבודות רבות דנות בבחירת **C** הראוי משיקולים שונים, ורבים מהם חורגים מהאלגברה ונוגעים כבר לתורת השערוך, אליה לא ניכנס כאן. עבודות רבות גם עוזבות את השימוש ב- ℓ^2 לטובת נורמות אחרות, אך בכך גם מאבדות את התמימות והפשטות המאפיינים את השיטה הנ"ל. חשוב להבהיר כי לרגולריזציה מקום חשוב גם במקרים בהם ה-LS מוביל לפתרון יחיד, אז תפקידו להכניס ייצוב מחוזק לבעיה.

יש מקום לומר מילה אחת על בחירת הפרמטר λ. עבור ערך ההולך לאינסוף, הפתרון נוטה להתאפס. בהמשך נגלה כי עבור $\lambda \to 0$ לפתרון יש מבנה מיוחד הקשור להיפוך מדומה של מטריצות. אנו נפגוש נושא זה בהמשך.

אם נחזור לרגולריזציה פשוטה, עולה נקודה מעניינת: התחלנו את סיפורנו בבעיית LS בה יש ריבוי פתרונות והכנסנו רגולריזציה שתביא למיקוד הבעיה לקבלת פתרון יחיד. האם הפתרון החדש הוא אחד מאינסוף הפתרונות שהיו בידינו קודם? מסתבר כי באופן כללי התשובה היא שלילית! הפתרונות המקוריים חייבים לקיים את הקשר

$$\mathbf{A}^T\mathbf{A}\underline{x}^*_{Orig} = \mathbf{A}^T\underline{b} ,$$

ואילו הפתרון החדש (והיחיד!) מקיים

$$\left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)\underline{x}^* = \mathbf{A}^T\underline{b} .$$

מניפולציה פשוטה על צמד קשרים אלו נותנת כי אם נרצה שהפתרון החדש יקיים את התנאי המקורי של ריבוי הפתרונות, יש לדרוש

$$\lambda\underline{x}^* = 0 .$$

וברור כי זה מחזיר אותנו לאחור למצב בו אין פתרון יחיד. למעשה, יכולנו להגדיר בעיה תחליפית בה נבטיח כי הפתרון היחיד שיצמח יבוא מתוך הפתרונות הרבים האפשריים. נגדיר את הבעיה הבאה:

$$\min_{\underline{x}} \|\underline{x}\|_2^2 \quad \text{subject to} \quad \mathbf{A}\underline{x} = \underline{b}.$$

למעשה, פתרונה של בעיה זו ניתן להמרה לבעיה נטולת אילוצים ע"י כופלי לגרנז' ולדרוש

$$\min_{\underline{x}} \|\underline{x}\|_2^2 + \lambda\|\mathbf{A}\underline{x} - \underline{b}\|_2^2.$$

ידוע לנו כי קיים ערך של λ אשר ייתן כי פתרון בעיה זו שקול לבעיה עם האילוץ – הדמיון לבעיה המקורית עם רגולריזציה ניכר. אנו נעצור את הדיון הזה כאן ונשוב אליו כשנתחמש בידע בכופלי לגרנז' ובשימוש בהם.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 4 – אורתוגונאליות

נושאים שייסקרו:

1. בסיסים אורתונורמליים וחשיבותם
2. מטריצות אורתונורמליות
3. תהליך Gram-Schmidt ווריאציות עליו
4. פירוק QR
5. שימוש בפירוק QR לפתרון מערכות משוואות
6. שימוש בפירוק QR לפתרון LS

1. בסיסים אורתונורמליים וחשיבותם

אנו יודעים כי בהינתן סט של וקטורים

$$\{\underline{v}_k\} \in \Re^n,\ k = 1,2,\dots,K_0,$$

ניתן להשתמש בהם להגדרת תת מרחב ל-$\Re^n$, ע"י איסוף כל הקומבינציות הליניאריות האפשריות. לפעולה זו אנו קוראים ה-span של סט הוקטורים. אם סט זה של וקטורים תלוי ליניארית, פירוש הדבר שנוכל להשמיט מאוסף זה וקטורים ועדיין לפרוס את אותו תת-מרחב. למעשה, בתהליך האלימינציה הגאוסית הצלחנו לגלות קיום תלויות כאלה (כשורות במערכות ליניאריות), כך שבידינו כלי לניפוי קבוצת הוקטורים למציאת הקבוצה המינימאלית שתפרוס את אותו תת-מרחב, וממילא למצוא את מימדו של תת-מרחב זה.

נמחיש זאת דרך דוגמה – נניח כי שורות המטריצה הבאה הן 5 וקטורים ממימד 3 הנתונים לנו:

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

תהליך האלימינציה ישתמש ב- a_{11} כאיבר ציר וייתן

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

שימוש ב- a_{22} החדש כציר ייתן

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

רואים כי הוצאת הווקטורים השלישי והחמישי מהאוסף לא תפגע ב-span של סט הוקטורים, וכך מיקדנו את תשומת הלב בשלשת הווקטורים הנותרים,

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

נשים לב כי תוצאת תהליך האלימינציה היא רק לסמן לנו את זהותן של הווקטורים (שורות) המיותרים. בשעה שאלו נמצאו, יש לחזור לווקטורים המקוריים לאפיון תת-המרחב (שבמקרה זה הוא המרחב המלא)!

נניח לפיכך כי בידינו קבוצת וקטורים בלתי תלויים ליניארית, $\{\underline{v}_k\}$ בה יש K_0 וקטורים ממימד $n \geq K_0$ כל אחד. אנו מכנים סט זה בסיס בהיותו סט מינימאלי הפורס את תת-המרחב שמימדו K_0. לכן, כל אלמנט בתת המרחב $\text{span}\{\underline{v}_k\}$ ניתן לרישום כ:

$$\underline{b} = \sum_{k=1}^{K_0} x_k \underline{v}_k \quad \underline{b} \in \text{span}\{\underline{v}_k\}.$$

יתרה מכך, ידוע לנו כי סט המקדמים הבונים את הווקטור $\underline{b}$ הוא יחיד. זאת כיוון שמציאתם כרוכה בפתרון מערכת משוואות מהצורה

$$\mathbf{A}\underline{x} = \begin{bmatrix} | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \cdots & \underline{v}_{K_0} \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K_0} \end{bmatrix} = \underline{b},$$

ולמערכת זו של משוואות ליניאריות פתרון יחיד! זאת כיוון שבמערכת זו יש n משוואות ב-K_0 נעלמים, ודרגת המטריצה **A** הינה בהגדרה K_0 (כיוון שעמודותיה בלתי תלויות ליניארית). לכן, קיימות K_0 משוואות בסט זה שהינן בלתי תלויות (דרגת שורות שווה לדרגת עמודות), והיתר תלויות בהן. לכן בהכרח יש פתרון יחיד או אף פתרון, וכיוון ש-$\underline{b}$ מצוי ב-span של וקטורים אלה, יהיה זה המקרה של פתרון יחיד.

כל האמור לעיל ידוע לנו כבר. למעשה, נקודת המוצא שלנו היא מערכת המשוואות שתוארה למעלה. בהינתן וקטור $\underline{b}$ עלינו לפתור מערכת משוואות מייגעת על מנת לקבוע אם אמנם הוא ב-span של וקטורי הבסיס שלנו, ועל מנת למצוא את מקדמי

ייצוגו, $\underline{x}$. השאלה הטבעית היא – האם נוכל לארגן את הבסיס כך שתהליך זה יהיה קל? מסתבר שהתשובה חיובית, ומובילה לבסיסים אורתונורמליים.

נניח כי ידוע לנו כי הסט אורתונורמלי. פירוש הדבר הוא ש-

$$\underline{v}_j^T \underline{v}_k = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}.$$

אגב, לו היה הסט אורתוגונאלי, היינו דורשים מכפלה פנימית מאופסת בין שני וקטורים שונים וכלשהי שונה מאפס אחרת. נחזור כעת למערכת המשוואות שהוגדרה קודם,

$$\mathbf{A}\underline{x} = \begin{bmatrix} | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \cdots & \underline{v}_{K_0} \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K_0} \end{bmatrix} = \underline{b}.$$

נניח כי אנו פותרים אותה בעזרת LS, דהיינו, מביאים למינימום את השגיאה

$$\left\| \mathbf{A}\underline{x} - \underline{b} \right\|_2^2.$$

כבר ראינו כי הפתרון נתון ע"י המשוואה

$$\underline{x}^* = \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \mathbf{A}^T \underline{b}.$$

היות אוסף עמודות $\mathbf{A}$ אורתונורמלי גורמת לכך ש- $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. זאת כיוון שבמטריצת Gram זו כל איבר אינו אלא מכפלה פנימית בין צמדי וקטורים מהאוסף שלנו, וכל מכפלה כזו מתאפסת אם מדובר בוקטורים שונים, ונותנת ערך יחידה עבור וקטורים זהים,

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} \underline{v}_1^T \underline{v}_1 & \underline{v}_1^T \underline{v}_2 & \underline{v}_1^T \underline{v}_3 & \cdots & \underline{v}_1^T \underline{v}_{K_0} \\ \underline{v}_2^T \underline{v}_1 & \underline{v}_2^T \underline{v}_2 & \underline{v}_2^T \underline{v}_3 & \cdots & \underline{v}_2^T \underline{v}_{K_0} \\ \underline{v}_3^T \underline{v}_1 & \underline{v}_3^T \underline{v}_2 & \underline{v}_3^T \underline{v}_3 & \cdots & \underline{v}_3^T \underline{v}_{K_0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \underline{v}_{K_0}^T \underline{v}_1 & \underline{v}_{K_0}^T \underline{v}_2 & \underline{v}_{K_0}^T \underline{v}_3 & \cdots & \underline{v}_{K_0}^T \underline{v}_{K_0} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}$$

לכן, פתרון ה-LS נתון ע"י

$$\underline{x}^* = \mathbf{A}^T \underline{b} = \begin{bmatrix} - & \underline{v}_1^T & - \\ - & \underline{v}_2^T & - \\ - & \underline{v}_3^T & - \\ & \vdots & \\ - & \underline{v}_{K_0}^T & - \end{bmatrix} \underline{b}.$$

פירוש הביטוי הנ"ל הוא שלחישוב מקדמי הייצוג של הווקטור $\underline{b}$ מעל הבסיס הנתון, כל שעלינו לעשות הוא חישוב מכפלה פנימית מהצורה

$\forall 1 \le k \le K_0 \quad x_k = \underline{v}_k^T \underline{b}$.

לכן, אם נחזור להתחלה, בהינתן וקטור $\underline{b}$ המצוי ב-span של תת-המרחב הנפרס ע"י סט K_0 הווקטורים האורתונורמליים $\{\underline{v}_k\}$, ניתן לרושמו כ-

$$\underline{b} \in \text{span}\{\underline{v}_k\} \quad \underline{b} = \sum_{k=1}^{K_0} x_k \underline{v}_k = \sum_{k=1}^{K_0} \left(\underline{v}_k^T \underline{b}\right) \underline{v}_k .$$

מעניין לציין כי התיאור הנ"ל מקל עלינו מאוד גם להוכיח כי סט וקטורים אורתונורמליים הינו בהכרח בלתי-תלוי ליניארית. אם סט זה תלוי ליניארית, פירוש הדבר שקיים פתרון לא טריוויאלי (מאופס) למערכת ההומוגנית

$$\mathbf{A}\underline{x} = \begin{bmatrix} | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \cdots & \underline{v}_{K_0} \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K_0} \end{bmatrix} = 0 .$$

שימוש בגישת ה-LS שהוצגה קודם יביא לכך שהפתרון למערכת זו הינו בהכרח

$$\underline{x}^* = \mathbf{A}^T \underline{0} = \begin{bmatrix} - & \underline{v}_1^T & - \\ - & \underline{v}_2^T & - \\ - & \underline{v}_3^T & - \\ & \vdots & \\ - & \underline{v}_{K_0}^T & - \end{bmatrix} \underline{0} = \underline{0} ,$$

כך שלא קיים כל צירוף שאינו טריוויאלי לאיפוס הקומבינציה הליניארית של איברי הבסיס.

למעשה, גישת ה-LS נותנת לנו הרבה יותר מאשר מקדמי הייצוג של $\underline{b}$ במונחי הבסיס האורתונורמלי. אם אמנם $\underline{b}$ מצוי ב-span של הבסיס הנתון, הרי שמערכת המשוואות

$$\mathbf{A}\underline{x} = \begin{bmatrix} | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \cdots & \underline{v}_{K_0} \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K_0} \end{bmatrix} = \underline{b}$$

תתקיים, עובדה שתוביל לערך אפס בפונקציה $\|\mathbf{A}\underline{x} - \underline{b}\|_2^2$. לכן, בחינת ערך הפונקציה תאמר לנו אם אמנם $\underline{b}$ שייך לתת-המרחב. ראינו בעבר כי ערך הפונקציה בפתרון האופטימאלי נתון ע"י

$$\begin{aligned} \|\mathbf{A}\underline{x}^* - \underline{b}\|_2^2 &= \underline{b}^T \underline{b} - (\underline{x}^*)^T \left(\mathbf{A}^T \mathbf{A}\right)(\underline{x}^*) = \\ &= \underline{b}^T \underline{b} - \left(\mathbf{A}^T \underline{b}\right)^T \left(\mathbf{A}^T \mathbf{A}\right)\left(\mathbf{A}^T \underline{b}\right) = \underline{b}^T \underline{b} - \underline{b}^T \mathbf{A}\mathbf{A}^T \underline{b} \end{aligned}$$

בפיתוח הנ"ל השתמשנו בעובדה ש- $\mathbf{A}^T\mathbf{A}=\mathbf{I}$ ובתוצאה $\underline{x}^*=\mathbf{A}^T\underline{b}$. האיבר השני בביטוי הנ"ל ניתן לרישום אחר כ:

$$\underline{b}^T\mathbf{A}\mathbf{A}^T\underline{b}=\sum_{k=1}^{K_0}x_k^2,$$

כש- x_k הם מקדמי הייצוג שמצאנו ע"י מכפלות פנימיות ($x_k=\underline{v}_k^T\underline{b} \quad \forall 1\le k\le K_0$).

לכן, התאפסות שגיאת ה-LS משמעה השוויון

$$\underline{b}^T\underline{b}=\|\underline{b}\|_2^2=\sum_{k=1}^{K_0}x_k^2.$$

זוהי תכונה מוכרת לנו היטב, ופירושה שהאנרגיה בווקטור המקור זהה לאנרגיה האויקלידית בווקטור מקדמי הייצוג. אגב, אם $K_0=n$, דהיינו כמות הוקטורים בבסיס הנתון זהה למימד n, הרי ששוויון זה מתקיים לכל וקטור $\underline{b}$, מכיוון שכל וקטור ב- $\Re^n$ מצוי ב- span של בסיס זה, בהיותו פורס את מלוא המרחב.

אם השגיאה ב-LS אינה אפס, פירוש הדבר שהווקטור $\underline{b}$ אינו מצוי בתת-המרחב שלנו! אם כך מה המשמעות של המקדמים x_k שנמצאו? כבר ראינו בדיון ב-LS שפתרון LS במקרה כזה ניתן לאינטרפרטציה כמציאת הנקודה בתת-המרחב הקרובה ביותר לנקודה הנתונה. יתרה מזו – מידת השגיאה $\|\mathbf{A}\underline{x}^*-\underline{b}\|_2^2$ תאמר לנו כמה רחוקה הנקודה $\underline{b}$ מתת-המרחב.

2. מטריצות אורתונורמליות

למעשה, בהגדרת מערכת המשוואות

$$\mathbf{A}\underline{x}=\begin{bmatrix} | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \cdots & \underline{v}_{K_0} \\ | & | & & | \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K_0}\end{bmatrix}=\underline{b},$$

יצרנו מטריצה בה העמודות אורתונורמליות. מטריצה **Q** קרויה מטריצה אורתונורמלית אם היא ריבועית ועמודותיה מהוות בסיס אורתונורמלי. כלומר, עלינו להתייחס למקרה בו יש בידינו בסיס למלוא המרחב $\Re^n$. מקובל גם להשתמש במושג מטריצה יוניטרית כשהכוונה למטריצות מעל המרוכבים, ואנו נשתמש במושג זה מפעם לפעם.

אם אמנם **Q** אורתונורמלית, נקבל כי $\mathbf{Q}^T\mathbf{Q}=I$. למעשה כבר ראינו תכונה זו, כיוון שאיברי המכפלה הינם מכפלות פנימיות בין עמודות המטריצה **Q**,

$$\mathbf{Q}^T\mathbf{Q} = \begin{bmatrix} \underline{q}_1^T\underline{q}_1 & \underline{q}_1^T\underline{q}_2 & \underline{q}_1^T\underline{q}_3 & \cdots & \underline{q}_1^T\underline{q}_n \\ \underline{q}_2^T\underline{q}_1 & \underline{q}_2^T\underline{q}_2 & \underline{q}_2^T\underline{q}_3 & \cdots & \underline{q}_2^T\underline{q}_n \\ \underline{q}_3^T\underline{q}_1 & \underline{q}_3^T\underline{q}_2 & \underline{q}_3^T\underline{q}_3 & \cdots & \underline{q}_3^T\underline{q}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \underline{q}_n^T\underline{q}_1 & \underline{q}_n^T\underline{q}_2 & \underline{q}_n^T\underline{q}_3 & \cdots & \underline{q}_n^T\underline{q}_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}.$$

בניגוד לקשר דומה שנכתב קודם, כאן מדובר במטריצה ריבועית במימד מלא. לכן, מכאן נובע ישירות

$$\mathbf{Q}^T\mathbf{Q} = I \;\Rightarrow\; \mathbf{Q}^{-1} = \mathbf{Q}^T \;\Rightarrow\; \mathbf{Q}\mathbf{Q}^T = \mathbf{I},$$

כלומר, היפוך של מטריצה אורתונורמלית קל ומתקבל ע"י שחלוף (Transpose). יתרה מכך, מכיוון ש-**Q** ו- $\mathbf{Q}^T$ הם צמד של מטריצה והיפוכה, ניתן לכופלם בסדר הפוך ולהגיע למטריצת היחידה שוב. זאת אומרת שמטריצה כזו שעמודותיה אורתונורמליות נותנת כי גם שורותיה כאלה!

מהקשר הנ"ל נובע גם ישירות שהדטרמיננטה של מטריצה אורתונורמלית הינו 1 בערך מוחלט, כיוון ש:

$$\begin{aligned} &\det\{\mathbf{Q}^T\mathbf{Q}\} = \det\{\mathbf{I}\} = 1 = \det\{\mathbf{Q}^T\}\det\{\mathbf{Q}\} = \det^2\{\mathbf{Q}\} \\ &\Rightarrow \det\{\mathbf{Q}\} = \pm 1. \end{aligned}$$

תכונה מעניינת נוספת היא שמטריצות אורתונורמליות הכופלות זו את זו מניבות מטריצה אורתונורמלית חדשה, כיוון שאם

$$\mathbf{Q}_1^T\mathbf{Q}_1 = \mathbf{I} = \mathbf{Q}_2^T\mathbf{Q}_2,$$

אזי

$$(\mathbf{Q}_1\mathbf{Q}_2)^T(\mathbf{Q}_1\mathbf{Q}_2) = \mathbf{Q}_2^T\mathbf{Q}_1^T\mathbf{Q}_1\mathbf{Q}_2 = \mathbf{Q}_2^T\mathbf{Q}_2 = \mathbf{I}.$$

תכונה אחרונה שנזכיר כאן בהקשר למטריצות אורתונורמליות היא שימור האורך של וקטורים. נניח כי נתון לנו וקטור $\underline{x}$ כלשהו, והכפלנו אותו במטריצה אורתונורמלית **Q**. אזי מתקיים

$$\|\mathbf{Q}\underline{x}\|_2^2 = \underline{x}^T\mathbf{Q}^T\mathbf{Q}\underline{x} = \underline{x}^T\underline{x},$$

כלומר, אורכו של הווקטור לא משתנה, כשהוא מדוד בעזרת ℓ^2. נאמר גם כי נורמת ℓ^2 הינה אינווריאנטית לפעולה אורתונורמלית. מבחינה גיאומטרית תוצאה זו צפויה כיוון שהכפלה ב-**Q** אינה אלא סיבוב במרחב של וקטור המקור, וסיבוב לא אמור לשנות אורך.

ניתן דוגמה למטריצה 2-על-2 אורתונורמלית – מהן הדרישות שעל 4 כניסות המטריצה לקיים כל מנת שהמטריצה

$$\mathbf{Q} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

תהיה אורתונורמלית? יש לדרוש עמודות מנורמלות ואנכיות זו לזו – 3 משוואות מהצורה

$$\begin{aligned} a^2 + c^2 &= 1 \\ b^2 + d^2 &= 1 \\ ab + cd &= 0 \end{aligned}$$

אין צורך לדרוש תכונות דומות מהשורות (נרמול ואנכיות) – זה אמור לקרות מאליו. נניח כי קבענו $a = \cos(\theta)$, כאשר $\theta \in [0, \pi)$ - אין בכך משום אובדן הכלליות כיוון שברור ש-a חייב להיות באינטרוול [-1,1] כפי שאמנם פונקצית ה-cos מקיימת באופן חד-חד-ערכי באינטרוול הזויות הנתון. מהמשוואה הראשונה נובע כי

$$\cos^2(\theta) + c^2 = 1 \;\Rightarrow\; c = \pm\sin(\theta).$$

כלומר, יש לנו חופש בקביעת הסימן של c. אנו נניח כי בחרנו את הסימן (+). נניח באופן דומה כי $b = \sin(\varphi)$ (זווית כלשהי אחרת בתחום $\varphi \in [-\pi/2, \pi/2)$, משיקולים דומים). באופן דומה יתקבל כי

$$\sin^2(\varphi) + d^2 = 1 \;\Rightarrow\; d = \pm\cos(\varphi).$$

שוב נניח כי בחרנו את הסימן (-). כעת עלינו לדאוג לקיומה של המשוואה השלישית,

$$ab + cd = 0 = \cos(\theta)\sin(\varphi) - \sin(\theta)\cos(\varphi) = \sin(\theta - \varphi).$$

לכן, על ההפרש בין הזויות להיות מכפלה שלמה של π, כגון 0 או π. הפרש 0 יכול להתקבל ע"י $\varphi = \theta$ ובתחום $\varphi \in [0, \pi/2)$ החופף בין השניים. הפרש π יכול להתקבל ע"י $\varphi + \pi = \theta$ ובתחום $\varphi \in [-\pi/2, 0)$. כל הפרש אחר אינו אפשרי.

במקרה הראשון המטריצה תהיה

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix}$$

המטריצה שהתקבלה במקרה של 2-על-2 מוכרת כמטריצת סיבוב בדו-מימד. בהינתן וקטור דו-מימדי כלשהו, הכפלתו במטריצה זו מסובבת אותו בשיעור θ סביב הראשית.

במקרה השני נקבל את המטריצה

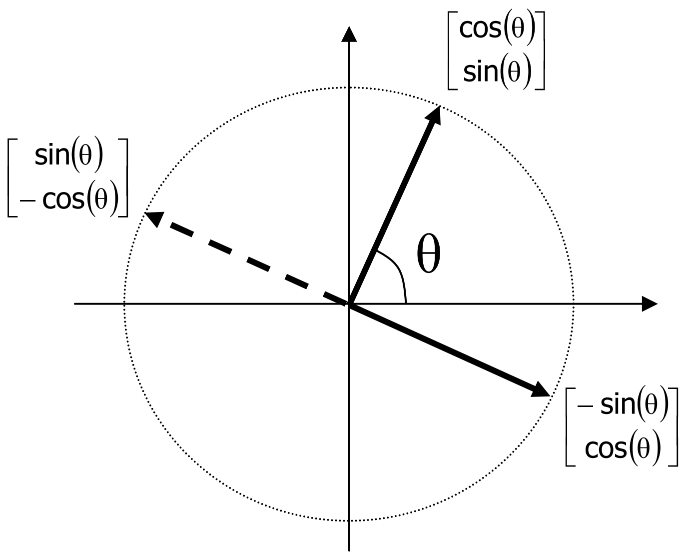
$$\begin{bmatrix} \cos(\theta) & \sin(\theta - \pi) \\ \sin(\theta) & -\cos(\theta - \pi) \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}.$$

ניתן לחזור על התהליך הנ"ל ולבחון את 4 האפשריות לסימנים. כל אלו יובילו למטריצות הבאות, ולהן בלבד:

$$\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ \sin(\theta) & -\cos(\theta)\end{bmatrix} \text{ או } \begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix} \text{ לכל } \theta\in[0,2\pi).$$

למעשה, שתי האפשרויות הן אחת, כיוון שבהינתן סט וקטורים אורתונורמלי, הכפלת אחד מהם ב-(1-) משאירה אותם קבוצה אורתונורמלית, ופשוט משקפת את הווקטור במערכת הצירים. זהו למעשה כל ההבדל בין שתי מטריצות אלה (עמודה שנייה מוכפלת ב-(1-)), ולכן ניתן לומר כי מטריצות אורתונורמליות בגודל 2-על-2 מאופיינות ע"י פרמטר יחיד באינטרוול $[0,2\pi)$. יתכן ותוצאה זו הייתה צפויה לנוכח תנאי הפתיחה – 4 פרמטרים ושלוש משוואות מאלצות. משיקול זה נובע כי מטריצה אורתונורמלית בגודל n-על-n כוללת n^2 פרמטרים חופשיים, ומולם סט של $n(n+1)/2$ אילוצי מכפלה פנימית, עובדה המותירה אותנו עם $n(n-1)/2$ דרגות חופש.

הציור הבא ממחיש את עמודות המטריצה הסיבובית שהתקבלה.



דוגמה שנייה למטריצה אורתונורמלית מעניינת היא מטריצת סיבוב Givens, הנתונה ע"י

$$\mathbf{G}=\begin{bmatrix}1 & & & & & & \\ & \ddots & & & & & \\ & & \cos(\theta) & & & -\sin(\theta) & \\ & & & \ddots & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & \sin(\theta) & & & \cos(\theta) & \\ & & & & & & \ddots \\ & & & & & & & 1\end{bmatrix}$$

זו מטריצה שנבנית ממטריצת יחידה בה שורות ועמודות i ו-j מוחלפות בתבנית 2-על-2 מסובבת מהסוג שראינו קודם. לדוגמה, המטריצה הבאה

$$\mathbf{G}=\begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & \cos(\theta) & 0 & -\sin(\theta) & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & \sin(\theta) & 0 & \cos(\theta) & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}$$

פועלת במרחב 7-מימדי, ומסובבת את הקואורדינאטות השלישית והחמישית.

דוגמה שלישית ומעניינת למטריצה אורתונורמלית ידועה בשם מטריצת השיקוף של Householder. בהינתן וקטור כלשהו $\underline{v}$ באורך יחידה ($\underline{v}^T\underline{v}=1$), מטריה זו נתונה ע"י

$$\mathbf{H}=\mathbf{I}-2\underline{v}\underline{v}^T.$$

נוכיח כי היא אמנם אורתונורמלית ע"י הכפלתה בשחלוף שלה וקבלת תוצאת יחידה. למעשה זו גם מטריצה סימטרית, ולכן

$$\begin{aligned}\mathbf{H}^T\mathbf{H}&=\left(\mathbf{I}-2\underline{v}\underline{v}^T\right)^T\left(\mathbf{I}-2\underline{v}\underline{v}^T\right)=\\ &=\mathbf{I}-2\underline{v}\underline{v}^T-2\underline{v}\underline{v}^T+4\underline{v}\underline{v}^T\underline{v}\underline{v}^T=\\ &=\mathbf{I}-4\underline{v}\underline{v}^T+4\underline{v}\underline{v}^T=\mathbf{I}\end{aligned}$$

מה קורה אם מפעילים את **H** על הווקטור $\underline{v}$? נקבל

$$\mathbf{H}\underline{v}=\left(\mathbf{I}-2\underline{v}\underline{v}^T\right)\underline{v}=\underline{v}-2\underline{v}=-\underline{v},$$

ומכאן השם שיקוף. מה קורה אם נפעיל מטריצה זו על וקטור אנך ל-$\underline{v}$? קל יהיה להשתכנע כי וקטורים כאלה יחזרו כלעומת שבאו ללא כל שינוי. לכן, בדומה לסיבוב Givens, מטריצת Householder פועלת בצירים מסוימים ואדישה באחרים. אנו עשויים לפגוש מטריצה זו בהמשך – יש לה חשיבות עצומה באלגוריתמים מבוססי מטריצות.

3. תהליך Gram-Schmidt ווריאציות עליו

ראינו כמה קל יותר לעבוד עם בסיסים אורתונורמליים. השאלה בה נעסוק כעת היא – בהינתן בסיס שאינו כזה, או אף סט וקטורים כלשהו שעבור ה-span שלו אנו מעוניינים בבסיס אורתונורמלי, כיצד נמצא כזה? תהליך Gram-Schmidt (GS) מציע לנו תהליך מסודר להשגת יעד זה.

נתון לנו סט של וקטורים $\underline{w}_1, \underline{w}_2, \underline{w}_3, \dots, \underline{w}_n$ וברצוננו לבנות בסיס אורתוגונאלי (בקלות ניתן להסב את הדיון לווקטורים אורתונורמליים, ע"י נרמול האוסף לבסוף) $\underline{v}_1, \underline{v}_2, \underline{v}_3, \dots, \underline{v}_n$ לתת-המרחב הנפרס מהם. אגב, אין כל הכרח שסט הווקטורים ההתחלתי יהיה בסיס, ואין כל הכרח שכמותם תהיה n כמימד הווקטורים – אלה גמישים לחלוטין. תהליך GS הוא תהליך סדרתי הבונה את הוקטורים אחד אחרי השני. הווקטור הראשון מתקבל בקלות על-פי

$$\underline{v}_1 = \underline{w}_1 .$$

הווקטור השני יתבסס על $\underline{w}_2$ אך ייקח בחשבון את העובדה ש- $\underline{v}_1^T \underline{v}_2 = 0$. לכן, נקבע את הווקטור השני כ-

$$\underline{v}_2 = \underline{w}_2 - c\underline{v}_1 ,$$

כלומר נחסיר מהווקטור השני הנתון לנו מרכיב מהווקטור הראשון בסט שאנו בונים. נחפש את הפרמטר c לקיום האילוץ $\underline{v}_1^T \underline{v}_2 = 0$. מתקבל

$$\underline{v}_1^T \underline{v}_2 = 0 = \underline{v}_1^T \left(\underline{w}_2 - c\underline{v}_1\right) \Rightarrow \quad c = \frac{\underline{v}_1^T \underline{w}_2}{\underline{v}_1^T \underline{v}_1} = \frac{\underline{v}_1^T \underline{w}_2}{\|\underline{v}_1\|_2^2}$$

לכן,

$$\underline{v}_2 = \underline{w}_2 - \frac{\underline{v}_1^T \underline{w}_2}{\|\underline{v}_1\|_2^2} \underline{v}_1 .$$

נשים לב לכך שתת-המרחב הנפרס ע"י הצמד $\underline{w}_1, \underline{w}_2$ ייפרס במידה זהה ע"י הצמד החדש, $\underline{v}_1, \underline{v}_2$. זהו בעצם הצידוק לכל התהליך. למה זה נכון? אנו נראה הצדקה בהמשך, כשניתן לתהליך זה פרשנות מטריצית.

תהליך GS ממשיך באופן רקורסיבי דומה, בו כל וקטור נבנה כשהוא נשען על תוצאות קודמות. נקבע כי

$$\underline{v}_3 = \underline{w}_3 - c_1 \underline{v}_1 - c_2 \underline{v}_2 ,$$

ואילוצי האנכיות של הוקטורים עד כה ייתנו

$$0 = \underline{v}_1^T \underline{v}_3 = \underline{v}_1^T \left(\underline{w}_3 - c_1 \underline{v}_1 - c_2 \underline{v}_2 \right) = \underline{v}_1^T \left(\underline{w}_3 - c_1 \underline{v}_1 \right) \;\; \Rightarrow \;\; c_1 = \frac{\underline{v}_1^T \underline{w}_3}{\left\| \underline{v}_1 \right\|_2^2}$$

$$0 = \underline{v}_2^T \underline{v}_3 = \underline{v}_2^T \left(\underline{w}_3 - c_1 \underline{v}_1 - c_2 \underline{v}_2 \right) = \underline{v}_2^T \left(\underline{w}_3 - c_2 \underline{v}_2 \right) \;\; \Rightarrow \;\; c_2 = \frac{\underline{v}_2^T \underline{w}_3}{\left\| \underline{v}_2 \right\|_2^2}$$

תופעה נוחה למדי שהתקבלה היא שחישוב c_1 מנותק מחישובו של c_2. לכן,

$$\underline{v}_3 = \underline{w}_3 - \frac{\underline{v}_1^T \underline{w}_3}{\left\| \underline{v}_1 \right\|_2^2} \underline{v}_1 - \frac{\underline{v}_2^T \underline{w}_3}{\left\| \underline{v}_2 \right\|_2^2} \underline{v}_2 .$$

קל להשתכנע כי ניתן להמשיך בתהליך זה ובשלב ה-k שלו לקבל נוסחה לפיה יהיה הווקטור ה-k-י נתון ע"י

$$\underline{v}_k = \underline{w}_k - \sum_{j=1}^{k-1} \frac{\underline{v}_j^T \underline{w}_k}{\left\| \underline{v}_j \right\|_2^2} \underline{v}_j \qquad k = 1,2,3, \ldots ,n .$$

בשלב ה-k, תת הרחב הנפרס מהקבוצה $\underline{w}_1, \underline{w}_2, \underline{w}_3, \ldots , \underline{w}_k$ יהיה זהה לתת-המרחב הנפרס ע"י הסט החדש, $\underline{v}_1, \underline{v}_2, \underline{v}_3, \ldots , \underline{v}_k$.

למעשה, תהליך GS מהווה הוכחה דרך בנייה לטענה שלכל סט של וקטורים קיים בסיס אורתונורמלי תחליפי אשר יפרוס את אותו תת-מרחב הנובע מה-span של הסט המקורי.

אם רצוננו בסט אורתונורמלי, כל שעלינו לעשות הוא לנרמל כל וקטור $\underline{v}_1, \underline{v}_2, \underline{v}_3, \ldots , \underline{v}_n$ בסט, לפי עוצמתו. אם ננרמל תוך כדי ריצה (כלומר כל וקטור ינורמל מיד עם היווצרו), נחסוך בחישובים כיוון שנוסחת העדכון לא תכלול את הנרמול ותהיה

$$\underline{v}_k = \underline{w}_k - \sum_{j=1}^{k-1} \left(\underline{v}_j^T \underline{w}_k \right) \underline{v}_j \qquad k = 1,2,3, \ldots ,n .$$

אם באחד השלבים יתקבל וקטור אפס ב-$\underline{v}_k$, פירוש הדבר שהווקטור $\underline{w}_k$ תלוי ליניארית בקודמיו, ולכן יש להורידו מהתהליך.

נראה עתה גישה מעט שונה המובילה לאותה תוצאה כבתהליך GS. אנו נקרא לשיטה זו Modified GS (MGS). כעת נניח כי רצוננו באורתונורמליזציה, כשאנו יוצאים מ-$\underline{w}_1, \underline{w}_2, \underline{w}_3, \ldots , \underline{w}_n$ ומייצרים בסיס אורתונורמלי $\underline{u}_1, \underline{u}_2, \underline{u}_3, \ldots , \underline{u}_n$. נניח כי מערכת המשוואות הבאה מתארת את תהליך הבניה,

$$
\begin{aligned}
\underline{w}_1 &= r_{11}\underline{u}_1 \\
\underline{w}_2 &= r_{21}\underline{u}_1 + r_{22}\underline{u}_2 \\
\underline{w}_3 &= r_{31}\underline{u}_1 + r_{32}\underline{u}_2 + r_{33}\underline{u}_3 \\
\vdots \quad & \quad \vdots \qquad \vdots \qquad \ddots \\
\underline{w}_n &= r_{n1}\underline{u}_1 + r_{n2}\underline{u}_2 + r_{n3}\underline{u}_3 + \cdots + r_{nn}\underline{u}_n
\end{aligned}
$$

ונראה כי אמנם ניתן לקבוע את ערכי המקדמים r_{ij} להשגת סט ראוי של וקטורים.

מערכת משוואות זו דומה מאוד לנעשה ב-GS, כיוון שהווקטור הראשון נבנה רק על סמך $\underline{w}_1$, והמקדם נועד רק לצורכי נרמול. באופן דומה, הווקטור השני נשען בבנייתו על קומבינציה ליניארית של $\underline{w}_2$ והווקטור הקודם, $\underline{v}_1$, ולכן ברור הדמיון.

מה צריכים המקדמים r_{ij} לקיים? נניח כי כבר חישבנו את k-1 הוקטורים הראשונים, $\underline{u}_1, \underline{u}_2, \underline{u}_3, \ldots, \underline{u}_{k-1}$, וכעת אנו מתמקדים בחישובו של $\underline{u}_k$. נתייחס למשוואה ה- k במערכת הנ"ל ונחשב

$$
\underline{u}_j^T \underline{w}_k = \underline{u}_j^T \left(r_{k1}\underline{u}_1 + r_{k2}\underline{u}_2 + r_{k3}\underline{u}_3 + \cdots + r_{kk}\underline{u}_k \right) = r_{kj} \qquad \forall j = 1,2,\ldots, k-1 .
$$

השתמשנו כאן בעובדה שאיברי הסדרה $\underline{u}_1, \underline{u}_2, \underline{u}_3, \ldots, \underline{u}_k$ אורתונורמליים. לכן, מבלי מאמץ רב קיבלנו את k-1 המקדמים הראשונים בצירוף הנ"ל. אם נדע גם את r_{kk} נוכל לחלץ את $\underline{u}_k$ בקלות ע"י

$$
\underline{u}_k = \frac{\underline{w}_k - \left(r_{k1}\underline{u}_1 + r_{k2}\underline{u}_2 + r_{k3}\underline{u}_3 + \cdots \right)}{r_{kk}} .
$$

באשר ל- r_{kk}, חילוצו אפשרי מהתכונה

$$
\begin{aligned}
\|\underline{w}_k\|_2^2 &= \|r_{k1}\underline{u}_1 + r_{k2}\underline{u}_2 + r_{k3}\underline{u}_3 + \cdots + r_{kk}\underline{u}_k\|_2^2 = \sum_{j=1}^{k} r_{kj}^2 \\
\Rightarrow r_{kk} &= \sqrt{\|\underline{w}_k\|_2^2 - \sum_{j=1}^{k-1} r_{kj}^2}
\end{aligned} .
$$

לכן, לסיכום, האלגוריתם המתקבל כאן נראה כך: בשלב הראשון נקבע $\underline{u}_1$ ע"י נירמולו של $\underline{w}_1$. בשלב ה-k (k=2 והלאה), בהנחה שכבר חושבו הוקטורים $\underline{u}_1, \underline{u}_2, \underline{u}_3, \ldots, \underline{u}_{k-1}$, יש לחשב את r_{kj} עבור $j = 1,2,\ldots, k-1$ לפי הנוסחא

$$
r_{kj} = \underline{u}_j^T \underline{w}_k .
$$

בהינתן אלו, יש לחשב את r_{kk} לפי הנוסחא

$$r_{kk}=\sqrt{\left\|\underline{w}_k\right\|_2^2-\sum_{j=1}^{k-1}r_{kj}^2}\,.$$

ואז לקבוע את הווקטור ה-k כ-

$$\underline{u}_k=\frac{\underline{w}_k-\left(r_{k1}\underline{u}_1+r_{k2}\underline{u}_2+r_{k3}\underline{u}_3+\cdots\right)}{r_{kk}}\,.$$

אלגוריתם זה ייתן תיאורטית את אותה התוצאה אליה הוביל תהליך GS. ההבדל ביניהם הוא שהאלגוריתם השני יעיל יותר בחישובים משום שיש בו פחות חלוקות במהלך הדרך. מסתבר, עם זאת, ששני אלגוריתמים אלו מובילים לבעיות נומריות קשות כשמדובר במימדים גדולים. סטיות שנובעות משגיאות העגלה גורמות לרוב לתוצאה שאינה סט אורתונורמלי. לכן מוצע אלגוריתם שלישי – ה-Stable GS (SGS) - המתגבר על מכשלה זו.

הצעד הראשון ייעשה כמקודם, ע"י הקביעה

$$\underline{u}_1=\frac{\underline{w}_1}{\left\|\underline{w}_1\right\|_2}\,,$$

אשר מספקת וקטור מנורמל. כעת, לפני המעבר לחישובו של הווקטור השני, "ננקה" את כל וקטורי הנתונים ע"י הפעולה

$$\underline{w}_k \leftarrow \underline{w}_k-\left(\underline{w}_k^T\underline{u}_1\right)\underline{u}_1 \quad \text{לכל} \quad k=2,3,\ \dots,\ n\,.$$

באופן זה אנו מוודאים כי בכל הוקטורים הנותרים אין כל מרכיב של הווקטור $\underline{u}_1$. כעת נמשיך ע"י הקביעה

$$\underline{u}_2=\frac{\underline{w}_2}{\left\|\underline{w}_2\right\|_2}\,,$$

כיוון שכבר הורדנו את חלקו של $\underline{u}_1$ בווקטור זה. באופן דומה, בשלב הבא "ננקה" שוב את הווקטורים הנותרים, והפעם ממרכיבי $\underline{u}_2$, ע"י

$$\underline{w}_k \leftarrow \underline{w}_k-\left(\underline{w}_k^T\underline{u}_2\right)\underline{u}_2 \quad \text{לכל} \quad k=3,4,\ \dots,\ n\,,$$

וכך הלאה. התוצאה מבחינה תיאורטי צריכה להיות זהה לזו של שני האלגוריתמים הקודמים, אך עבור כמות וקטורים גדולה, שיטה זו יציבה ומדויקת הרבה יותר.

לסיום חלק זה ניתן דוגמה שתמחיש את בעיית היציבות. נניח כי נתונים לנו הוקטורים הבאים

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

הפרמטר e הינו ערך קטן מאוד – כזה בו המחשב נותן כי $1+e=1$ בשל דיוק סופי. נתחיל בתהליך GS רגיל. נקבל

$$\underline{v}_1=\underline{w}_1=\begin{bmatrix}1 & e & 0 & 0\end{bmatrix}^T$$

$$\underline{v}_2=\underline{w}_2-\frac{\underline{v}_1^T\underline{w}_2}{\|\underline{v}_1\|_2^2}\underline{v}_1=\begin{bmatrix}1 & 0 & e & 0\end{bmatrix}^T-\frac{1}{1+e^2}\cdot\begin{bmatrix}1 & e & 0 & 0\end{bmatrix}^T$$

$$=\begin{bmatrix}0 & -e & e & 0\end{bmatrix}^T$$

$$\underline{v}_3=\underline{w}_3-\frac{\underline{v}_1^T\underline{w}_3}{\|\underline{v}_1\|_2^2}\underline{v}_1-\frac{\underline{v}_2^T\underline{w}_3}{\|\underline{v}_2\|_2^2}\underline{v}_2=$$

$$=\begin{bmatrix}1 & 0 & 0 & e\end{bmatrix}^T-\frac{1}{1+e^2}\cdot\begin{bmatrix}1 & e & 0 & 0\end{bmatrix}^T-\frac{0}{2e^2}\cdot\begin{bmatrix}0 & -e & e & 0\end{bmatrix}^T$$

$$=\begin{bmatrix}0 & -e & 0 & e\end{bmatrix}^T$$

ניכר כי התוצאה גרועה – הוקטורים השני והשלישי אינם ניצבים כלל! במקרה זה מטריצת ה- Gram על הוקטורים הנ"ל לאחר נרמול תהיה

$$\mathbf{G}=\begin{bmatrix}\underline{u}_1^T\underline{u}_1 & \underline{u}_1^T\underline{u}_2 & \underline{u}_1^T\underline{u}_3\\ \underline{u}_2^T\underline{u}_1 & \underline{u}_2^T\underline{u}_2 & \underline{u}_2^T\underline{u}_3\\ \underline{u}_3^T\underline{u}_1 & \underline{u}_3^T\underline{u}_2 & \underline{u}_3^T\underline{u}_3\end{bmatrix}=\begin{bmatrix}1 & -0.707e & -0.707e\\ -0.707e & 1 & 0.5\\ -0.707e & 0.5 & 1\end{bmatrix}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0.5\\ 0 & 0.5 & 1\end{bmatrix}.$$

אם נכניס נרמול לתוך האלגוריתם התוצאה לא תשתנה. כמו כן, אם נפעיל את גרסת ה- MGS התוצאה לא תשתנה.

נפעיל כעת את השיטה SGS. ההתחלה זהה ונותנת

$$\underline{u}_1=\underline{w}_1=\begin{bmatrix}1 & e & 0 & 0\end{bmatrix}^T$$

כמקודם, ונרמול לא נחוץ - של דיוק מוגבל וקטור זה מנורמל בעיני המחשב. עדכון שני וקטורי הנתונים ייתן

$$\underline{w}_2^{new1} = \underline{w}_2 - \frac{\underline{u}_1^T \underline{w}_2}{\|\underline{u}_1\|_2^2}\underline{u}_1 = [1 \quad 0 \quad e \quad 0]^T - \frac{1}{1+e^2}\cdot[1 \quad e \quad 0 \quad 0]^T$$
$$= [0 \quad -e \quad e \quad 0]^T$$
$$\underline{w}_3^{new1} = \underline{w}_3 - \frac{\underline{u}_1^T \underline{w}_3}{\|\underline{u}_1\|_2^2}\underline{u}_1 = [1 \quad 0 \quad 0 \quad e]^T - \frac{1}{1+e^2}\cdot[1 \quad e \quad 0 \quad 0]^T$$
$$= [0 \quad -e \quad 0 \quad e]^T$$

הווקטור השני כעת יהיה פשוט נרמול של $\underline{w}_2^{new}$, כלומר

$$\underline{u}_2 = [0 \quad -0.707 \quad 0.707 \quad 0]^T,$$

וניכר כי הוא זהה לזה אשר התקבל קודם בשיטת GS רגילה. כעת יבוא השינוי: הווקטור השלישי יחושב ע"י עדכון הכניסה לפי

$$\underline{w}_3^{new2} = \underline{w}_3^{new1} - \frac{\underline{u}_2^T \underline{w}_3^{new1}}{\|\underline{u}_2\|_2^2}\underline{u}_2 =$$
$$= [0 \quad -e \quad 0 \quad e]^T - \frac{0.707e}{1}\cdot[0 \quad -0.707 \quad 0.707 \quad 0]^T$$
$$= [0 \quad -0.5e \quad -0.5e \quad e]^T$$

כעת עלינו לנרמל ונקבל

$$\underline{u}_3 = \frac{1}{\sqrt{6}}[0 \quad -1 \quad -1 \quad 2]^T.$$

וקטור זה שונה מהתוצאה הקודמת, והוא ניצב יפה לשני קודמיו, עם סטייה בשיעור של שגיאת המחשב ולא יותר. מטריצת ה- Gram על הוקטורים הנ"ל תהיה

$$\mathbf{G} = \begin{bmatrix} \underline{u}_1^T\underline{u}_1 & \underline{u}_1^T\underline{u}_2 & \underline{u}_1^T\underline{u}_3 \\ \underline{u}_2^T\underline{u}_1 & \underline{u}_2^T\underline{u}_2 & \underline{u}_2^T\underline{u}_3 \\ \underline{u}_3^T\underline{u}_1 & \underline{u}_3^T\underline{u}_2 & \underline{u}_3^T\underline{u}_3 \end{bmatrix} = \begin{bmatrix} 1 & -0.707e & -e/\sqrt{6} \\ -0.707e & 1 & 0 \\ -e/\sqrt{6} & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

מה קרה כאן? מדוע הדברים עבדו באופן זה? כל ההבדל הוא בטיפול בווקטור השלישי. נוסחת החישוב של השיטה המקורית הינה

$$\underline{v}_3 = \underline{w}_3 - \frac{\underline{v}_1^T \underline{w}_3}{\|\underline{v}_1\|_2^2}\underline{v}_1 - \frac{\underline{v}_2^T \underline{w}_3}{\|\underline{v}_2\|_2^2}\underline{v}_2,$$

בעוד שנוסחת העדכון באלגוריתם המשופר דומה מאוד והינה

$$\underline{w}_3^{new} = \underline{w}_3 - \frac{\underline{v}_1^T \underline{w}_3}{\|\underline{v}_1\|_2^2}\underline{v}_1 \quad \text{and then} \quad \underline{v}_3 = \underline{w}_3^{new} - \frac{\underline{v}_2^T \underline{w}_3^{new}}{\|\underline{v}_2\|_2^2}\underline{v}_2.$$

הצבת המשוואה הראשונה בשנייה נותנת

$$\underline{v}_3 = \underline{w}_3^{new} - \frac{\underline{v}_2^T \underline{w}_3^{new}}{\|\underline{v}_2\|_2^2}\underline{v}_2 = \underline{w}_3 - \frac{\underline{v}_1^T \underline{w}_3}{\|\underline{v}_1\|_2^2}\underline{v}_1 - \frac{\underline{v}_2^T\left(\underline{w}_3 - \frac{\underline{v}_1^T \underline{w}_3}{\|\underline{v}_1\|_2^2}\underline{v}_1\right)}{\|\underline{v}_2\|_2^2}\underline{v}_2 =$$
$$= \underline{w}_3 - \frac{\underline{v}_1^T \underline{w}_3}{\|\underline{v}_1\|_2^2}\underline{v}_1 - \frac{\underline{v}_2^T \underline{w}_3}{\|\underline{v}_2\|_2^2}\underline{v}_2 + \frac{\underline{v}_1^T \underline{w}_3}{\|\underline{v}_1\|_2^2}\frac{\underline{v}_2^T \underline{v}_1}{\|\underline{v}_2\|_2^2}\underline{v}_2 .$$

שלושת הגורמים הראשונים זהים לנוסחא המקורית. בעולם מושלם הגורם האחרון שהצטרף היה נופל בשל היות $\underline{v}_2^T \underline{v}_1 = 0$. אם, לעומת זאת, מכפלה פנימית זו אינה מתאפסת בדיוק, התיקון הנגרם משמעותי. בדוגמה שלנו מתקבל כי גורם זה שווה

$$\frac{\underline{v}_1^T \underline{w}_3}{\|\underline{v}_1\|_2^2} \cdot \frac{\underline{v}_2^T \underline{v}_1}{\|\underline{v}_2\|_2^2} = \frac{1}{1+e^2} \cdot \frac{-e^2}{2e^2} = -0.5 .$$

4. פירוק QR

בפרק ראשון ראינו כיצד אלגוריתם האלימינציה של גאוס ניתן לפרשנות ממנה נולד פירוק ה-LU. אנו נראה כעת כי פירוק דומה ניתן להשגה מאלגוריתמי Gram-Schmidt. שלושת השיטות שתוארו קודם מובילות באלגוריתמיקה שונה לאותה תוצאה (מבחינה תיאורטית). לכן, נתייחס לשיטה השנייה בשל הקלות היחסית של המרתה למבנה וקטורי-מטריצי. ראינו כי ניתן לכתוב את מערכת המשוואות הבא לשם תיאור הקשר בין הוקטורים הנתונים לוקטורים המתקבלים:

$$\begin{aligned}
\underline{w}_1 &= r_{11}\underline{u}_1 \\
\underline{w}_2 &= r_{21}\underline{u}_1 + r_{22}\underline{u}_2 \\
\underline{w}_3 &= r_{31}\underline{u}_1 + r_{32}\underline{u}_2 + r_{33}\underline{u}_3 \\
\vdots \quad & \quad \vdots \qquad \vdots \qquad \ddots \\
\underline{w}_n &= r_{n1}\underline{u}_1 + r_{n2}\underline{u}_2 + r_{n3}\underline{u}_3 + \cdots + r_{nn}\underline{u}_n
\end{aligned} .$$

אם נגדיר מטריצה בה סט הווקטורים $\underline{u}_1, \underline{u}_2, \underline{u}_3, \ldots, \underline{u}_n$ מהווה עמודות, הרי שזו מטריצה אורתונורמלית מהגדרתה. נסמן מטריצה זו ב-**Q**:

$$\mathbf{Q} = \begin{bmatrix} | & | & | & & | \\ \underline{u}_1 & \underline{u}_2 & \underline{u}_3 & \cdots & \underline{u}_n \\ | & | & | & & | \end{bmatrix} .$$

באופן דומה, נגדיר את המטריצה **A** כמטריצה בה הוקטורים $\underline{w}_1, \underline{w}_2, \underline{w}_3, \ldots, \underline{w}_n$ הם העמודות,

$$\mathbf{A} = \begin{bmatrix} | & | & | & & | \\ \underline{w}_1 & \underline{w}_2 & \underline{w}_3 & \cdots & w_n \\ | & | & | & & | \end{bmatrix}.$$

מה הקשר בין שתי מטריצות אלה? מערכת המשוואות הנ"ל ניתנת לרישום כ:

$$\mathbf{A} = \begin{bmatrix} | & | & | & & | \\ \underline{w}_1 & \underline{w}_2 & \underline{w}_3 & \cdots & w_n \\ | & | & | & & | \end{bmatrix} =$$

$$= \begin{bmatrix} | & | & | & & | \\ \underline{u}_1 & \underline{u}_2 & \underline{u}_3 & \cdots & \underline{u}_n \\ | & | & | & & | \end{bmatrix} \begin{bmatrix} r_{11} & r_{21} & r_{31} & \cdots & r_{n1} \\ 0 & r_{22} & r_{32} & \cdots & r_{n2} \\ 0 & 0 & r_{33} & \cdots & r_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & r_{nn} \end{bmatrix} = \mathbf{QR}.$$

כאן הגדרנו את המטריצה המשולשת העליונה **R** בה מאוכלסים המקדמים במערכת המשוואות. עמודה ראשונה משמאל ב-**A** תתקבל מהמכפלה

$$\begin{bmatrix} | \\ \underline{w}_1 \\ | \end{bmatrix} = \begin{bmatrix} | & | & | & & | \\ \underline{u}_1 & \underline{u}_2 & \underline{u}_3 & \cdots & \underline{u}_n \\ | & | & | & & | \end{bmatrix} \begin{bmatrix} r_{11} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = r_{11} \begin{bmatrix} | \\ \underline{u}_1 \\ | \end{bmatrix}.$$

עמודה שנייה תתקבל ע"י

$$\begin{bmatrix} | \\ \underline{w}_2 \\ | \end{bmatrix} = \begin{bmatrix} | & | & | & & | \\ \underline{u}_1 & \underline{u}_2 & \underline{u}_3 & \cdots & \underline{u}_n \\ | & | & | & & | \end{bmatrix} \begin{bmatrix} r_{21} \\ r_{22} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = r_{21} \begin{bmatrix} | \\ \underline{u}_1 \\ | \end{bmatrix} + r_{22} \begin{bmatrix} | \\ \underline{u}_2 \\ | \end{bmatrix},$$

וכך הלאה, ואנו רואים את התאימות למערכת המשוואות המקורית. קיבלנו אם כך את המשפט הבא:

משפט 1: כל מטריצה ריבועית ולא סינגולארית **A** ניתנת לפירוק יחיד מהצורה **A**=**QR** כאשר **Q** מטריצה אורתונורמלית ו-**R** מטריצה משולשת עליונה, בה איברי האלכסון חיוביים.

עניין היות איברי האלכסון חיוביים ב-**R** נובע מדרך חישובם באלגוריתם ה-MGS, לפי הנוסחא

$$r_{kk} = \sqrt{\|\underline{w}_k\|_2^2 - \sum_{j=1}^{k-1} r_{kj}^2}\,.$$

אם נוותר על דרישה זו, ייתכנו פירוקים אחרים של מכפלת מטריצה אורתונורמלית במטריצה משולשת עליונה. באשר ליחידות הפירוק, לא נוכיח זאת כאן.

נשים לב לכך שבהינתן מטריצה שאינה מקיימת את תנאי המשפט (ריבועיות ואי-סינגולאריות), אין פירוש הדבר שלא ניתן להשיג פירוק דומה. למעשה, מתהליך GS ניכר כי ניתן לטפל במטריצות אלה ויחולו מספר הבדלים מחויבי המציאות:

- ייתכנו אפסים לאורך האלכסון הראשי של R בשל היות וקטור מסוים תלוי בקודמיו. לדוגמה, עבור המטריצה הבאה,

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

ברור שעמודתה השנייה תלויה בראשונה. הנה פירוק QR במקרה כזה (התקבל ב-Maltab ע"י פקודת qr):

$$\mathbf{Q} = \begin{bmatrix} 0.5 & -0.866 & 0 & 0 \\ 0.5 & 0.289 & -0.816 & 0 \\ 0.5 & 0.289 & 0.408 & 0.707 \\ 0.5 & 0.289 & 0.408 & -0.707 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 2 & 4 & 1 & 1 \\ 0 & 0 & 0.577 & 0.577 \\ 0 & 0 & 0.8165 & -0.408 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- אם המטריצה מלבנית עם יותר עמודות משורות, פירוש הדבר שיש לנו יותר וקטורים מהממד הנדון. אם נניח כי אוסף n העמודות הראשונות בלתי-תלוי ליניארית, הפירוק יביא למטריצה משולשת עליונה בחלקה העליון אשר קטומה בתחתיתה, כפי שהציור הבא ממחיש.

$$\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix} = \begin{bmatrix} & | & \\ - & \mathbf{Q} & - \\ & | & \end{bmatrix} \begin{bmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix}.$$

- אם המטריצה מלבנית עם יותר שורות מעמודות, פירוש הדבר שיש לנו פחות וקטורים מהממד המלא. אם נניח כי אוסף העמודות בלתי-תלוי ליניארית, הפירוק יביא למטריצה משולשת עליונה בחלקה העליון ואפסים לאחר מכן, כפי שהציור הבא ממחיש.

$$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} & | & \\ - & \mathbf{Q} & - \\ & | & \end{bmatrix} \begin{bmatrix} * & * & * \\ 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

למעשה, העמודות האחרונות במטריצה **Q** פורסות תת-מרחב שלא מצוי בעמודות **A**, ולכן יש אינסוף דרכים לייצגו בבסיס אורתונורמלי. לעיתים אף מקובל להשמיט עמודות אלה ולתת מבנה אלטרנטיבי מהצורה

$$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} & | & \\ - & \mathbf{Q} & - \\ & | & \end{bmatrix} \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix},$$

וגישה זו קרויה Economy QR decomposition.

כהערה אחרונה בהקשר של פירוק QR, יש להזכיר יש מקובל גם בפירוק QR שימוש בפרמוטציה שתכליתה שינוי סדר שורות או עמודות, כדי לארגן טוב יותר את הפירוק. למשל, ישנו עניין רב בפירוק בו איברי האלכסון של **R** מאורגנים בסדר יורד. אנו לא ניכנס לאופציות אלה כאן.

5. שימוש בפירוק QR לפתרון מערכת משוואות

אם אנו יכולים לפרק מטריצה **A** ל- QR, בידינו הכוח לפתור בעיות כגון מערכות של משוואות וריבועים פחותים. נראה זאת, ונמחיש את הייחודיות העולה מהשימוש ב-QR.

נתייחס תחילה למערכות משוואות ליניאריות, $\mathbf{A}\underline{x} = \underline{b}$, ונניח כי ביכולתנו לפרק את **A** ל-**QR**, כאשר **Q** מטריצה ריבועית ואורתונורמלית. לכן נוכל לכתוב

$$\mathbf{A}\underline{x} = \mathbf{QR}\underline{x} = \underline{b} \quad \Rightarrow \quad \mathbf{R}\underline{x} = \mathbf{Q}^T\underline{b} = \hat{\underline{b}}.$$

הכפלת $\underline{b}$ בהיפוך של המטריצה **Q** קל כיוון שהיפוכה אינו אלא ביצוע שחלוף. כעת בידינו מערכת משוואות חדשה בה **R** מטריצה משולשת עליונה. אם היא מטריצה ריבועית ולא סינגולארית, הפתרון יתקבל בקלות ע"י תהליך החלפה אחורית, בדומה לנעשה ב-LU.

אם המדובר היה במערכת עם יותר שורות מעמודות, נקבל תבנית מהצורה

$$\mathbf{R}\underline{x}=\begin{bmatrix} * & * & * \\ 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\underline{x}=\mathbf{Q}^T\underline{b}=\hat{\underline{b}},$$

ובנקל נוכל לקבוע אם יש בכלל פתרון למערכת – כל שעלינו לעשות הוא להתייחס לערכים בתחתית הווקטור $\hat{\underline{b}}$ (אלה העומדים מול שורות האפסים) ולוודא שהם מתאפסים על מנת למנוע סתירה. והיה וזה אמנם כך, נותרנו שוב עם תהליך החלפה אחורית למציאת הפתרון.

אם מדובר במטריצה בה יש יותר עמודות משורות, נקבל תבנית מהצורה

$$\mathbf{R}\underline{x}=\begin{bmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix}\underline{x}=\mathbf{Q}^T\underline{b}=\hat{\underline{b}}.$$

תבנית זו נותנת לנו מיד ידיעה על דרגות החופש שבידינו (במקרה זה שני משתנים אחרונים ייבחרו בחופשיות), והתלות בהם לפתרון המערכת.

6. שימוש בפירוק QR לפתרון LS

נניח כי עלינו להביא למינימום את הפונקציה

$$p(\underline{x})=\|\mathbf{A}\underline{x}-\underline{b}\|_2^2$$

ונניח כי בידינו בעיה בת פתרון יחיד (כלומר **A** עשויה להיות ריבועית בדרגה מלאה או מלבנית עם יותר שורות מעמודות ודרגה מלאה). נניח כמו כן שבידינו פירוק QR של **A**. נוכל לכתוב

$$\begin{aligned} p(\underline{x}) &= \|\mathbf{A}\underline{x}-\underline{b}\|_2^2 = \|\mathbf{QR}\underline{x}-\underline{b}\|_2^2 = \\ &= \left\|\mathbf{Q}\left(\mathbf{R}\underline{x}-\mathbf{Q}^T\underline{b}\right)\right\|_2^2 = \\ &= \left\|\mathbf{R}\underline{x}-\hat{\underline{b}}\right\|_2^2 \end{aligned}$$

בפיתוח הנ"ל השתמשנו בעובדה שנורמת ℓ^2 אינווריאנטית לסיבוב אורתונורמלי, ולכן ההכפלה ב-**Q** יכולה להיות מושמטת. התבנית בתוך השגיאה נראית כפי שהציור הבא מציע.

$$p(\underline{x}) = \left\| \mathbf{R}\underline{x} - \hat{\underline{b}} \right\|_2^2 = \left\| \begin{bmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \underline{x} - \begin{bmatrix} \\ \\ \hat{\underline{b}} \\ \\ \\ \end{bmatrix} \right\|_2^2 =$$

$$= \left\| \begin{bmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & * \end{bmatrix} \underline{x} - \begin{bmatrix} \\ \hat{\underline{b}}_1 \\ \\ \end{bmatrix} \right\|_2^2 + \left\| \hat{\underline{b}}_2 \right\|_2^2$$

הווקטור $\hat{\underline{b}}_1$ הינו החלק העליון של $\hat{\underline{b}}$ העומד מול שורות לא ריקות מתוכן של **R**, ואילו $\hat{\underline{b}}_2$ מתייחס לשורות המאופסות. החלק הראשון ניתן לאיפוס מדויק ע"י פתרון המשוואה

$$\mathbf{R}_1 \underline{x} = \hat{\underline{b}}_1 ,$$

אשר ייעשה בשיטת ההחלפה האחורית. לשיטה זו יתרון עצום בהשוואה לשימוש במשוואות הנורמאליות. זאת כיוון שבמשוואות הנורמאליות אנו מחשבים את $\mathbf{A}^T\mathbf{A}$ - אשר יכול לקבל ערכים בטווח דינאמי גדול, ועם זה נקבל סטיות נומריות קשות. כאן לעומת זאת, אנו עובדים ישירות עם מטריצה המיוחסת ל-**A**.

הגורם הנותר משקף את השגיאה שאין אפשרות לאפס. לכן הביטוי $\left\| \hat{\underline{b}}_2 \right\|_2^2$ אינו אלא גובה הפונקציה במינימום.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 5 - ערכים עצמיים וסינגולאריים

נושאים שייסקרו:

1. ערכים עצמיים ווקטורים עצמיים - יסודות
2. לכסון של מטריצות
3. מטריצות סימטריות וחיוביות מוגדרות
4. ערכים עצמיים כפתרון בעיות אופטימיזציה
5. ערכים עצמיים מוכללים
6. ערכים סינגולאריים ופירוק ה- SVD
7. שימוש ב- SVD למערכות של משוואות ול-LS
8. קירוב מטריצות עם אילוץ דרגה
9. בעיית ה- TLS

1. ערכים עצמיים ווקטורים עצמיים - יסודות

נתחיל בהגדרת המושגים ערך עצמי ווקטור עצמי: בהינתן מטריצה ריבועית (עם איברים ממשיים – זו תהיה ההנחה לאורך כל פרק זה, אם כי היא אינה חיונית) **A**, הסקלר λ והווקטור $\underline{v} \neq 0$ נקראים ערך עצמי ווקטור עצמי, אם הם מקיימים את המשוואה

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

אם נחשוב על המטריצה כפעולה ליניארית כללית ומגוונת, הרי שבכיוונים מסוימים (אליהם מצביעים הוקטורים העצמיים) פעולת המטריצה אינה משנה דבר פרט למתיחת הווקטור – הכפלה בסקלר λ שהינו הערך העצמי. כלומר, בכיוונים אלה פעולת המטריצה ניתנת לתיאור פשוט מאוד וקל ליישום. לכן כיוונים אלה ושיעור המתיחה עליהם כה מעניינים וחשובים.

יש לשים לב לכך שהדרישה $\underline{v} \neq 0$ חיונית! בלעדיה נקבל את הפתרון הטריוויאלי $\underline{v} = 0$ וכל סקלר שהוא כערך עצמי תואם, וברור שלפתרון כזה אין כל משמעות או תרומה להבנת המטריצה ודרך פעולתה. מעבר לכך, אם מצאנו וקטור עצמי שאינו טריוויאלי, כל הכפלה שלו בקבוע תותיר אותו כפתרון ראוי. לכן מקובל לדרוש $\|\underline{v}\|_2^2 = 1$.

בהינתן המטריצה **A** נרצה למצוא את הפתרון (או פתרונות) של המשוואה $\mathbf{A}\underline{v} = \lambda\underline{v}$. לו ידענו את λ, הייתה זו מערכת ליניארית שאת פתרונה נוכל לחשב. אבל אנו לא יודעים את λ, ולכן לפנינו בעיה לא ליניארית. למרות זאת, אנו נראה מיד כי ניתן לפתור אותה בצורה שיטתית. לצורך כך, נשים לב כי אותה משוואה ניתנת לרישום כ:

$$(\mathbf{A}-\lambda\mathbf{I})\underline{v}=0 .$$

כיוון שאנו דורשים $\underline{v}\neq 0$, ברור כי λ צריכה להיות כזו שתיתן כי המטריצה $(\mathbf{A}-\lambda\mathbf{I})$ סינגולארית. לכן, נוכל להשתמש בתכונה שדטרמיננטה של מטריצה סינגולארית היא אפס, ולחפש את λ ע"י המשוואה

$$\det\{\mathbf{A}-\lambda\mathbf{I}\}=0 .$$

משוואה זו ידועה בשם "המשוואה האופיינית" (Characteristic Polynomial). פיתוחה יוביל לפולינום מדרגה n.

לכן, נוכל להתרכז קודם במציאת λ שיתנו את קיום המשוואה הנ"ל, ולאחר מכן למצוא את הוקטורים העצמיים כפתרון של מערכות משוואות ליניאריות. כדוגמה, נתייחס למטריצה הבאה, ונמצא את ערכיה ווקטוריה העצמיים.

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

המשוואה האופיינית תהיה

$$\begin{aligned}\det\{\mathbf{A}-\lambda\mathbf{I}\}&=\det\begin{bmatrix}-\lambda & -1 & -1\\ 1 & 2-\lambda & 1\\ 1 & 1 & 2-\lambda\end{bmatrix}=\\ &=-\lambda\cdot\left[(2-\lambda)^2-1\right]+1\cdot[(2-\lambda)-1]-1\cdot[1-(2-\lambda)]=\\ &=-(2-\lambda)^2\lambda+\lambda+2(2-\lambda)-2=\\ &=-\lambda^3+4\lambda^2-5\lambda+2=-(\lambda-1)^2(\lambda-2)\end{aligned} .$$

לכן הערכים העצמיים הם 1 ו-2. עבור $\lambda_1=1$ את הווקטור העצמי נמצא לפי המשוואה

$$(\mathbf{A}-\lambda_1\mathbf{I})\underline{v}=\begin{bmatrix}-\lambda_1 & -1 & -1\\ 1 & 2-\lambda_1 & 1\\ 1 & 1 & 2-\lambda_1\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix}-1 & -1 & -1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix}0\\ 0\\ 0\end{bmatrix} .$$

למעשה קיבלנו שלוש משוואות זהות בהן נדרש שסכום איברי $\underline{v}$ יהיה אפס. לכן יש שני פתרונות אפשריים כוקטורים עצמיים,

$$\underline{v}_{1a}=\begin{bmatrix}1\\ -1\\ 0\end{bmatrix}\quad\text{and}\quad\underline{v}_{1b}=\begin{bmatrix}0\\ 1\\ -1\end{bmatrix} ,$$

וכל קומבינציה ליניארית של שני וקטורים אלה תהיה אף היא וקטור עצמי, כלומר קיבלנו תת מרחב דו-מימדי של פתרונות לווקטור העצמי הראשון.

באשר לווקטור העצמי השני, מתקבל

$$(\mathbf{A}-\lambda_2\mathbf{I})\underline{v}=\begin{bmatrix}-\lambda_2 & -1 & -1\\ 1 & 2-\lambda_2 & 1\\ 1 & 1 & 2-\lambda_2\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix}-2 & -1 & -1\\ 1 & 0 & 1\\ 1 & 1 & 0\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}.$$

מערכת זו ניתנת לפישוט ע"י אלימינציה גאוסית לקבלת

$$\begin{bmatrix}-2 & -1 & -1\\ 0 & -0.5 & 0.5\\ 0 & 0.5 & -0.5\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}\quad\Rightarrow\quad\begin{bmatrix}-2 & -1 & -1\\ 0 & -1 & 1\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}.$$

לכן הפתרון יהיה

$$\underline{v}_2=\begin{bmatrix}-1\\ 1\\ 1\end{bmatrix}$$

או כל מכפלה שלו בסקלר.

באופן כללי יותר נוכל לומר כי עבור מטריצה **A** ממשית כללית בגודל n-על-n, המשוואה האופיינית תוביל לפולינום מסדר n אשר לו n שורשים – חלקם ממשיים וחלקם מדומים. כמות הערכים העצמיים השונים יכולה לנוע בין 1 (במקרה שכל הערכים העצמיים זהים) ועד n (במקרה בו כל הערכים העצמיים שונים זה מזה).

כאשר נקבל ערך עצמי מרוכב $\lambda=\mu+i\eta$, גם הצמוד לו, $\bar{\lambda}=\mu-i\eta$, יופיע כערך עצמי. זאת כיוון שבשל היות המטריצה **A** ממשית, כל מקדמי הפולינום האופייני ממשיים. אם לערך העצמי $\lambda=\mu+i\eta$ נמצא הווקטור העצמי $\underline{v}$, נקבל

$$\mathbf{A}\underline{v}=\lambda\underline{v}\quad\Rightarrow\quad\overline{\mathbf{A}\underline{v}}=\mathbf{A}\bar{\underline{v}}=\bar{\lambda}\bar{\underline{v}}$$

ולכן הווקטור העצמי התואם ל-$\bar{\lambda}=\mu-i\eta$ יהיה הצמוד של $\underline{v}$.

כשערך עצמי מופיע k פעמים, נאמר כי לערך עצמי זה יש ריבוי אלגברי $r_a=k$. כבר ראינו קודם מקרה בו יש ריבוי אלגברי 2 לערך עצמי, והדבר גזר שהוקטורים העצמיים התואמים פרסו תת-מרחב דו-מימדי. למימד תת-מרחב זה נקרא ריבוי גיאומטרי, r_g. בכל מקרה, $r_g\geq 1$ כיוון שאם נמצא ערך עצמי ההופך את המטריצה $(\mathbf{A}-\lambda\mathbf{I})$ לסינגולארית, הרי שבהכרח יש וקטור שונה מאפס בגרעין שלה. במקרה הכללי קיים לכל ערך עצמי הקשר - $r_a\geq r_g\geq 1$ (מובא ללא הוכחה). דוגמה מעניינת בהקשר זה היא המטריצה הבאה, המוכרת בשל בלוק-ג'ורדן:

$$\mathbf{A} = \begin{bmatrix} \lambda_0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \lambda_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \lambda_0 \end{bmatrix}$$

הפולינום האופייני כאן הינו

$$\det\{\mathbf{A} - \lambda\mathbf{I}\} = (\lambda - \lambda_0)^n \,.$$

לכן יש לנו כאן מקרה בו הערך העצמי λ_0 מופיע בריבוי עצמי $r_a = n$. באשר לוקטורים העצמיים התואמים לו, נקבל

$$0 = (\mathbf{A} - \lambda\mathbf{I})\underline{v} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ \vdots \\ v_6 \end{bmatrix} \,.$$

לכן הווקטור העצמי היחיד שפותר מערכת זו הוא $\underline{v}^T = [1,0,0, \ \dots \ 0,0,0]$. לכן במקרה זה הריבוי הגיאומטרי הוא $r_g = 1$.

ישנן שתי תכונות מעניינות הקושרות בין הערכים העצמיים ובין איברי המטריצה המקורית **A**. תכונות אלו הן פועל יוצא ישיר של פיתוח המשוואה האופיינית לפולינום, מחד, ופיתוח הפולינום כאוסף של מכפלת מונומיאלים. אנו נביא תכונות אלו ללא הוכחה:

$$\lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdots \lambda_n = \prod_{k=1}^{n} \lambda_k = \det\{\mathbf{A}\}$$
$$\lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_n = \sum_{k=1}^{n} \lambda_k = \sum_{k=1}^{n} a_{kk} = \mathrm{trace}\{\mathbf{A}\} \,.$$

מקבץ תכונות אחר שלא יידון כאן הוא חישוב פולינומים של מטריצות. אנו יודעים להעלות מטריצה בריבוע, ובכל חזקה שלמה אחרת ולכן חישוב של פולינום כזה אינו קשה. מסתבר כי הפולינום האופייני של מטריצה הוא פולינום אותו המטריצה מאפסת! תכונה זו פירושה שחזקות גבוהות של המטריצה ניתנות לרישום כקומבינציה ליניארית של חזקות נמוכות (עד דרגה n). מכך גם נובע שהיפוך מטריצה כזו ניתן לתיאור כקומבינציה כזו.

2. לכסון של מטריצות

מסתבר כי אוסף הוקטורים העצמיים יכול לסייע לנו רבות בהבנת המטריצה, ובפירוקה למכפלה מלכסנת. הבעיה היא שתהליך כזה יכול לפעול רק למטריצות מסוימות, בהן הריבויים הגיאומטרי והאלגברי מתלכדים. נראה זאת כעת, ונוביל לפירוקים מלכסנים.

נניח שעבור מטריצה **A**, בידינו שני וקטורים עצמיים $\underline{v}_1, \underline{v}_2$, והערכים העצמיים התואמים להם אשר שונים זה מזה, $\lambda_1 \neq \lambda_2$. האם יתכן ששני הוקטורים העצמיים הללו תלויים ליניארית? נניח שלא! לכן קיימים מקדמים $c_1, c_2 \neq 0$ כך ש:

$$c_1 \underline{v}_1 + c_2 \underline{v}_2 = 0 .$$

נכפיל משוואה זו במטריצה **A** ונקבל

$$\mathbf{A}\left(c_1 \underline{v}_1 + c_2 \underline{v}_2\right) = c_1 \lambda_1 \underline{v}_1 + c_2 \lambda_2 \underline{v}_2 = 0 .$$

מצד שני, נוכל לכפול את אותה משוואה מקורית ב- λ_1 (נניח שהוא אינו אפס – אחד מהשניים חייב להיות כי הם שונים) ולקבל

$$\lambda_1\left(c_1 \underline{v}_1 + c_2 \underline{v}_2\right) = c_1 \lambda_1 \underline{v}_1 + c_2 \lambda_1 \underline{v}_2 = 0 .$$

החסרת שני המשוואות המתקבלות מובילה לדרישה

$$c_2\left(\lambda_1 - \lambda_2\right)\underline{v}_2 = 0 .$$

מכיוון שהמקדם c_2 שונה מאפס, וכך גם ההפרש $\left(\lambda_1 - \lambda_2\right)$, מתקבלת דרישה בלתי אפשרית, ממנה אנו מסיקים כי שני הוקטורים העצמיים בלתי תלויים ליניארית. בדרך זהה ניתן להוכיח את המשפט הבא,

משפט 1: אם $\lambda_1, \lambda_2, \lambda_3, \cdots, \lambda_k$ הם ערכים עצמיים שונים של המטריצה **A**, אזי הוקטורים העצמיים התואמים להם בלתי תלויים ליניארית.

נובעת ממשפט זה באופן ישיר הטענה הבאה:

משפט 2: אם מטריצה **A** בעלת n ערכים עצמיים שונים זה מזה, אזי n הוקטורים העצמיים התואמים להם בלתי תלויים ליניארית, ולכן מהווים בסיס למרחב ה-n מימדי.

מה קורה כאשר ערכים עצמיים מופיעים בריבוי אלגברי? אם במקרה כזה יתקבל כי הריבוי האלגברי והריבוי הגיאומטרי זהים, נוכל שוב לטעון שהווקטורים העצמיים בונים בסיס.

למה כל זה חשוב לנו? מכיוון שאם בידינו אוסף וקטורים המהווים בסיס מחד, אבל כל אחד מהם הוא גם וקטור עצמי של המטריצה **A**, נוכל לתאר את המכפלה $\mathbf{A}\underline{x}$ בצורה

אחרת, נוחה יותר. נניח כי עבור מטריצה **A** יש לנו סט של וקטורים עצמיים בלתי-תלויים ליניארית,

$$\mathbf{A}\underline{v}_k = \lambda_k \underline{v}_k \qquad n\,, \ldots ,k = 1,2.$$

חשוב להבהיר, מרגע שאמרנו כי הוקטורים העצמיים מהווים בסיס, ריבוי אלגברי של ערך עצמי כלשהו חסר חשיבות לחלוטין לענייננו.

כעת בידינו וקטור $\underline{x}$ אותו אנו מעוניינים לכפול ב-**A**. הווקטור $\underline{x}$ יכול להיכתב אחרת כצירוף ליניארי של הוקטורים העצמיים (בהיותם בסיס). לכן נרשום

$$\underline{x} = \sum_{k=1}^{n} c_k \underline{v}_k .$$

ולכן, המכפלה הרצויה תהיה

$$\mathbf{A}\underline{x} = \sum_{k=1}^{n} c_k \mathbf{A}\underline{v}_k = \sum_{k=1}^{n} c_k \lambda_k \underline{v}_k .$$

אנו רואים כי לו רק יכולנו לייצג כל וקטור $\underline{x}$ בו יש לנו עניין כצירוף ליניארי של הווקטורים העצמיים, ההכפלה ב- **A** הופכת להיות קלה מאוד – למעשה גורם סקלר מתקן לכל מקדם ולא יותר.

את שנאמר למעלה נוכל לרשום מעט אחרת, ולהגיע לתובנה מעניינת. נגדיר מטריצה **V** בה העמודות הם הוקטורים העצמיים,

$$\mathbf{V} = \begin{bmatrix} | & | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \underline{v}_3 & \cdots & \underline{v}_5 \\ | & | & | & & | \end{bmatrix} .$$

אזי הווקטור $\underline{x}$ נתון כמכפלה

$$\underline{x} = \mathbf{V}\underline{c} = \begin{bmatrix} | & | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \underline{v}_3 & \cdots & \underline{v}_5 \\ | & | & | & & | \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_5 \end{bmatrix} .$$

וההכפלה ב-**A** נותנת

$$\mathbf{A}\underline{x} = \mathbf{AV}\underline{c} = \begin{bmatrix} | & | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \underline{v}_3 & \cdots & \underline{v}_5 \\ | & | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_5 \end{bmatrix} = \mathbf{V}\Lambda\underline{c} \,.$$

כיוון שהקשר הנ"ל נכון לכל וקטור מקדמים $\underline{c}$, נציע סידרת משוואות כאלה בהן $\underline{c}$ מקבל את וקטורי היחידה הטריוויאליים $\underline{e}_k$,

$$\mathbf{AV}\underline{e}_k = \mathbf{V}\Lambda\underline{e}_k \quad k = 1,2, \dots ,n \,.$$

קיבוץ כל משוואות אלה יחדיו נותנת

$$\mathbf{AV}\begin{bmatrix} | & | & | & | & | \\ \underline{e}_1 & \underline{e}_2 & \underline{e}_3 & \cdots & \underline{e}_n \\ | & | & | & & | \end{bmatrix} = \mathbf{AVI} = \mathbf{V}\Lambda\begin{bmatrix} | & | & | & | & | \\ \underline{e}_1 & \underline{e}_2 & \underline{e}_3 & \cdots & \underline{e}_n \\ | & | & | & & | \end{bmatrix} = \mathbf{V}\Lambda\mathbf{I} \,.$$

ההכפלה ב-**I** בשני האגפים ניתנת להשמטה, ומביאה לקשר

$$\mathbf{AV} = \mathbf{V}\Lambda \quad \Rightarrow \quad \mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1} \,,$$

אשר אומר כי את המטריצה **A** הצלחנו ללכסן ע"י העתקת דמיון. המטריצה האלכסונית שהתקבלה כוללת באלכסונה את הערכים העצמיים, ופעולת הדמיון (הכפלה ב- **V** מימין ובהיפוכה משמאל) בנויה על מטריצה בה העמודות הם הוקטורים העצמיים.

אגב, עיון במשוואה $\mathbf{AV} = \mathbf{V}\Lambda$ מראה קשר ברור לוקטורים והערכים העצמיים – אם ניקח משני אגפי המשוואה את העמודה ה-k, זו תהיה

$$\mathbf{A}\underline{v}_k = \mathbf{V}\begin{bmatrix} 0 \\ \vdots \\ \lambda_k \\ \vdots \\ 0 \end{bmatrix} = \lambda_k \underline{v}_k \,.$$

כלומר, הקשר $\mathbf{AV} = \mathbf{V}\Lambda$ אינו אלא קיבוץ של כל המשוואות $\mathbf{A}\underline{v}_k = \lambda_k \underline{v}_k$ למערכת אחת. העובדה שניתן להפוך את המטריצה **V** לטובת הלכסון נובעת מהיות עמודותיה בלתי-תלויות. קיבלנו למעשה את המשפט הבא:

משפט 3: מטריצה ריבועית **A** נקרית "לכסינה" (diagonalizable) אם קיימת מטריצה לה סינגולארית **S** ומטריצה אלכסונית **D** כך ש:

$$\mathbf{S}^{-1}\mathbf{AS} = \mathbf{D} \,.$$

במקרה כזה, המטריצה **S** מכילה בהכרח את הווקטורים העצמיים של **A**, והאלכסון של **D** מכיל את ערכיהם העצמיים התואמים. יתרה מכך – פירוק זה הוא יחיד (עד כדי פרמוטציה של הוקטורים העצמיים וסימנם).

מהנאמר למעלה ברור כי מטריצה תהיה לכסינה רק אם כל ערכיה העצמיים שונים זה מזה (דרך אחת להבטיח כי הוקטורים העצמיים בלתי תלויים) או אם יש בהם ריבויים אלגבריים וגיאומטריים זהים.

אם נחזור לדוגמה שעלתה קודם עם המטריצה

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

הערכים\וקטורים העצמיים הם $\lambda_1 = 1$ בריבוי 2, ועם שני וקטורים עצמיים,

$$\underline{v}_{1a} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \text{and} \quad \underline{v}_{1b} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

ו- $\lambda_1 = 2$ עם הווקטור העצמי התואם

$$\underline{v}_2 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}.$$

לכן, נוכל להציע פירוק מהצורה

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

משפט 3 למעלה מציע שימוש בטרנספורמציית דמיון על מנת להגיע למבנה פשוט – אלכסוני במקרה זה. ישנו פירוק אחר מאוד מעניין הקשור לפירוק הנ"ל, ומוכר בשם פירוק Schur (Schur decomposition). בפירוק Schur המטריצה **S** מוחלפת במטריצה אורתונורמלית. לכן במקרה הכללי היא אינה יכולה להביא ללכסון! מסתבר כי ניתן להבטיח הגעה למטריצה משולשת עליונה, כפי שהמשפט הבא טוען:

משפט 4: בהינתן מטריצה **A** בגודל n-על-n לכסינה, ניתן למצוא מטריצה אורתונורמלית **Q** כך ש:

$$\mathbf{Q}^{\top}\mathbf{AQ} = \mathbf{T}$$

כש-**T** מטריצה משולשת עליונה.

מסתבר כי הוכחת משפט זה אינה קשה, והיא ניתנת לבניה ע"י שילוב של ה-QR ולכסון לעיל. בשל היות המטריצה לכסינה, יש אפשרות לקבל את הפירוק

$$\mathbf{S}^{-1}\mathbf{AS} = \mathbf{D}.$$

כש-**S** מטריצה לא סינגולארית. פירוק QR מבטיח לנו כי למטריצה **S** נוכל לתת פירוק מהצורה **S=QR**. לכן מתקבל,

$$\mathbf{S}^{-1}\mathbf{AS} = (\mathbf{QR})^{-1}\mathbf{A}(\mathbf{QR}) = \mathbf{R}^{-1}\mathbf{Q}^{\top}\mathbf{AQR} = \mathbf{D}.$$

מכיוון ש-**R** אינה סינגולארית נוכל להכפיל בהיפוכה ולקבל

$$\mathbf{Q}^{\top}\mathbf{AQ} = \mathbf{RDR}^{-1} = \mathbf{T}.$$

כבר ראינו כי היפוך של מטריצה משולשת עליונה גם הוא מטריצה משולשת עליונה, מכפלתן נותרת כזו, והכפלתן במטריצה אלכסונית מותירה מבנה זה גם כן. אם כך, קיבלנו כי **T** היא מטריצה משולשת עליונה כפי שנטען.

פירוק Schur פשוט לאין-ערוך מלכסון, ועם זאת הוא מספק לנו ישירות את הערכים העצמיים של המטריצה – אלה יהיו איברי האלכסון הראשי של **T**.

3. מטריצות סימטריות וחיוביות מוגדרות

לנו יש עניין מיוחד במטריצות סימטריות ובכאלה שהינן חיוביות (חצי) מוגדרות, כיוון שהן עולות לעיתים קרובות במדעים ובהנדסה. נתחיל את הדיון במטריצות סימטריות. מה קורה כאשר מטריצה **A** סימטרית? נניח כי מטריצה זו ממשית. אזי מתקבלות התוצאות הבאות אשר יובאו כאן ללא הוכחה:

משפט 5: אם **A** מטריצה סימטרית וממשית בגודל n-על-n, אז

- כל ערכיה ווקטוריה העצמיים ממשיים,
- וקטורים עצמיים המתייחסים לערכים עצמיים שונים יהיו אורתוגונאליים,
- המטריצה לכסינה. כלומר, גם אם ערכך עצמי מופיע בריבוי, ריבוייו הגיאומטרי יהיה זהה.

לדוגמה, נתייחס למטריצה הסימטרית הבאה:

$$\mathbf{A} = \begin{bmatrix} 5 & -4 & 2 \\ -4 & 5 & 2 \\ 2 & 2 & -1 \end{bmatrix}.$$

מאמץ קטן יוביל לסט הערכים והווקטורים העצמיים הבאים:

$$\lambda_1 = -3, \underline{v}_1 = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \qquad \lambda_1 = 3, \underline{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad \lambda_1 = 9, \underline{v}_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

וניכר כי וקטורים אלה אורתוגונאליים.

העובדה שהוקטורים העצמיים מהווים בסיס כבר נטענה קודם למטריצות לכסינות. כעת בסיס זה הינו גם אורתוגונאלי, ולכך יתרון עצום! בהינתן וקטור $\underline{x}$ ייצוגו ע"י בסיס זה (מציאת וקטור המקדמים $\underline{c}$) קל לקבלה ע"י סידרת מכפלות פנימיות פשוטות, ולא כמקודם. מעבר לכך, תוצאה זו מובילה למשפט הבא:

משפט 6: אם **A** מטריצה סימטרית וממשית בגודל n-על-n, אז היא ניתנת ללכסון אורתוגונאלי מהצורה

$$\mathbf{Q}^T\mathbf{AQ} = \mathbf{D},$$

כאשר **Q** מטריצה אורתונורמלית אשר עמודותיה הם הווקטורים העצמיים, ו-**D** מטריצה אלכסונית עם הערכים העצמיים באלכסונה.

תכונת הדמיון האורתוגונאלי הנ"ל ברורה, ומזכירה את פירוק Schur (אם כי הפעם לא מתקבלת מטריצה משולשת עליונה אלא אלכסונית). כל שעלינו לעשות על מנת לקבלה הוא לקחת את הווקטורים העצמיים ולנרמלם, עובדה שלא פוגעת בזהותם כוקטורים עצמיים. כעת הקשר שקיבלנו קודם על מטריצה לכסינה,

$$\mathbf{S}^{-1}\mathbf{AS} = \mathbf{D},$$

תקף, אך הפעם גם מתקבל כי **S** מטריצה אורתונורמלית, ולכן היפוכה ניתן לקבלה ע"י שחלוף פשוט. תכונה זו גם מוכרת בשם פירוק ספקטראלי, כיוון שאוסף הערכים העצמיים מכונה גם ה-"ספקטרום" של המטריצה.

הערה: יש דמיון מטעה בין הפירוק הנ"ל לפירוק LDV למטריצות סימטריות. אין קשר בין שני הפירוקים, והמטריצה האלכסונית ב-LDV אינה מכילה את הערכים העצמיים.

מה קורה עבור מטריצות חיוביות (חצי) מוגדרות? נניח כי בידינו מטריצה חיובית מוגדרת **K**. פירוש הדבר שמתקיים (לפי הגדרה):

$$\underline{x}^T\mathbf{K}\underline{x} > 0 \quad \forall \underline{x} \neq 0 \, .$$

נוכל כעת להשתמש בפירוק הספקטראלי המשתמש בווקטוריה וערכיה העצמיים, ולקבל

$$\underline{x}^T\mathbf{K}\underline{x} = \underline{x}^T\mathbf{QDQ}^T\underline{x} = \left(\underline{x}^T\mathbf{Q}\right)\mathbf{D}\left(\mathbf{Q}^T\underline{x}\right) > 0 \quad \forall \underline{x} \neq 0 \, .$$

אם נגדיר $\underline{u} = \mathbf{Q}^T\underline{x}$, הדרישה הנ"ל שקולה לדרישה

$$\underline{x}^T\mathbf{K}\underline{x} = \underline{u}\mathbf{D}\underline{u} = \sum_{k=1}^{n}\lambda_k u_k^2 > 0 \quad \forall \underline{u} \neq 0 \, .$$

על מנת שדרישה זו תתקיים, יש לדרוש שכל הערכים העצמיים של **K** הינם חיוביים ממש. לסיכום, קיימת התכונה החשובה הבאה:

משפט 7: אם **A** מטריצה סימטרית, ממשית וחיובית מוגדרת, בגודל n-על-n, אז כל ערכיה העצמיים ממשיים וחיוביים. אם מטריצה זו חיובית חצי מוגדרת, ערכיה העצמיים ממשיים ואי-שליליים.

תכונה זו מעניינת וחשובה. כזכור, כשנתונה מטריצה סימטרית וברצוננו לברר האם היא חיובית מוגדרת, דרך אפשרית לפעולה היא פירוק Cholesky. כעת אנו רואים אלטרנטיבה מעניינת – חישוב הערכים העצמיים ובחינת היותם חיובים (או אי-שליליים, במקרה של מטריצה חיובית חצי מוגדרת).

הערה: ישנה משפחת מטריצות רחבה יותר ממשפחת המטריצות הסימטריות הקרויה "מטריצות נורמאליות". מטריצה **A** היא נורמאלית אם $\mathbf{A}^H\mathbf{A} = \mathbf{A}\mathbf{A}^H$, כשעשינו שימוש בשחלוף צמוד. ברור כי כל מטריצה אורתוגונאלית (גם מרוכבת!) הינה נורמאלית כיוון ש- $\mathbf{Q}^H\mathbf{Q} = \mathbf{Q}\mathbf{Q}^H = \mathbf{D}$, כש-**D** מכיל את נורמת הוקטורים בעמודות המטריצה **Q**. גם מטריצות סימטריות (במובן הכללי יותר המתייחס לאפשרות של כניסות מרוכבות) הן נורמאליות, וממילא גם המטריצות החיוביות מוגדרות. מסתבר כי היכולת ללכסן ע"י דמיון יוניטרי (אורתונורמלי מרוכב) קיימת במקור למשפחה רחבה זו, וכך גם תכונות אחרות. אנו לא נתייחס יותר למשפחה זו של מטריצות, ונתעמק בתת-משפחות שלה אשר להן ערך מעשי ביישומים רבים.

4. ערכים עצמיים כפתרון בעיות אופטימיזציה

כשדנו במערכות של משוואות, ראינו כי ניתן להמיר את פתרונן בבעיית מינימיזציה, ולהפיק מכך תועלת (כגון קיום פתרון מקורב, גם כשאין פתרון למערכת המקורית). האם ניתן למצוא אנלוגיה דומה בערכים עצמיים? האם יתכן שערכים או וקטורים

עצמיים יצמחו כפתרונה של בעיית אופטימיזציה? מסתבר שהתשובה חיובית, ומוכרת כתכונת Rayleigh-Ritz. אנו נפתח תכונה זו כשאנו מתחילים במטריצות פשוטות, ומשם נסבך את העניינים יותר ויותר עד לטיפול במגוון רחב יותר ומעניין של מטריצות.

נניח כי בידינו מטריצה אלכסונית עם ערכים ממשיים על האלכסון,

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix},$$

כשאיברי האלכסון מסודרים באופן יורד, $d_1 \ge d_2 \ge d_3 \ge \cdots \ge d_n$. נניח כי ברצוננו להביא למקסימום את הביטוי הריבועי

$$q(\underline{y}) = \underline{y}^T \mathbf{D} \underline{y} = \sum_{k=1}^{n} d_k y_k^2 .$$

ברור כי פונקציה זו לא חסומה, וניתן להביאה לאינסוף ע"י ניפוח איברי הווקטור $\underline{y}$. על מנת למנוע מצב זה נוכל לדרוש כי $\underline{y}$ יימצא על פני כדור היחידה, כלומר,

$$\underline{y}^* = \underset{\underline{y} |\ \|\underline{y}\|_2^2 = 1}{\arg\max}\ \underline{y}^T \mathbf{D} \underline{y}$$

מהו הפתרון במקרה זה? נשתמש באי-השוויון הבא:

$$q(\underline{y}) = \underline{y}^T \mathbf{D} \underline{y} = \sum_{k=1}^{n} d_k y_k^2 \le \sum_{k=1}^{n} d_1 y_k^2 = d_1 \cdot \sum_{k=1}^{n} y_k^2 = d_1 .$$

אנו רואים כי תמיד נקבל כי ערך הפונקציה קטן או שווה לערך העצמי הגדול ביותר ב-**D**, הלא הוא d_1. יתרה מזו, עבור הבחירה

$$\underline{y}^T = [1 \quad 0 \quad 0 \quad \cdots \quad 0]$$

ערך זה מושג בדיוק, עובדה שפירושה שזהו הפתרון של בעיית האופטימיזציה שבידינו.

באופן דומה, פתרון של בעיית מינימיזציה דומה,

$$\underline{y}^* = \underset{\underline{y} |\ \|\underline{y}\|_2^2 = 1}{\arg\min}\ \underline{y}^T \mathbf{D} \underline{y}$$

מוביל לערך מינימאלי שהוא לא אחר מהערך העצמי הקטן ביותר, כיוון שהבחירה

$$\underline{y}^T = [0 \quad \cdots \quad 0 \quad 0 \quad 1]$$

ייתן את אי-השוויון הבא

$$.\, q(\underline{y}) = \underline{y}^T \mathbf{D} \underline{y} = \sum_{k=1}^{n} d_k y_k^2 \geq \sum_{k=1}^{n} d_n y_k^2 = d_n \cdot \sum_{k=1}^{n} y_k^2 = d_n$$

בשוויון.

נניח כעת כי אנו עוסקים במטריצה סימטרית כלשהי **A**, וברצוננו להביא לפתרון את הבעיות הנ"ל (נניח מקסימיזציה, כיוון שהדברים דומים). נשתמש בעובדה שלמטריצה **A** יש פירוק אורתונורמלי מלכסן ולכן,

$$.\, \underline{x}^* = \underset{\underline{x} \mid \|\underline{x}\|_2^2 = 1}{\arg\max}\ \underline{x}^T \mathbf{A} \underline{x} = \underset{\underline{x} \mid \|\underline{x}\|_2^2 = 1}{\arg\max}\ \underline{x}^T \mathbf{Q} \mathbf{D} \mathbf{Q}^T \underline{x}$$

שימוש בסימון $\underline{y} = \mathbf{Q}^T \underline{x}$ (שממנו נובע גם $\mathbf{Q}\underline{y} = \underline{x}$) נותן רישום אחר לאותה בעיה,

$$.\, \underline{x}^* = \underset{\underline{x} \mid \|\underline{x}\|_2^2 = 1}{\arg\max}\ \underline{x}^T \mathbf{Q} \mathbf{D} \mathbf{Q}^T \underline{x} = \underset{\underline{y} \mid \|\mathbf{Q}\underline{y}\|_2^2 = 1}{\arg\max}\ \underline{y}^T \mathbf{D} \underline{y}$$

האילוץ $\|\mathbf{Q}\underline{y}\|_2^2 = 1$ שקול לחלוטין לאילוץ $\|\underline{y}\|_2^2 = 1$ כפי שכבר ראינו בעבר. לכן התנקזנו לאותה בעיה בדיוק ונוכל לומר את המשפט הבא:

משפט 8: למטריצה ריבועית וסימטרית בגודל n-על-n, **A**, בעלת צמדים עצמיים $\{\lambda_k, \underline{v}_k\}_{k=1}^{n}$ המאורגנים בסדר יורד של הערכים העצמיים, מתקבל כי

$$\left\{ \underline{v}_1 = \underset{\underline{x} \mid \|\underline{x}\|_2^2 = 1}{\arg\max}\ \underline{x}^T \mathbf{A} \underline{x} \right\} \quad \left\{ \lambda_1 = \max_{\underline{x} \mid \|\underline{x}\|_2^2 = 1} \underline{x}^T \mathbf{A} \underline{x} \right\}$$
$$.\, \left\{ \underline{v}_n = \underset{\underline{x} \mid \|\underline{x}\|_2^2 = 1}{\arg\min}\ \underline{x}^T \mathbf{A} \underline{x} \right\} \quad \left\{ \lambda_n = \min_{\underline{x} \mid \|\underline{x}\|_2^2 = 1} \underline{x}^T \mathbf{A} \underline{x} \right\}$$

למעשה, בעיות אלו ניתנות לרישום אלטרנטיבי נטול אילוץ על נורמת הנעלם. אם נחליף את הופעת $\underline{x}$ בבעיות אלו בביטוי $\underline{x}/\|\underline{x}\|_2$ הרי שהווקטור מנורמל מהגדרתו. אז יתקבל למשל בבעיית המקסימיזציה

$$\underline{x}^* = \underset{\underline{x} \mid \|\underline{x}\|_2^2 = 1}{\arg\max}\ \underline{x}^T \mathbf{A} \underline{x} = \underset{\underline{x}}{\arg\max}\ \frac{\underline{x}^T}{\|\underline{x}\|_2} \mathbf{A} \frac{\underline{x}}{\|\underline{x}\|_2} =$$
$$.\, = \underset{\underline{x}}{\arg\max}\ \frac{\underline{x}^T \mathbf{A} \underline{x}}{\|\underline{x}\|_2^2} = \underset{\underline{x}}{\arg\max}\ \frac{\underline{x}^T \mathbf{A} \underline{x}}{\underline{x}^T \underline{x}}$$

קיבלנו מנה של שתי תבניות ריבועיות כפונקצית המטרה כאן.

ממש כשם שנעשה הדבר בהקשר של ה-LS, גם כאן נוכל לאמץ גישת מקסימיזציה או מינימיזציה ישירה שתכליתה להשיג את הפתרון משיקולי איפוס וקטור הנגזרות. כיוון שזוהי פונקצית מנה (עבורה יש לנו נוסחת נגזרת), וכיוון שבמונה ובמכנה יש ביטויים ריבועיים שאותם אנו יודעים לגזור, מתקבל

$$\mathrm{grad}\{f(\underline{x})\} = \begin{bmatrix} \frac{\partial f(\underline{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\underline{x})}{\partial x_n} \end{bmatrix} = \frac{\underline{x}^T \mathbf{A} \underline{x}}{\underline{x}^T \underline{x}} = \frac{(2\mathbf{A}\underline{x})(\underline{x}^T \underline{x}) - (2\underline{x})(\underline{x}^T \mathbf{A} \underline{x})}{(\underline{x}^T \underline{x})^2} = 0 \ .$$

נשים לב כי נוסחת הנגזרת תואמת מבחינת מימדים את מימד הגרדיאנט הצפוי. איפוס ביטוי זה מוביל לאפשרות להתעלם מהמכנה בהיותו שונה מאפס (למניעת פתרון טריוויאלי), ולכן

$$(2\mathbf{A}\underline{x})(\underline{x}^T \underline{x}) - (2\underline{x})(\underline{x}^T \mathbf{A} \underline{x}) = 0 \quad \Rightarrow \quad \mathbf{A}\underline{x} = \frac{\underline{x}^T \mathbf{A} \underline{x}}{\underline{x}^T \underline{x}} \underline{x} = f(\underline{x})\underline{x} \ .$$

קיבלנו כי על הפתרון להיות וקטור עצמי של המטריצה **A**, כשהערך העצמי יוצא ערך הפונקציה. לכן, בידיעה שיש לנו n ערכים עצמיים, בחירת הווקטור העצמי המתייחס לערך העצמי המקסימאלי תיתן את הפתרון, ממש כפי שראינו קודם!

עד כה התמקדנו במציאת הערכים העצמיים הקיצוניים. מה באשר לאחרים שביניהם? כיצד ננסח בעיית אופטימיזציה שתיתן אותם כפתרון? זה קל מאוד! נניח כי מצאנו את j הוקטורים העצמיים המתייחסים לערכים העצמיים הגבוהים ביותר, ורצוננו למצוא את ה-j+1 (ואת הערך העצמי התואם לו). נארגן את המטריצה U בגודל n שורות על j עמודות, בה הוקטורים שנמצאו משמשים כעמודות, ונפתור את הבעיה

$$\underline{x}^* = \underset{\underline{x} |\ \mathbf{U}^T \underline{x} = 0}{\arg\max} \ \frac{\underline{x}^T \mathbf{A} \underline{x}}{\underline{x}^T \underline{x}} \ .$$

מטבע הדברים, כיוון שעבור מטריצות סימטריות הוקטורים העצמיים יוצרים בסיס אורתונורמלי, האילוץ יבטיח (תיאורטית) כי הווקטור שיימצא מתייחס לערך העצמי ה-j+1.

אם נפעיל תהליך GS על אוסף הוקטורים שנמצאו ונגדיר את המטריצה $\bar{\mathbf{U}}$ המשלימה לה (בה יש n-j עמודות אורתוגונאליות לאלה שמצאנו עד כה), נוכל לייצג את הבעיה מעט אחרת כיוון שהאילוץ $\mathbf{U}^T \underline{x} = 0$ שקול לדרישה

$$\underline{x} = \bar{\mathbf{U}} \underline{\alpha} \ .$$

זאת כיוון שהווקטור $\underline{x}$ שבו אנו מעוניינים חייב להיות צירוף ליניארי של יתר וקטורי הבסיס האורתונורמלי. הצבת ביטוי זה בפונקציה שלנו תיתן

$$\underline{x}^* = \arg\max_{\underline{x} |\ \mathbf{U}^T \underline{x}=0} \frac{\underline{x}^T \mathbf{A} \underline{x}}{\underline{x}^T \underline{x}} = \arg\max_{\underline{\alpha}} \frac{\underline{\alpha}^T \bar{\mathbf{U}}^T \mathbf{A} \bar{\mathbf{U}} \underline{\alpha}}{\underline{\alpha}^T \bar{\mathbf{U}}^T \bar{\mathbf{U}} \underline{\alpha}} = \arg\max_{\underline{\alpha}} \frac{\underline{\alpha}^T \tilde{\mathbf{A}} \underline{\alpha}}{\underline{\alpha}^T \underline{\alpha}} .$$

המכפלה $\bar{\mathbf{U}}^T \bar{\mathbf{U}}$ נותנת מטריצת יחידה בגודל (n-j)-על-(n-j). המכפלה $\bar{\mathbf{U}}^T \mathbf{A} \bar{\mathbf{U}}$ יוצרת מטריצה חדשה $\tilde{\mathbf{A}}$ בגודל (n-j)-על-(n-j). לכן קיבלנו בעיה מהמבנה הקודם עם גודל קטן יותר. שוב ערכים עצמיים ווקטורים עצמיים (והפעם של המטריצה המוקטנת) ישחקו תפקיד.

5. ערכים עצמיים מוכללים

מה היה קורה לו רצינו להביא למקסימום את הפונקציה המעט כללית יותר,

$$\underline{x}^* = \arg\max_{\underline{x}} \frac{\underline{x}^T \mathbf{A} \underline{x}}{\underline{x}^T \mathbf{B} \underline{x}} .$$

מה משמעות פונקציה זו ותוצאות ערכיה הקיצוניים? אנו נניח כי המטריצה **B** חיובית מוגדרת על מנת להבטיח כי אין כאן חלוקה באפס. משמעות ביטוי זה אינטואיטיבית ואומרת כי בבואנו להביא למקסימום את הביטוי $\underline{x}^T \mathbf{A} \underline{x}$, לא נגביל את הפתרונות לכדור היחידה אלא לאליפסויד n-מימדי המאופיין ע"י $\underline{x}^T \mathbf{B} \underline{x} = 1$. המנה המתוארת בפונקצית המטרה מוכרת בשם Rayleigh Quotient.

נוכל לפעול בדרך דומה לנעשה קודם ולהשוואת הנגזרת לאפס. פעולה זו תיתן:

$$\operatorname{grad}\{f(\underline{x})\} = \begin{bmatrix} \frac{\partial f(\underline{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\underline{x})}{\partial x_n} \end{bmatrix} = \frac{\underline{x}^T \mathbf{A} \underline{x}}{\underline{x}^T \mathbf{B} \underline{x}} = \frac{(2\mathbf{A}\underline{x})(\underline{x}^T \mathbf{B} \underline{x}) - (2\mathbf{B}\underline{x})(\underline{x}^T \mathbf{A} \underline{x})}{(\underline{x}^T \mathbf{B} \underline{x})^2} = 0 .$$

כמקודם, ניתן להתעלם מהמכנה, ולכן

$$(2\mathbf{A}\underline{x})(\underline{x}^T \mathbf{B} \underline{x}) - (2\mathbf{B}\underline{x})(\underline{x}^T \mathbf{A} \underline{x}) = 0 \quad \Rightarrow \quad \mathbf{A}\underline{x} = \frac{\underline{x}^T \mathbf{A} \underline{x}}{\underline{x}^T \mathbf{B} \underline{x}} \mathbf{B}\underline{x} = f(\underline{x})\underline{x} .$$

קיבלנו תבנית חדשה ומעניינת מהצורה

$$\mathbf{A}\underline{x} = \lambda \mathbf{B}\underline{x} ,$$

אשר מכלילה את קונספט הערך והווקטור העצמי. ביטוי זה קרוי עפרון (Pencil), ופתרונות מערכת זו קרויים ערכים עצמיים ווקטורים עצמיים מוכללים (-Generalized eigenvalues and eigenvectors). לאלה יישומים רבים במדעים ובהנדסה. מסתבר כי

תכונות רבות של ערכים ווקטורי עצמיים ניתנים להכללה כזו, ובכלל תוצאות אלה קיומם של n צמדי פתרונות עבור **A** סימטרית, והיותם ממשיים. לכן בחירת הגדול מכולם (מבין הערכים העצמיים) כערכה של הפונקציה במקסימום, ומולו את הווקטור התואם לו מביאים לפתרון בעיה זו.

6. ערכים סינגולאריים ופירוק ה- SVD

למטריצה ריבועית הראינו כי ניתן להגדיר ערכים עצמיים, ולאלה חשיבות רבה בהבנת המטריצה ודרך פעולה על וקטורים. מה באשר למטריצות לא ריבועיות? לא יתכן שלא נוכל לומר דבר דומה! ברור כי הרעיון של וקטור וערך עצמי לא ניתנים להכללה פשוטה למטריצה מלבנית כיוון ש- $\mathbf{A}\underline{x}$ הוא וקטור שונה במימדו מ- $\lambda\underline{x}$, ולכן לא ניתן להשוותם כלל. מה אם כך ניתן לעשות?

כבר ההקשרים קודמים (LS, או בנית מטריצות Gram למשל) ראינו כי ממטריצה כלשהי ניתן להגיע למטריצה חיובית חצי מוגדרת **K** ע"י

$$\mathbf{K} = \mathbf{A}^T\mathbf{A}.$$

ננסה אם לאפיין מטריצות מלבניות ע"י טיפול במטריצת ה-Gram הנוצרת מהן. יופייה של גישה זו בכך שעבור מטריצה **A** <u>כלשהי</u> בגודל m שורות על n עמודות, למטריצת ה-Gram שלה תמיד יהיו n וקטורים עצמיים, ואלה אפילו יהוו בסיס אורתונורמלי. נניח אם כן כי בידינו אוסף הוקטורים והערכים העצמיים של **K**, דהיינו

$$\mathbf{K}\underline{u}_k = \mathbf{A}^T\mathbf{A}\underline{u}_k = \lambda_k\underline{u}_k.$$

אנו נניח כי הוקטורים העצמיים נורמלו. מה נוכל לומר על תמונתם - $\mathbf{A}\underline{u}_k$? מסתבר כי הם גם אורתוגונאליים! זאת כיוון ש:

$$\left(\mathbf{A}\underline{u}_k\right)^T\left(\mathbf{A}\underline{u}_k\right) = \underline{u}_j^T\mathbf{A}^T\mathbf{A}\underline{u}_k = \lambda_k\underline{u}_j^T\underline{u}_k = \lambda_k.$$

אם כך, נגדיר (שרירותית) עבור המטריצה **A** את הערכים הסינגולאריים של **A**, $\sigma_1, \sigma_2, \ldots, \sigma_n$ ע"י

$$\sigma_k = \sqrt{\lambda_k}.$$

מכיוון שהערכים העצמיים של **K** ממשיים ואי-שליליים, גם הערכים הסינגולאריים שהוגדרו הם כאלה. בדרך כלל מקובל לארגן את הערכים הסינגולאריים בסדר יורד, כך שמתקיים

$$\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_n \ge 0.$$

ברור כי בינתיים הגדרה זו נשמעת שרירותית לחלוטין וחסרת תרומה.

נניח כעת כי אמנם הערכים הסינגולאריים (וגם הערכים העצמיים בהתאמה) סודרו בסדר יורד, ונניח כי r הראשונים שונים מאפס, וכל היתר (n-r ערכים) מאופסים. נגדיר סט של r וקטורים אורתונורמליים,

$$\underline{v}_k = \frac{\mathbf{A}\underline{u}_k}{\left\|\mathbf{A}\underline{u}_k\right\|_2} = \frac{\mathbf{A}\underline{u}_k}{\sigma_k} \quad k = 1,2,\ \dots\ ,r .$$

כלומר, עבור וקטורים אלה שהינם במימד m מתקבל הקשר

$$\mathbf{A}\underline{u}_k = \sigma_k \underline{v}_k \quad \text{for} \quad k = 1,2,\ \dots\ ,r .$$

קשר זה "מזכיר מאוד" את היחס המאפיין ערכים עצמיים ($\mathbf{A}\underline{x} = \lambda\underline{x}$), כשהשוני הוא בכך ששני האגפים מתייחסים לקבוצות ווקטורים אורתונורמליות שונות.

עבור $k > r$ כל המכפלות $\mathbf{A}\underline{u}_k$ מתאפסות בהגדרה (בשל היותם וקטורים עצמיים עם ערך עצמי 0). לאלה נגדיר אוסף $\{v_k\}_{k=r+1}^{m}$ אשר ימשיך את האוסף הראשון $\{v_k\}_{k=1}^{r}$ לכלל בסיס אורתונורמלי. לכן, עבור אלה יתקיים הקשר

$$\mathbf{A}\underline{u}_k = 0 \cdot \underline{v}_k \quad \text{for} \quad k = r+1, r+2,\ \dots\ ,m .$$

כעת ניקח את כל המשוואות הללו ונאחדן למשוואה אחת ונקבל

$$\begin{bmatrix} | & & | & | & & | \\ \mathbf{A}\underline{u}_1 & \cdots & \mathbf{A}\underline{u}_r & \mathbf{A}\underline{u}_{r+1} & \cdots & \mathbf{A}\underline{u}_n \\ | & & | & | & & | \end{bmatrix} = \mathbf{A}\begin{bmatrix} | & & | & | & & | \\ \underline{u}_1 & \cdots & \underline{u}_r & \underline{u}_{r+1} & \cdots & \underline{u}_n \\ | & & | & | & & | \end{bmatrix} = \mathbf{AU} =$$

$$= \begin{bmatrix} | & & | & | & & | \\ \sigma_1\underline{v}_1 & \cdots & \sigma_r\underline{v}_r & 0\cdot\underline{v}_{r+1} & \cdots & 0\cdot\underline{v}_m \\ | & & | & | & & | \end{bmatrix} =$$

$$= \begin{bmatrix} | & & | & | & & | \\ \underline{v}_1 & \cdots & \underline{v}_r & \underline{v}_{r+1} & \cdots & \underline{v}_k \\ | & & | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \mathbf{0} \\ & & \sigma_r & & \\ & & & 0 & \\ \mathbf{0} & & & & \ddots \\ & & & & & 0 \end{bmatrix} = \mathbf{V}\Sigma$$

ולכן, קיבלנו קשר כולל מהצורה

$$\mathbf{AU} = \mathbf{V}\Sigma \quad \Rightarrow \quad \mathbf{A} = \mathbf{V}\Sigma\mathbf{U}^T .$$

פירוק זה אומר לנו כי מטריצה **A** כללית ניתנת להצגה כמכפלה של שתי מטריצות אורתונורמליות ומטריצה אלכסונית. חשוב להבהיר כי המטריצה Σ אינה ריבועית –

גודלה m שורות על n עמודות, ועל אלכסונה הראשי מתקבלים הערכים הסינגולאריים בסדר יורד. לדוגמה, למטריצה **A** בגודל 4 שורות על 6 עמודות פירוק זה ייראה כך:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \end{bmatrix} =$$

$$= \begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \\ v_{41} & v_{42} & v_{43} & v_{44} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_{11} & u_{21} & u_{31} & u_{41} & u_{51} & u_{61} \\ u_{12} & u_{22} & u_{32} & u_{42} & u_{52} & u_{62} \\ u_{13} & u_{23} & u_{33} & u_{43} & u_{53} & u_{63} \\ u_{14} & u_{24} & u_{34} & u_{44} & u_{54} & u_{64} \\ u_{15} & u_{25} & u_{35} & u_{45} & u_{55} & u_{65} \\ u_{16} & u_{26} & u_{36} & u_{46} & u_{56} & u_{66} \end{bmatrix}$$

ובאופן דומה, עבור **A** בגודל 6 שורות על 4 עמודות יתקבל

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{51} & a_{52} & a_{53} & a_{54} \\ a_{61} & a_{62} & a_{63} & a_{64} \end{bmatrix} =$$

$$= \begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} & v_{15} & v_{16} \\ v_{21} & v_{22} & v_{23} & v_{24} & v_{25} & v_{26} \\ v_{31} & v_{32} & v_{33} & v_{34} & v_{35} & v_{36} \\ v_{41} & v_{42} & v_{42} & v_{44} & v_{45} & v_{46} \\ v_{51} & v_{52} & v_{53} & v_{54} & v_{55} & v_{56} \\ v_{61} & v_{62} & v_{63} & v_{64} & v_{65} & v_{66} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_{11} & u_{21} & u_{31} & u_{41} \\ u_{12} & u_{22} & u_{32} & u_{42} \\ u_{13} & u_{23} & u_{33} & u_{43} \\ u_{14} & u_{24} & u_{34} & u_{44} \end{bmatrix}$$

נסכם תוצאה זו במשפט:

משפט 9: כל מטריצה ממשית שהיא **A** בגודל m שורות על n עמודות ניתנת לפירוק (Singular Value Decomposition – SVD) מהצורה $\mathbf{A} = \mathbf{V}\Sigma\mathbf{U}^T$ כאשר,

- **V** מטריצה אורתונורמלית בגודל m-על-m,
- **U** מטריצה אורתונורמלית בגודל n-על-n, ו-
- Σ מטריצה "אלכסונית" בגודל m שורות על n עמודות, בה איברי האלכסון ממשיים ואי-שליליים, ומסודרים באופן יורד, ולכל היותר min(m,n) מהם שונים מאפס.

כדוגמה מעניינת, נניח כי בידינו מטריצה **A** ריבועית וסימטרית. במקרה כזה נוכל לחשב

גם את ערכיה העצמיים וגם את ערכיה הסינגולאריים. מה הקשר בין השניים? במקרה זה הערכים העצמיים מקיימים

$$\mathbf{A}\underline{x} = \lambda\underline{x} \,,$$

והכפלה בשחלוף של **A** תיתן

$$\mathbf{A}^T\mathbf{A}\underline{x} = \lambda\mathbf{A}^T\underline{x} = \lambda\mathbf{A}\underline{x} = \lambda^2\underline{x} \,.$$

לכן, הערכים הסינגולאריים של **A** במקרה זה יקיימו $\sigma_k\{\mathbf{A}\} = |\lambda_k\{\mathbf{A}\}|$.

למה $\sigma_1, \sigma_2, \ldots, \sigma_n$ נקראים ערכים סינגולאריים? כיוון שבכוחם לגלות האם המטריצה **A** סינגולארית (אם היא ריבועית) ואת דרגתה במקרה הכללי. למשל, עבור מטריצה ריבועית **A**, קיים הקשר הבא:

$$\det\{\mathbf{A}\} = \det\{\mathbf{V}\Sigma\mathbf{U}^T\} = \det\{\mathbf{V}\}\det\{\Sigma\}\det\{\mathbf{U}^T\} = \pm\det\{\Sigma\} = \pm\prod_{k=1}^{n}\sigma_k$$

הספק באשר לסימן נובע מכך שלמטריצה אורתונורמלית הדטרמיננטה ידועה להיות בערך מוחלט 1, אך סימנה לא ידוע. לכן, די בערך סינגולארי מאופס על מנת לקבוע את היות המטריצה סינגולארית.

7. שימוש ב- SVD למערכות של משוואות ול-LS

נניח שלפנינו מערכת משוואות ליניארית המשתמשת במטריצה **A** ולזו ידוע לנו פירוק ה-SVD שלה. לכן,

$$\mathbf{A}\underline{x} = \mathbf{V}\Sigma\mathbf{U}^T\underline{x} = \underline{b} \,.$$

נגדיר

$$\begin{aligned} \underline{y} = \mathbf{U}^T\underline{x} \quad &\Rightarrow \quad \mathbf{U}\underline{y} = \underline{x} \\ \underline{c} = \mathbf{V}^T\underline{b} \quad &\Rightarrow \quad \mathbf{V}\underline{c} = \underline{b} \end{aligned} \,,$$

בשני המקרים מדובר בפעולה אורתונורמלית שלא מאבדת מהוקטורים הללו תוכן, ורק מסובבת אותם במרחבם (n ו- m מימדי בהתאמה). מתקבלת מערכת משוואות אלטרנטיבית קלה לאין-ערוך,

$$\Sigma\underline{y} = \underline{c} \,.$$

אם **A** ריבועית, קל לראות מיד כי פתרון למערכת זו קיים רק עבור המקרה בו כל הערכים הסינגולאריים חיוביים ממש. במקרה כזה חישוב $\underline{y}$ קל, וממנו $\underline{x}$ יתקבל ע"י $\mathbf{U}\underline{y} = \underline{x}$. למעשה, במקרה כזה השגנו היפוך של המטריצה לפי הנוסחה

$$\mathbf{A}\underline{x} = \mathbf{V}\Sigma\mathbf{U}^T\underline{x} = \underline{b} \quad \Rightarrow \quad \mathbf{A}^{-1}\underline{b} = \left(\mathbf{V}\Sigma\mathbf{U}^T\right)^{-1}\underline{b} = \mathbf{U}\cdot\Sigma^{-1}\cdot\mathbf{V}^T\underline{b} = \mathbf{U}\cdot\Sigma^{-1}\cdot\underline{c} \,.$$

כאשר המטריצה מלבנית עם m<n, הערכים הסינגולאריים יאמרו לנו כמה משוואות פעילות יש בפועל, לפי השורות שלא מאופסות במטריצה Σ. כיוון שהמערכת $\Sigma\underline{y} = \underline{c}$ שקולה לחלוטין למערכת המשוואות המקורית $\mathbf{A}\underline{x} = \underline{b}$, מתקבל כי מספר הערכים הסינגולאריים השונים מאפס קובע את מספר השורות הבלתי תלויות ליניארית ב-**A**, ולכן את דרגתה של המטריצה.

כאשר המטריצה מלבנית עם m>n, שוב מספר הערכים הסינגולאריים השונים מאפס יספרו לנו את מספר המשוואות הפעילות – דרגת המטריצה. מעבר לכך, בהנחה כי יש n ערכים סינגולאריים שונים מאפס, אם איברי $\underline{c}$ מול שורות האפסים אינם מאופסים, ניכר מיד כי לא קיים פתרון למערכת המשוואות הנתונה. אם אלה אפסים, הפתרון מתקבל בצורה פשוטה.

נסכם זאת במשפט הבא:

משפט 10: דרגת המטריצה **A** הינה מספר הערכים הסינגולאריים שלה השונים מאפס.

נעבור כעת לפתרונן של בעיות ריבועים פחותים. נניח כמקודם כי בידינו מטריצה **A** ובעזרתה הוגדרה בעיית מינימיזציה

$$\min_{\underline{x}} \left\|\mathbf{A}\underline{x} - \underline{b}\right\|_2^2 = \min_{\underline{x}} \left\|\mathbf{V}\Sigma\mathbf{U}^T\underline{x} - \underline{b}\right\|_2^2 .$$

כמקודם, נגדיר

$$\begin{aligned} \underline{y} = \mathbf{U}^T\underline{x} \quad &\Rightarrow \quad \mathbf{U}\underline{y} = \underline{x} \\ \underline{c} = \mathbf{V}^T\underline{b} \quad &\Rightarrow \quad \mathbf{V}\underline{c} = \underline{b} \end{aligned} ,$$

ונקבל

$$\min_{\underline{x}} \left\|\mathbf{A}\underline{x} - \underline{b}\right\|_2^2 = \min_{\underline{x}} \left\|\mathbf{V}\left(\Sigma\mathbf{U}^T\underline{x} - \mathbf{V}^T\underline{b}\right)\right\|_2^2 = \min_{\underline{y}} \left\|\Sigma\underline{y} - \underline{c}\right\|_2^2 .$$

נתייחס תחילה למקרה המעניין בו המטריצה **A** הינה מטריצה בדרגה מלאה עם יותר שורות מעמודות. במקרה כזה נוכל לפרק את פונקצית המחיר לשני חלקים (דומה לנעשה בעזרת QR) ולקבל בעיה מהצורה

$$\min_{\underline{y}} \left\|\Sigma\underline{y} - \underline{c}\right\|_2^2 = \min_{\underline{y}} \left\|\Sigma_1\underline{y} - \underline{c}_1\right\|_2^2 + \left\|\underline{c}_2\right\|_2^2 .$$

המטריצה Σ_1 הינה מטריצה ריבועית ואלכסונית בה מצויים כל הערכים הסינגולאריים (השונים מאפס) על האלכסון. מול מטריצה זו עומד חלקו העליון של הווקטור $\underline{c}$. חלקו התחתון של הווקטור $\underline{c}$ באורך m-n לא ניתן לאיפוס ועוצמתו מגדירה את גובה המינימום של הפונקציה. לכן, במקרה כזה הפתרון יהיה

$$\underline{y}^* = \Sigma_1^{-1}\underline{c}_1 = [\Sigma_1^{-1} \quad 0]\begin{bmatrix}\underline{c}_1\\ \underline{c}_2\end{bmatrix} = [\Sigma_1^{-1} \quad 0]\mathbf{V}^T\underline{b}\ .$$

הפתרון של הבעיה המקורית נתון אם כך ע"י

$$\underline{x}^* = \mathbf{U}\underline{y}^* = \mathbf{U}[\Sigma_1^{-1} \quad 0]\mathbf{V}^T\underline{b}$$

ונוסחא זו מגדירה "היפוך מוכלל" של המטריצה **A**. לו הייתה המטריצה Σ ריבועית, היינו משתמשים בהיפוך רגיל בו כל איבר אלכסון מוחלף בהיפוכו. אך כשמטריצה זו מלבנית בגודל m-על-n (ולא חשוב מה היחס בין מספר עמודותיה לשורותיה), היפוך מוכלל שלה מוגדר ע"י מטריצה בגודל n-על-m בה איברי האלכסון השונים מאפס מוחלפים בהיפוכם, והיתר אפס. אנו נסמן פעולה זו כ- Σ^+ ונקראה לה Pseudo-Inverse:

$$\begin{bmatrix}\sigma_1 & 0 & 0 & 0 & 0 & 0\\ 0 & \sigma_2 & 0 & 0 & 0 & 0\\ 0 & 0 & \sigma_3 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix} = \begin{bmatrix}\sigma_1 & 0 & 0 & 0\\ 0 & \sigma_2 & 0 & 0\\ 0 & 0 & \sigma_3 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}^+ .$$

לכן, פתרון של בעיית ה- LS הנ"ל אינו אלא

$$\underline{x}^* = \mathbf{U}\underline{y}^* = \mathbf{U}\Sigma^+\mathbf{V}^T\underline{b} = \mathbf{A}^+\underline{b}\ .$$

אם **A** מטריצה מלבנית בגודל m-על-n עם פחות שורות מעמודות ועם דרגה מלאה (m), ברור לנו כי יש אינסוף פתרונות המביאים לאפס את המחיר $\|\mathbf{A}\underline{x} - \underline{b}\|_2^2$. ביטוי זה ניתן לרישום מהצורה

$$\min_{\underline{y}} \|\Sigma\underline{y} - \underline{c}\|_2^2 = \min_{\underline{y}} \|\Sigma_1\underline{y}_1 - \underline{c}\|_2^2,$$

כאשר Σ_1 מטריצה ריבועית ואלכסונית בגודל m-על-m עם הערכים הסינגולאריים השונים מאפס על האלכסון. מטריצה זו כופלת את m האיברים הראשונים ב- $\underline{y}$. את איבר זה נוכל לאפס ע"י הבחירה

$$\Sigma_1\underline{y}_1 = \underline{c} \quad\Rightarrow\quad \underline{y}_1 = \Sigma_1^{-1}\underline{c}\ .$$

באשר ליתר איברי $\underline{y}$, הם יכולים להיבחר באופן חופשי ללא השפעה על ביטוי ה-LS. לכן, הם מכילים בתוכם את דרגות החופש בבעיה. בחירתם כאפסים תיתן כי מצאנו מבין כל הפתרונות את זה שנותן את האורך המינימאלי לווקטור $\underline{y}$, אך כיוון ש- $\underline{x} = \mathbf{U}\underline{y}$ פתרנו בפועל את הבעיה החליפית

$$, \min_{\underline{x}|\ \mathbf{A}\underline{x}=\underline{b}} \|\underline{x}\|_2^2$$

עם אורך מינימאלי על הווקטור המקורי $\underline{x}$. פתרון זה קשור לרגולריזציה שהוצגה בהקשר ל-LS. גם הפעם ניתן לרשום פתרון זה ע"י ההיפוך המוכלל,

$$, \underline{x}^* = \mathbf{U}\Sigma^+\mathbf{V}^T\underline{b} = \mathbf{A}^+\underline{b}$$

כשהפעם המטריצה Σ^+ כוללת בחלקה העליון את הנתח השונה מאפס, ובתחתיתה איברים מאופסים.

לסיכום חלק זה, מצאנו כי פתרון LS נותן מוטיבציה להגדרת היפוך מוכלל של מטריצה בגודל כלשהו ולאו דווקא ריבועי, והיפוך זה פשוט ממיר את ההיפוך הנחוץ במטריצה Σ בהיפוך האיברים באלכסון השונים מאפס והשארת האחרים כאפסים.

מסתבר כי במקביל ללימוד עניין ה-SVD, אנשי אלגברה ליניארית חיפשו הכללה מוצלחת להיפוך מטריצה אשר יתלכד עם ההיפוך במקרה הריבועי הלא-סינגולארי, אך יוכל להתמודד עם מטריצות סינגולאריות ועם מטריצות מלבניות. בהינתן **A**, קבעו Moore ו-Penrose את התנאים הבאים שהיפוך מוכלל **X** צריך לקיים:

$$\begin{aligned}
\mathbf{AXA} = \mathbf{A} \quad &\Rightarrow \quad \mathbf{V}\Sigma\mathbf{U}^T\mathbf{U}\Sigma^+\mathbf{V}^T\mathbf{V}\Sigma\mathbf{U}^T = \mathbf{V}\Sigma\Sigma^+\Sigma\mathbf{U}^T = \mathbf{V}\Sigma\mathbf{U}^T \\
\mathbf{XAX} = \mathbf{X} \quad &\Rightarrow \quad \mathbf{U}\Sigma^+\mathbf{V}^T\mathbf{V}\Sigma\mathbf{U}^T\mathbf{U}\Sigma^+\mathbf{V}^T = \mathbf{U}\Sigma^+\Sigma\Sigma^+\mathbf{V}^T = \mathbf{U}\Sigma^+\mathbf{V}^T \\
(\mathbf{AX})^T = \mathbf{AX} \quad &\Rightarrow \quad \left(\mathbf{V}\Sigma\mathbf{U}^T\mathbf{U}\Sigma^+\mathbf{V}^T\right)^T = \mathbf{V}\Sigma\Sigma^+\mathbf{V}^T = \mathbf{V}\Sigma\mathbf{U}^T\mathbf{U}\Sigma^+\mathbf{V}^T \\
(\mathbf{XA})^T = \mathbf{XA} \quad &\Rightarrow \quad \left(\mathbf{U}\Sigma^+\mathbf{V}^T\mathbf{V}\Sigma\mathbf{U}^T\right)^T = \mathbf{U}\Sigma^+\Sigma\mathbf{U}^T = \mathbf{U}\Sigma^+\mathbf{V}^T\mathbf{V}\Sigma\mathbf{U}^T
\end{aligned}$$

וכפי שמומחש כאן,השיטה המבוססת על SVD עונה על כל הדרישות. מסתבר (עובדה זו מובאת ללא הוכחה) כי זוהי גם ההגדרה היחידה האפשרית אשר עונה לכל אותן אקסיומות. לכן גם לא אחת נקרא ההיפוך המוכלל Moore-Penrose Pseudo Inverse.

8. קירוב מטריצות עם אילוץ דרגה

הפירוק $\mathbf{A} = \mathbf{V}\Sigma\mathbf{U}^T$ ניתן גם לרישום אחר כ:

$$. \mathbf{A} = \mathbf{V}\Sigma\mathbf{U}^T = \sum_{k=1}^{n} \sigma_k \underline{v}_k \underline{u}_k^T$$

מטריצה מהצורה $\underline{a}\underline{b}^T$ הינה מטריצה בת דרגה 1, כיוון שכל שורותיה הם שכפולים של $\underline{b}$ עם ניפוח בסקלרים הלקוחים מ-$\underline{a}$. להן הרישום הנ"ל מציע את בניית המטריצה **A** כסכום של מטריצות בעלות דרגת יחידה, כשכל אחת מוכפלת במקדם – הערך הסינגולארי התואם.

במקרה הכללי, בהינתן מכפלה מהצורה

$$\mathbf{C}=\mathbf{AB}^T=\begin{bmatrix} | & | \\ \underline{a}_1 & \underline{a}_2 \\ | & | \end{bmatrix}\begin{bmatrix} - & \underline{b}_1^T & - \\ - & \underline{b}_2^T & - \end{bmatrix}=\underline{a}_1\underline{b}_1^T+\underline{a}_2\underline{b}_2^T ,$$

דרגת המטריצה **C** יכולה להיות 1 או 2. אם עמודות **A** בלתי תלויות ליניארית, וכך גם עמודות **B**, הדרגה תהיה מלאה – דהיינו – 2. תכונה זו נכונה גם למקבצים גדולים יותר, ולדוגמה, בפירוק ה-SVD שקיבלנו

$$\mathbf{A}=\mathbf{V}\Sigma\mathbf{U}^T=\sum_{k=1}^{n}\sigma_k\underline{v}_k\underline{u}_k^T$$

דרגת המטריצה **A** תהיה כמספר הערכים הסינגולאריים השונים מאפס (כפי שכבר טענו קודם) כיוון שכל מרכיב הינו בעל דרגת יחידה, ואלה בלתי תלויים.

נניח כעת כי נתונה לנו מטריצה **A** ועבורה ידוע פירוק ה-SVD הנ"ל. מהי המטריצה הקרובה ביותר ל-**A** אשר תהיה בעלת דרגה ידועה K_0? שאלה זו לא מוגדרת היטב עדיין, כיוון שלא הגדרנו עד כה מרחק בין שתי מטריצות. ממש כשם שיש מגוון דרכים להגדיר מרחק בין וקטורים, כך גם יש מגוון עצום המטפל במטריצות. נציע כאן מדד אחד מאוד פופולארי ומאוד אינטואיטיבי – נורמת Frobenius למטריצה **E** בגודל של m-על-n:

$$\|\mathbf{E}\|_F^2=\sum_{k=1}^{m}\sum_{j=1}^{n}e_{kj}^2 .$$

נורמה זו "חושבת" על המטריצה כווקטור ארוך ש-"התקפל" למטריצה, ולכן נורמה זו מתעלמת מהיות המטריצה פעולה ליניארית. בהמשך נכיר נורמות אחרות, אך לעת עתה זו תספיק לנו.

השאלה שלנו אם כך היא: בהינתן מטריצה **A** כלשהי בגודל של m-על-n, מהי המטריצה $\mathbf{A}_{K_0}$ באותו גודל הקרובה ביותר ל-**A** מחד, ובעלת דרגה מוגבלת K_0?

כלומר, אנו מעוניינים בפתרון בעיית האופטימיזציה

$$\underset{\mathbf{A}_{K_0}}{\arg\min}\left\|\mathbf{A}-\mathbf{A}_{K_0}\right\|_F^2 \quad \text{subject to } \operatorname{rank}\{\mathbf{A}_k\}=K_0 .$$

בזכות פירוק ה-SVD מוצע לנו פתרון סגור לבעיה זו, כפי שהמשפט הבא טוען:

משפט 11: עבור מטריצה A בעלת פירוק SVD מהצורה

$$\mathbf{A} = \mathbf{V}\Sigma\mathbf{U}^T = \sum_{k=1}^{n} \sigma_k \underline{v}_k \underline{u}_k^T$$

עם ערכים סינגולאריים בסדר יורד, פתרון הבעיה

$$\arg\min_{\mathbf{A}_{K_0}} \left\| \mathbf{A} - \mathbf{A}_{K_0} \right\|_F^2 \quad \text{subject to } \operatorname{rank}\{\mathbf{A}_k\} = K_0$$

הינו

$$\mathbf{A}_{K_0} = \mathbf{V}\Sigma\mathbf{U}^T = \sum_{k=1}^{K_0} \sigma_k \underline{v}_k \underline{u}_k^T .$$

והמרחק בין המטריצות נתון ע"י

$$\left\| \mathbf{A} - \mathbf{A}_{K_0} \right\|_F^2 = \sum_{k=K_0+1}^{n} \sigma_k^2 .$$

כלומר, כל שעלינו לעשות הוא לזרוק את "זנב" סידרת הערכים הסינגולאריים ולשמר את ה- K_0 הדומיננטיים. אנו נביא טענה זו ללא הוכחה בשלב זה.

נקודה מעניינת אשר עולה מהתוצאה הנ"ל היא שעבור מטריצה ריבועית בגודל n-על-n, ערכם של הערכים הסינגולאריים הקטנים שמצויים בתחתית הרשימה מגדיר את ריחוקה של המטריצה מהיותה סינגולארית. נניח כי בידינו מטריצה בגודל 10-על-10 עם הערכים הסינגולאריים:

$$\sigma = \{1,1,1,1,1,1,1e-15,1e-15,1e-15,1e-15\} .$$

לכאורה מטריצה זו בעלת דרגה מלאה, אך ברור כי דרגתה המעשית היא 6. זאת כיוון שהשמטת ארבעת הערכים הסינגולאריים האחרונים כמעט לא תיתן שוני במטריצה. בהקשר זה מקובל להשתמש ב-Condition Number, המסומן כ- $\kappa\{\mathbf{A}\} = \sigma_1 / \sigma_n$. גודל זה קובע כמה "חולנית" המטריצה וקרובה להיות סינגולארית. אם מספר זה גדול מידי ($\kappa\{\mathbf{A}\} \ge 1$ לפי הגדרה), נאמר כי המטריצה Ill-Conditions. אגב – ההגדרה חייבת לעשות שימוש ביחס בין ערכים סינגולאריים, כיוון שהתייחסות ל- σ_n חסרת משמעות – נוכל לכפול המטריצה במקדם גדול, ולכאורה לרפא המטריצה.

9. בעיית ה- TLS

אנו חוזרים כעת לבעיית התאמת עקומים, בה פגשנו בהקשר ל-LS. נניח כי נתון לנו אוסף צמדי נתונים $\{t_i, b_i\}_{i=1}^m$. צמדי נתונים אלו מתארים נקודות במישור. אנו מחפשים את הקו הישר הקרוב ביותר לכל הנקודות,

$$b = \alpha t + \beta .$$

בטיפול בעזרת LS הגדרנו את מערכת המשוואות

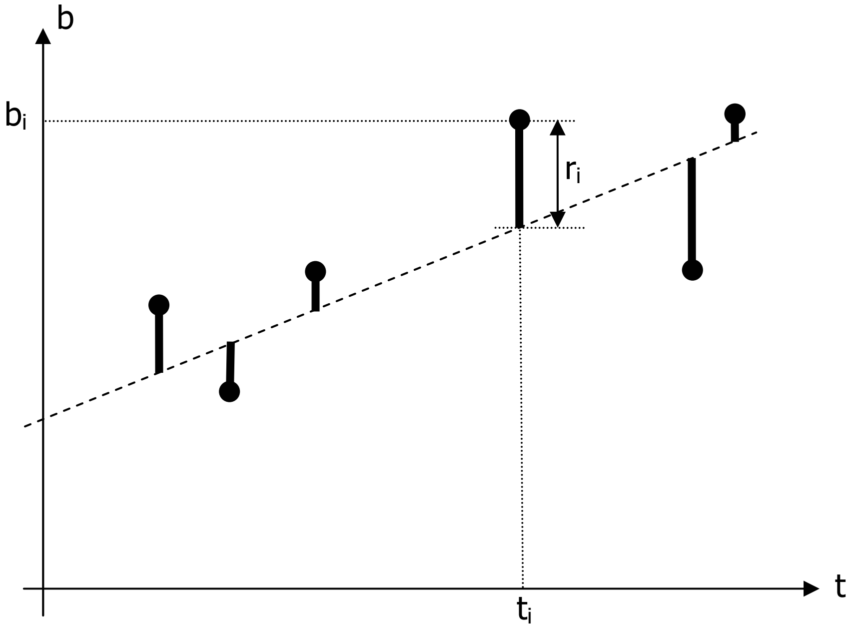
$$.\underline{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix} \approx \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \mathbf{A}\underline{x}$$

וחיפשנו פתרון α,β שייתן סטייה מינימאלית בין שני אגפי המשוואה. כלומר, עבור הנקודה $[t_i,b_i]$, הגובה הרצוי הוא b_i, והגובה המתקבל הוא $\alpha t_i + \beta = b_i + r_i$. שיטת ה-LS מביאה למינימום את סכום ריבועי הסטייה

$$.p(\underline{x}) = (\mathbf{A}\underline{x} - \underline{b})^T (\mathbf{A}\underline{x} - \underline{b}) = \|\mathbf{A}\underline{x} - \underline{b}\|_2^2 = \sum_{k=1}^{m} r_k^2$$

הציור הבא (שכבר ראינו) ממחיש סטיות אלה. ופתרונה, כפי שכבר ראינו, נתון ע"י

$$.\underline{x}^* = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\underline{b} = \begin{bmatrix} m & \sum_{i=1}^{m} t_i \\ \sum_{i=1}^{m} t_i & \sum_{i=1}^{m} t_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{m} b_i \\ \sum_{i=1}^{m} t_i b_i \end{bmatrix}$$



עובדה מטרידה בסיפור הנ"ל היא שמבלי שהצהרנו על כך, שיטת ה-LS קבעה שבאוסף הנתונים $\{t_i,b_i\}_{i=1}^{m}$, ערכי ה-t_i מדויקים, וסטיות, אם יחולו, ייוחסו ל-b_i. נניח שנהפוך את היוצרות ונפתור עבור הישר $t = \gamma b + \varphi$, ה-LS יוביל ל:

$$.\underline{t} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_m \end{bmatrix} \approx \begin{bmatrix} 1 & b_1 \\ 1 & b_2 \\ 1 & b_3 \\ \vdots & \vdots \\ 1 & b_m \end{bmatrix} \begin{bmatrix} \varphi \\ \gamma \end{bmatrix} = \mathbf{C}\underline{x}$$

והפתרון יהיה

$$.\underline{x}^* = \left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\underline{t} = \begin{bmatrix} m & \sum_{i=1}^{m} b_i \\ \sum_{i=1}^{m} b_i & \sum_{i=1}^{m} b_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{m} t_i \\ \sum_{i=1}^{m} t_i b_i \end{bmatrix}$$

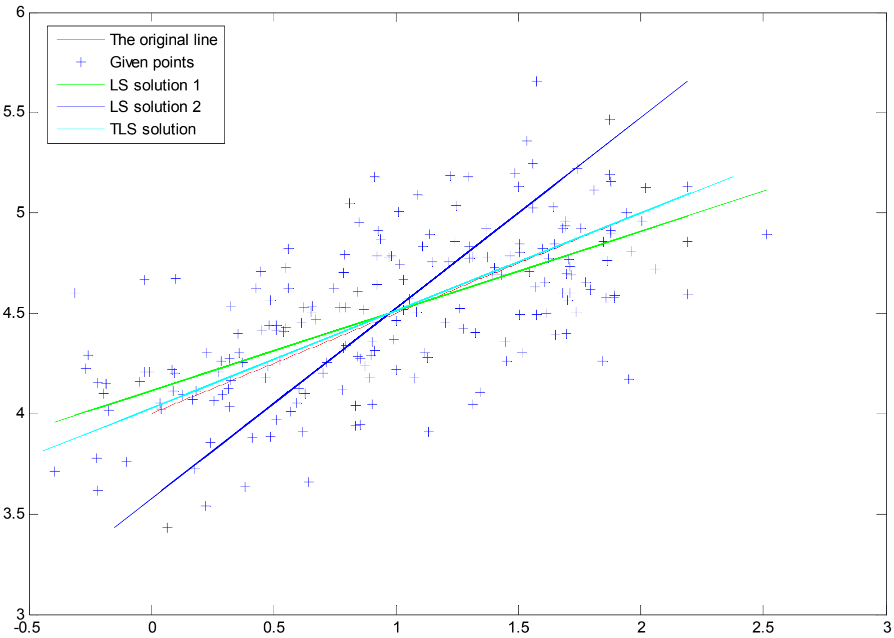
הישר המתקבל הקושר בין השניים יהיה אחר! זאת כיוון שהפעם פקפקנו בנכונותם של המדידות t_i, כשה- b_i מונחים כמדויקים.

לדוגמה, בניסוי הבא יצרנו סט של 21 צמדי מספרים היושבים על הישר $y = 0.5x + 4$, כשהם בפיזור אחיד בתחום $x = 0 : 0.1 : 2$ (שפת Matlab), ו-y תואמים. ערכים אלו (x ו-y) זוהמו ברעש אקראי, והוזנו לשתי השיטות הנ"ל לאיתור פרמטרי הישר. הגרף הבא ממחיש תוצאות אלה (יש להתעלם בשלב זה מהישר התכלכל) בהתבסס על התוכנית הבאה:

```
% Create the data
x=(0:0.01:2)'; y=0.5*x+4;
xn=x+randn(201,1)*0.3;
yn=y+randn(201,1)*0.3;
figure(1); clf; plot(x,y,'r'); hold on;
plot(xn,yn,'+');
% LS - version 1 - horizontal is fixed
A=[ones(201,1),xn]; b=yn;
param=inv(A'*A)*A'*b;
plot(xn,A*param,'g');
% LS - version 2 - vertical is fixed
C=[ones(201,1),yn]; t=xn;
param=inv(C'*C)*C'*t;
plot(C*param,yn,'b');
```

כאשר סט המדידות $\{t_i, b_i\}_{i=1}^{m}$ כולו מזוהם, לא נכון לעבוד לפי אחת מהשיטות הנ"ל, ויש לאמץ גישה שונה, בה מותרת סטייה בשתי עמודות הנתונים. אנו מניחים מודל ישר מעט אחר, לפיו

$$a\left(t_i - \bar{t}\right) = b_i - \bar{b}$$



כאשר $\bar{t},\bar{b}$ הם ממוצעי המדידות בכל משתנה, אותם ניתן לחשב בקלות,

$$\bar{t}=\frac{1}{m}\sum_{k=1}^{m}t_i \quad \text{and} \quad \bar{b}=\frac{1}{m}\sum_{k=1}^{m}b_i .$$

כעת ניתן לארגן את המשוואה שלפנינו ע"י

$$\mathbf{A}\underline{x}=\begin{bmatrix} t_1-\bar{t} & b_1-\bar{b} \\ t_2-\bar{t} & b_2-\bar{b} \\ t_3-\bar{t} & b_3-\bar{b} \\ \vdots & \vdots \\ t_m-\bar{t} & b_m-\bar{b} \end{bmatrix}\begin{bmatrix} a \\ -1 \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} .$$

כלומר, ידוע לנו כי המטריצה A שנבנתה צריכה הייתה להיות מטריצה בת דרגת יחידה, ואם היא אינה כזו, זה בשל הרעש שזיהם את הנתונים. אם רצוננו במטריצה אלטרנטיבית $\mathbf{A}_1$ בעלת דרגת יחידה אשר תהיה הכי קרובה ל-**A** הנתונה, פעולת ה-SVD מספקת זאת מן המוכן – כל שעלינו לעשות הוא פירוק SVD ל-**A**, איפוס הערך הסינגולארי השני, ואז הכפלה חזרה לבניית המטריצה המתקבלת $\mathbf{A}_1$. מטריצה זו תכיל בעמודותיה את ערכי ה-$\{t_i,b_i\}_{i=1}^{m}$ הנקיים, אשר יושבים על ישר. חשוב לזכור כי ערכים אלו מורכזו בשל הורדת הממוצע – זה הזמן להחזירו. גישה זו מוכרת בשם Total-Least-Squares. התוכנית הבאה ממחישה גישה זו, ותוצאותיה נתונות בגרף שראינו קודם.

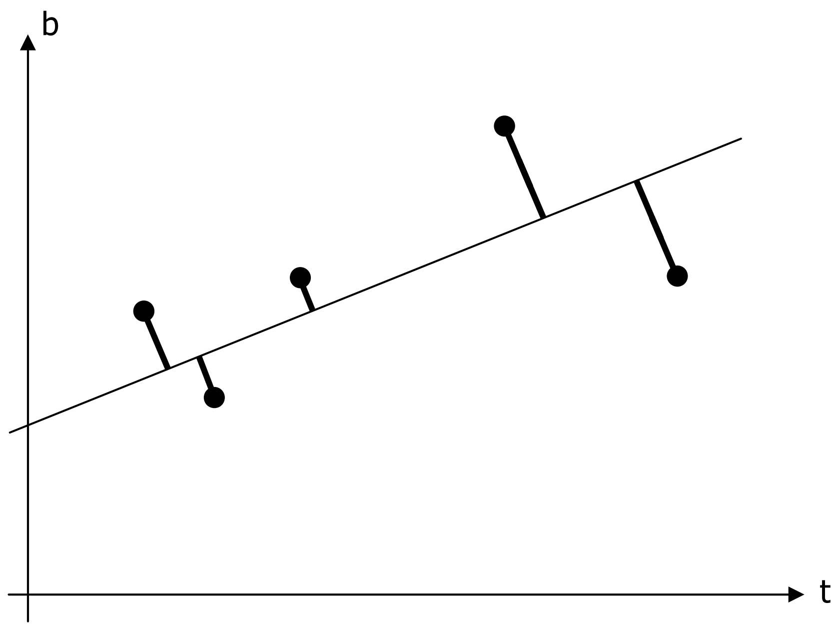
```
% TLS
```

```
xnM=mean(xn); ynM=mean(yn);
A=[xn-xnM,yn-ynM];
[U,D,V]=svd(A);
D(2,2)=0; Anew=U*D*V';
plot(Anew(:,1)+xnM,Anew(:,2)+ynM,'c');
legend({'The original line','Given points',…
        'LS solution 1','LS solution 2','TLS solution',});
```

כשמתבוננים במודל $\hat{b}_i = (b_i - \bar{b}) = a\hat{t}_i = a(t_i - \bar{t})$, ישנה משמעות גיאומטרית יפה לשיטת ה-TLS. במקרה כזה, פתרנו את בעיית האופטימיזציה הבאה:

$$\min_{a\underline{t}_0=\underline{b}_0} \left\| \begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix} - \begin{bmatrix} | & | \\ \underline{t}_0 & \underline{b}_0 \\ | & | \end{bmatrix} \right\|_F^2 = \min_{at_k^0=b_k^0} \sum_{k=1}^{m} \left(\hat{t}_k - t_k^0\right)^2 + \left(\hat{b}_k - b_k^0\right)^2 .$$

כלומר, חיפשנו ישר עליו יושבות הנקודות החדשות, $at_k^0 = b_k^0$, ואת אלה אנו מחפשים כך שיהיו הקרובות ביותר לנקודות הנתונות. לכן, המרחקים המובאים למינימום בשיטת ה-TLS הם המרחקים הקצרים ביותר אל הישר, כפי שהציור הבא מראה.



אגב, נראה כי השמטת הממוצע מסט המדידות מהווה בתיאור הנ"ל בחירה שרירותית ולא מוצדקת. למעשה זו הגישה האופטימאלית, כפי שהניתוח האלטרנטיבי הבא מראה: נוכל להציע פרמטריזציה של הישר ע"י $\alpha t_i + \beta b_i + \gamma = 0$. ברור כי שלושה פרמטרים הם יותר מידי ולכן נוכל לכפות על המקדמים את האילוץ $\alpha^2 + \beta^2 = 1$ (כי

הכפלה של המודל הנ"ל במספר שרירותי מותירה אותו נכון). כעת הבעיה שלנו מומרת לחיפוש השלשה α,β,γ. מה יהיה קריטריון האופטימאליות? כנגד כל נקודה נתונה מהסט $\{t_i,b_i\}_{i=1}^m$ נרצה למצוא לה את הקרובה אליה ביותר על הישר $\alpha t_i+\beta b_i+\gamma=0$ ולדרוש שהמרחק ביניהן יובא למינימום.

בהינתן נקודה t_i,b_i, מהי הנקודה הקרובה ביותר אליה t_0,b_0 על הישר $\alpha t_i+\beta b_i+\gamma=0$? התשובה נתונה ע"י בעיית האופטימיזציה

$$\min \frac{1}{2}\left((t_0-t_i)^2+(b_0-b_i)^2\right) \text{ s.t. } \alpha t_0+\beta b_0+\gamma=0$$

הפתרון נתון ע"י שיטת כופלי לגרנז':

$$L(t_0,b_0,\lambda)=\frac{1}{2}\left((t_0-t_i)^2+(b_0-b_i)^2\right)+\lambda(\alpha t_0+\beta b_0+\gamma)$$
$$\Rightarrow \frac{\partial L(t_0,b_0,\lambda)}{\partial t_0}=t_0-t_i+\lambda\alpha=0 \quad \text{and} \quad \frac{\partial L(t_0,b_0,\lambda)}{\partial b_0}=b_0-b_i+\lambda\beta=0$$

הצבת משוואות אלו באילוץ נותנת

$$0=\alpha t_0+\beta b_0+\gamma=\alpha(t_i-\lambda\alpha)+\beta(b_i-\lambda\beta)+\gamma \Rightarrow \lambda=\alpha t_i+\beta b_i+\gamma.$$

כאן השתמשנו בידיעה ש-$\alpha^2+\beta^2=1$. לסיכום, המרחק בין הישר לנקודה הנתונה הוא

$$(t_0-t_i)^2+(b_0-b_i)^2=\left(\alpha^2+\beta^2\right)(\alpha t_i+\beta b_i+\gamma)^2=(\alpha t_i+\beta b_i+\gamma)^2.$$

לכן, קריטריון האופטימאליות יהיה

$$\sum_{i=1}^m(\alpha t_i+\beta b_i+\gamma)^2=\left\|\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}+\gamma\underline{1}\right\|_2^2.$$

כעת ניפטר מ-γ ע"י גזירה ביחס אליו והשוואה לאפס. נקבל

$$\frac{\partial}{\partial\gamma}\left\|\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}+\gamma\underline{1}\right\|_2^2=\underline{1}^T\left(\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}+\gamma\underline{1}\right)=0 \Rightarrow \gamma=\frac{-1}{m}\underline{1}^T\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}.$$

הצבת ביטוי זה חזרה לקריטריון שלנו תיתן כי עלינו להביא למינימום את הביטוי

$$\left\|\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}+\gamma\underline{1}\right\|_2^2=\left\|\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}-\frac{1}{m}\underline{1}\underline{1}^T\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\right\|_2^2=\left\|\left(\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}-\frac{1}{m}\underline{1}\underline{1}^T\begin{bmatrix} | & | \\ \underline{\hat{t}} & \underline{\hat{b}} \\ | & | \end{bmatrix}\right)\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\right\|_2^2$$

המטריצה הריבועית $\frac{1}{m}\underline{1}\underline{1}^T$ יוצרת שתי עמודות חדשות בהן הממוצעים של שתי העמודות המקוריות של הנתונים, וכך הגענו בדיוק לאותה בעיה.

באופן כללי יותר נאמר את המשפט הבא:

משפט 12: בהינתן מערכת משוואות מקורבת $\mathbf{A}\underline{x} \approx \underline{b}$ עם **A** בגודל m-על-n ($m \geq n$) בה הסטייה משוויון נובעת משגיאות גם ב- **A** וגם ב-$\underline{b}$, שיטת ה-TLS מציעה את הפתרון הבא למציאת $\underline{x}$:

- בנה את המטריצה המוגדלת $[\mathbf{A},-\underline{b}]$,
- בצע פירוק SVD למטריצה זו, $[\mathbf{A},-\underline{b}] = \mathbf{UDV}^T$,
- אפס את הערך הסינגולארי σ_{n+1} לקבלת $\mathbf{D}_1$,
- המטריצה $\mathbf{UD}_1\mathbf{V}^T$ מהווה גרסה נקייה של $[\mathbf{A},-\underline{b}]$ עבורה קיים פתרון,
- דרך אחרת לקבלת הפתרון: שימוש ב-n האיברים הראשונים ב-$\underline{v}_{n+1}$ וחלוקה באיבר האחרון בווקטור זה.

פתרון זה אופטימלי במובן קירבת Frobenius בין המדידות בפועל ב-$[\mathbf{A},\underline{b}]$ ובין גרסתם הנקייה המביאה לשוויון.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 6 - שיטות איטרטיביות

נושאים שייסקרו:

1. תהליכים איטרטיביים
2. יציבות בתהליכים איטרטיביים
3. נורמות של מטריצות ויציבות
4. פתרון איטרטיבי של מערכות משוואות ליניאריות
5. פתרון איטרטיבי של בעיות אופטימיזציה ריבועיות
6. פתרון איטרטיבי של בעיות ערכים עצמיים

1. מערכות איטרטיביות

בפרק זה נדון בתהליך ליניארי סדרתי מהצורה

$$\underline{u}_{k+1} = \mathbf{T}\underline{u}_k$$

אם אתחול מסוים $\underline{u}_0$. ברור כי קיים הקשר

$$\underline{u}_k = \mathbf{T}\underline{u}_{k-1} = \mathbf{T}\cdot\mathbf{T}\underline{u}_{k-2} = \cdots = \mathbf{T}^k\underline{u}_0 .$$

השאלה הטבעית לשאול היא למה זה בכלל מעניין? מסתבר כי לתהליך מהסוג הנ"ל יש חשיבות מרכזית בפתרון איטרטיבי של מערכות משוואות, מציאת ערכים עצמיים, ופתרון בעיות אופטימיזציה. לכן נקדיש זמן תחילה להתנהגות העקרונית של תהליך כזה, ואז נתאים אותו לצרכים הנ"ל.

אם נניח כי המטריצה $\mathbf{T}$ ניתנת ללכסון, פירוש הדבר שיש לה n ערכים עצמיים λ_j ו-n וקטורים עצמיים תואמים להם $\underline{v}_j$ אשר הינם בלתי תלויים ליניארית. מתקבל

$$\mathbf{T} = \mathbf{V}\Lambda\mathbf{V}^{-1} \quad \Rightarrow \quad \begin{aligned}\underline{u}_k &= \mathbf{T}^k\underline{u}_0 = \left(\mathbf{V}\Lambda\mathbf{V}^{-1}\right)\left(\mathbf{V}\Lambda\mathbf{V}^{-1}\right)\cdots\left(\mathbf{V}\Lambda\mathbf{V}^{-1}\right)\underline{u}_0 = \\ &= \mathbf{V}\Lambda^k\mathbf{V}^{-1}\underline{u}_0\end{aligned}$$

אם נסמן $\mathbf{V}\underline{x}_0 = \underline{u}_0$, הרי שהווקטור $\underline{x}_0$ מתאר כיצד לבנות בצירוף ליניארי של הוקטורים העצמיים את וקטור האתחול. לכן,

$$\underline{u}_k = \mathbf{V}\Lambda^k\underline{x}_0 = \sum_{j=1}^{n} x_0(j)\lambda_j^k\underline{v}_j .$$

כלומר, קיבלנו כי בעזרת הערכים והוקטורים העצמיים של המטריצה $\mathbf{T}$ ניתן לתאר בצורה פשוטה למדי את התנהלות הסדרה $\{\underline{u}_k\}_k$. בשלב ה-k יהיה המוצא של תהליך זה צירוף ליניארי של הוקטורים העצמיים (זה לא אומר כלום - כל וקטור ניתן לתיאור כצירוף כזה), עם מקדמים שהינם חזקות מתקדמות של הערכים העצמיים.

כדוגמה, נניח כי

$$\mathbf{T} = \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix} \qquad \underline{u}_0 = \begin{bmatrix} a \\ b \end{bmatrix}$$

תחילה נמצא את הערכים והוקטורים העצמיים. אלה נתונים ע"י

$$\lambda_1 = 0.4 \quad \underline{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \lambda_2 = 0.2 \quad \underline{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

כעת נמצא את $\underline{x}_0$:

$$\mathbf{V}\underline{x}_0 = \underline{u}_0 = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_0(1) \\ x_0(2) \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \begin{bmatrix} x_0(1) \\ x_0(2) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} a+b \\ b-a \end{bmatrix}$$

ולכן,

$$\underline{u}_k = \sum_{j=1}^{n} x_0(j)\lambda_j^k \underline{v}_j = \frac{1}{2}(a+b)\cdot 0.4^k \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{1}{2}(b-a)\cdot 0.2^k \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} =$$

$$= \frac{1}{2}\begin{bmatrix} (a+b)\cdot 0.4^k - (b-a)\cdot 0.2^k \\ (a+b)\cdot 0.4^k + (b-a)\cdot 0.2^k \end{bmatrix}.$$

אנו רואים כי כש- k מתקדם, ערכי המוצא קטנים והולכים ומתכנסים לאפס, בשל היות הערכים העצמיים קטנים מ-1 בערכם המוחלט. לאתחול אין השפעה על התנהגות זו.

דוגמה נוספת קשורה לסדרת פיבונצ'י,

$$u^{(k+2)} = u^{(k+1)} + u^{(k)}, \qquad u^{(1)} = u^{(0)} = 1$$

לכן הסדרה היא 1, 1, 2, 3, 5, 8, 13, וכך הלאה. מסתבר שלסדרה זו חשיבות גדולה במדע, ותופעות רבות ניתנות לתיאור בעזרתה (פיבונצ'י עצמו קיבל אותה בניסיון לאמוד את כמות זוגות הארנבות כפונקציה של מחזור הרבייה שלהן). ע"י ההגדרה

$$\underline{u}_k = \begin{bmatrix} u^{(k)} \\ u^{(k+1)} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} u^{(k-1)} \\ u^{(k)} \end{bmatrix}$$

אנו מקבלים כי לפנינו תהליך המאופיין ע"י

$$\mathbf{T} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \qquad \underline{u}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

הערכים והוקטורים העצמיים במקרה זה הם

$$\lambda_1 = 1.618 \quad \underline{v}_1 = \begin{bmatrix} 0.618 \\ 1 \end{bmatrix} \quad \text{and} \quad \lambda_2 = -0.618 \quad \underline{v}_2 = \begin{bmatrix} -1.618 \\ 1 \end{bmatrix}.$$

לעניין האתחול נקבל

$$\mathbf{V}\underline{x}_0 = \underline{u}_0 = \begin{bmatrix} 0.618 & -1.618 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_0(1) \\ x_0(2) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} x_0(1) \\ x_0(2) \end{bmatrix} = \begin{bmatrix} 1.1708 \\ -0.1708 \end{bmatrix},$$

ואילו מוצא התהליך בשלב ה-k יהיה

$$\underline{u}_k = \sum_{j=1}^{n} x_0(j)\lambda_j^k \underline{v}_j = 1.1708 \cdot 1.618^k \cdot \begin{bmatrix} 0.618 \\ 1 \end{bmatrix} + 0.1708 \cdot (-0.618)^k \cdot \begin{bmatrix} -1.618 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 1.1708 \cdot 0.618 \cdot 1.618^k - 1.618 \cdot 0.1708 \cdot (-0.618)^k \\ 1.1708 \cdot 1.618^k + 0.1708 \cdot (-0.618)^k \end{bmatrix},$$

כנקודה מעניינת, קיבלנו נוסחא סגורה לחישובו של האיבר ה-k בסדרת פיבונצ'י ללא חישוב רקורסיבי. כמו כן, בדוגמה זו ניכר כי היות ערך עצמי אחד מעל 1 בערך מוחלט גורמת לוקטורי המוצא לגדול ללא גבול.

נשים לב לכך שכל הדיון הנ"ל התייחס למטריצות לכסינות. ניתוח המקרה הכללי יותר מורכב יותר, אך יכול להישען על כלים שיוצגו בהמשך (נורמות של מטריצות).

2. יציבות של תהליכים איטרטיביים

מסיבות שונות שיובהרו בהמשך, יש לנו עניין מיוחד במצבים בהם $\underline{u}_k \to \underline{0}$ (ההתכנסות היא בכל איבר). מצב כזה ייקרא "מצב יציב", ותהליך המביא אליו ייחשב כ-"תהליך אסימפטוטית יציב". ברור שיציבות תלויה בהתנהגות המטריצה $\mathbf{T}^k$. מטריצה $\mathbf{T}$ תיקרא יציבה (או מתכנסת) אם $\mathbf{T}^k \to \mathbf{0}$. התכנסות זו פירושה שכל אחד מהאיברים במטריצה המתקבלת שואף לאפס. מהדוגמאות למעלה וההיגיון העוטף אותן ניכרת התכונה הבאה:

משפט 1: תהליך ליניארי מבוסס $\mathbf{T}$ לכסינה יהיה אסימפטוטית יציב אם ערכיה העצמיים של המטריצה $\mathbf{T}$ מקיימים

$$|\lambda_k| < 1 \quad \forall k \in [1,n].$$

בהינתן מטריצה $\mathbf{T}$ מגדירים את הרדיוס הספקטראלי שלה כ:

$$\rho\{\mathbf{T}\} = \max_{1 \le k \le n} |\lambda_k|$$

וברור אם כן כי יציבות תושג כשהרדיוס הספקטראלי מתחת 1, כלומר שכל הערכים העצמיים בתוך מעגל היחידה (מותר להם להיות מרוכבים!). הוכחת משפט זה טריוויאלית – נרשום את הפתרון כפי שכבר קיבלנו לתהליך זה:

$$\underline{u}_k = \mathbf{V}\Lambda^k \underline{x}_0 = \sum_{j=1}^{n} x_0(j)\lambda_j^k \underline{v}_j$$

הפעלת נורמה על שני האגפים ושימוש באי-שוויונות פשוטים תביא ל:

$$\left\|\underline{u}_k\right\|_2^2=\left\|\sum_{j=1}^{n}x_0(j)\lambda_j^k\underline{v}_j\right\|_2^2\le\sum_{j=1}^{n}\left|x_0(j)\lambda_j^k\right|^2\cdot\left\|\underline{v}_j\right\|_2^2=$$

$$=\sum_{j=1}^{n}\left|x_0(j)\right|^2\cdot\left|\lambda_j\right|^{2k}\le\rho\{\mathbf{T}\}^{2k}\cdot\sum_{j=1}^{n}\left|x_0(j)\right|^2\xrightarrow[k\to\infty]{}0$$

חשוב להבהיר כי אם חלק מהערכים העצמיים קטנים מ-1 בערכם המוחלט, השפעתם על המוצא של התהליך הולכת ודועכת, אך אלה לא יביאו לתיקון היציבות, והערכים הגדולים מ-1 ישלטו בסופו של דבר בהתנהלות המוצא. בהקצנה נאמר כי אם הערכים מסודרים לפי ערכם המוחלט כש- $\rho\{\mathbf{T}\}=|\lambda_1|$, אזי מוצא התהליך עבור k גדול יהיה

$$\underline{u}_k=\sum_{j=1}^{n}x_0(j)\lambda_j^k\underline{v}_j\approx x_0(1)\lambda_1^k\underline{v}_1 .$$

כמו כן, ממשוואה זו נראה כי עבור אתחול בו $x_0(1)=0$, מרכיב זה יאופס זהותית. הדבר נכון תיאורטית, אך מבחינה מעשית, די בשגיאות העגלה של המחשב על מנת לעורר מרכיב זה לחיים, ובסדרה של צעדים יגדל זה וייקח את הבכורה.

מהניתוח הנ"ל קיבלנו גם מידע נוסף – הרדיוס הספקטראלי לא רק מוצא האם תחול התכנסות אלא גם מנבא באופן לא רע את קצבה. ההתכנסות תתנהג כמו הטור הגיאומטרי

$$\left\|\underline{u}_k\right\|_2^2\propto\rho\{\mathbf{T}\}^{2k},$$

ולכן, אם יש לנו שליטה בקביעת רדיוס זה, ננסה לעשותו קטן ככל האפשר. אגב, קצב התכנסות זה קרוי קצב אקספוננציאלי, אם כי לעיתים גם משתמשים בשם קצב ליניארי כדי לומר כי בגרף לוגריתמי העקום המתאר את הנורמה הריבועית של $\underline{u}_k$ הוא ישר.

3. נורמות של מטריצות ויציבות

בשלב הזה ברור לנו שבהינתן תהליך איטרטיבי, אם אנו רוצים לראות אותו מתכנס לאפס (ונרצה!!), כדאי שנבדוק את הרדיוס הספקטראלי שלו. אלא שחישוב זה קשה למדי ומעיק. האם יש דרכי קיצור? האם ישנן בדיקות פשוטות יותר שיעידו בדרך עקיפה על יציבות\אי-יציבות? מסתבר שהתשובה היא חיובית, וזו קשורה לנורמות של מטריצות.

נורמות של וקטורים מוכרים לנו היטב. אלה צריכות לקיים את הדרישות הבאות:

- חיוביות: $\|\underline{v}\|\ge 0$,$\forall\underline{v}$. אם $\|\underline{v}\|=0$ אזי $\underline{v}=0$.

- הומוגניות: $\|c\underline{v}\| = |c| \cdot \|\underline{v}\|$ $\forall \underline{v}$ and c, .
- אי-שוויון המשולש: $\|\underline{v}+\underline{w}\| \le \|\underline{v}\| + \|\underline{w}\|$ $\forall \underline{v}$ and $\underline{w}$, .

בפרט אנו מכירים את:

- נורמת ℓ^2: $\|\underline{v}\|_2 = \sqrt{\sum_{k=1}^{n} v_k^2}$,
- נורמת ℓ^p: $\|\underline{v}\|_p = \sqrt[p]{\sum_{k=1}^{n} v_k^p}$. עבור p=1 מתקבלת נורמה הסוכמת את הערכים המוחלטים - $\|\underline{v}\|_1 = \sum_{k=1}^{n} |v_k|$. עבור p שואף לאינסוף מתקבלת נורמת ה-max - $\|\underline{v}\|_\infty = \max_k |v_k|$.
- נורמת ℓ^2 ממושקלת: $\|\underline{v}\|_{\mathbf{W}} = \sqrt{\underline{v}^T \mathbf{W} \underline{v}}$ עם **W** חיובית מוגדרת.

ישנה תכונה של שקילות בין כל הנורמות האפשריות (נכון למימדים סופיים, אך זהו המקרה הנדון כאן)! שקילות זו אומרת שעבור שתי נורמות **כלשהן** שונות, קיימים שני קבועים כך ש:

$$\forall \underline{v},\ c_1 \cdot \|\underline{v}\|_A \le \|\underline{v}\|_B \le c_2 \cdot \|\underline{v}\|_A .$$

פירוש הדבר שנורמות שונות מוגבלות בפתיחת פערים ביניהן. אגב, בין היתר זה אומר שאם ישנה התכנסות ביחס לנורמה אחת, ההתכנסות תקפה לכל נורמה אחרת.

כבר קודם התלבטנו בשאלה כיצד לתת למטריצה נורמה. דרך אחת היא להתעלם מעובדת היותה מטריצה, ולחשוב עליה כעל וקטור מקופל. דרך זו הובילה למשל לנורמת Frobenius כהמשך טבעי לנורמת ℓ^2 המיוחסת לווקטורים. כאן נעשה דברים אחרת, מתוך מטרה להיות מסוגלים לומר דברים על יציבות התהליך האיטרטיבי.

נניח כי בידינו נורמה וקטורית כלשהי המסומנת $\|\underline{v}\|_A$. נגדיר בעזרתה נורמה מטריצית למטריצה ריבועית בגודל n-על-n לפי

$$\|\mathbf{T}\|_A = \max_{\underline{v} |\ \|\underline{v}\|_A \le 1} \|\mathbf{T}\underline{v}\|_A = \max_{\underline{v}} \frac{\|\mathbf{T}\underline{v}\|_A}{\|\underline{v}\|_A} .$$

נורמה זו נקראת נורמה אופרטורית, כיוון שהיא בוחנת כיצד המטריצה פועלת כהתמרה ליניארית, ומהו ה-"ניפוח" הגדול ביותר לו היא מסוגלת להגיע על וקטורי יחידה.

ניתן להראות (ואנו לא נעשה זאת) כי נורמה זו מקיימת את האקסיומות שהוזכרו קודם. יתרה מזו, נורמה אופרטורית זו מקיימת תכונות נוספות שנורמות "סתמיות" לא היו מקיימות:

- פירוק מכפלה 1 - $\|\mathbf{T}\underline{v}\|_A \le \|\mathbf{T}\|_A \cdot \|\underline{v}\|_A$ - נובע ישירות מההגדרה ע"י:

$$\|\mathbf{T}\underline{v}\|_A = \frac{\|\mathbf{T}\underline{v}\|_A}{\|\underline{v}\|_A}\|\underline{v}\|_A \le \left[\max_{\underline{v}} \frac{\|\mathbf{T}\underline{v}\|_A}{\|\underline{v}\|_A}\right] \cdot \|\underline{v}\|_A = \|\mathbf{T}\|_A \|\underline{v}\|_A .$$

- פירוק מכפלה 2 - $\|\mathbf{T}_1\mathbf{T}_2\|_A \le \|\mathbf{T}_1\|_A \cdot \|\mathbf{T}_2\|_A$ - נובע מהתכונה הקודמת, ע"י:

$$\|\mathbf{T}_1\mathbf{T}_2\|_A = \max_{\underline{v}} \frac{\|\mathbf{T}_1\mathbf{T}_2\underline{v}\|_A}{\|\underline{v}\|_A} \le \max_{\underline{v}} \frac{\|\mathbf{T}_1\|_A \|\mathbf{T}_2\underline{v}\|_A}{\|\underline{v}\|_A} = \|\mathbf{T}_1\|_A \|\mathbf{T}_2\|_A .$$

- פירוק מכפלה 3 - $\left\|\mathbf{T}^k\right\|_A \le \|\mathbf{T}\|_A^k$ - נובע מהתכונה הקודמת ישירות, וכמו-כן

- יחס לרדיוס ספקטראלי - $\rho\{\mathbf{T}\} \le \|\mathbf{T}\|_A$ - נובע ישירות מההגדרה, ע"י הצבת הווקטור העצמי, $\underline{x}_1$, הקשור לערך העצמי הגדול ביותר בערך מוחלט, λ_1:

$$\|\mathbf{T}\|_A = \max_{\underline{v}} \frac{\|\mathbf{T}\underline{v}\|_A}{\|\underline{v}\|_A} \ge \frac{\|\mathbf{T}\underline{x}_1\|_A}{\|\underline{x}_1\|_A} = \frac{\|\lambda_1 \underline{x}_1\|_A}{\|\underline{x}_1\|_A} = |\lambda_1| .$$

התכונה האחרונה קובעת שכל נורמה אופרטורית תוביל לערך גדול מהרדיוס הספקטראלי. מכאן עולה כי אם חישבנו למטריצה נורמה אופרטורית כלשהי וקיבלנו ערך קטן מ-1, ברור שהרדיוס הספקטראלי קטן מ-1, ולכן התהליך האיטרטיבי יהיה יציב. למעשה, יציבות זו נובעת גם ישירות מהתכונה השלישית בה אם למטריצה יש נורמה אופרטורית קטנה מ-1, ברור כי $\left\|\mathbf{T}^k\right\|_A$ מתכנס למטריצת אפסים, וזו הרי יציבות.

ראינו כי כתחליף לחישוב של ערך עצמי בעל נורמה מרבית, נוכל להסתפק בחישוב נורמת המטריצה. האם זה קל יותר? מסתבר שכן בחלק מהמקרים. לדוגמה, עבור הנורמה הווקטורית $\|\underline{v}\|_\infty = \max_k |v_k|$, נבחן כיצד נראית הנורמה האופרטורית:

$$\|\mathbf{T}\|_\infty = \max_k \left\{ \left| \sum_{j=1}^{n} t_{kj} v_j \right| \right\} .$$

ביטוי זה מחשב כל איבר במכפלה של **T** עם וקטור באורך יחידה (בנורמה הנדונה), ומחפש את הגדול שבהם בערך מוחלט. הכנסת הערך המוחלט לתוך הסכימה גורמת להגדלת הביטוי,

$$\|\mathbf{T}\|_\infty = \max_k \left\{ \left| \sum_{j=1}^{n} t_{kj} v_j \right| \right\} \le \max_k \left\{ \sum_{j=1}^{n} |t_{kj}| \cdot |v_j| \right\}.$$

החלפת איברי הווקטור v בגדול שבהם תיתן

$$\|\mathbf{T}\|_\infty = \max_k \left\{ \left| \sum_{j=1}^{n} t_{kj} v_j \right| \right\} \le \max_k \left\{ \sum_{j=1}^{n} |t_{kj}| \cdot |v_j| \right\} \le \max_k \left\{ \sum_{j=1}^{n} |t_{kj}| \cdot \max_j |v_j| \right\}.$$

נשתמש בעובדה ש- $\|\underline{v}\|_\infty = \max_k |v_k| = 1$, ונקבל

$$\|\mathbf{T}\|_\infty \le \max_k \left\{ \sum_{j=1}^{n} |t_{kj}| \cdot \max_j |v_j| \right\} = \max_k \left\{ \sum_{j=1}^{n} |t_{kj}| \right\}.$$

בדרך בה פיתחנו את המעברים הללו, קיבלנו ביטוי קל לחישוב (סכום ערכי איברי כל שורה בערך מוחלט, ולאחר מכן חישוב המקסימום ביניהם), אך נראה כי הוא חסם עליון על הנורמה המבוקשת. מסתבר כי זהו חסם הדוק, כלומר, זהו הביטוי הנותן את הנורמה במדויק. קל לראות זאת בדרך הבאה: נניח כי השורה הראשונה הגשימה את הסכום המקסימאלי בביטוי האחרון. אזי, בחירת הווקטור $\underline{v}$ להיות בעל ערכים ± 1 בהתאם לסימני האיברים בשורה זו, תיתן כי כל מעברי האי-שוויון שנעשו קודם מתקיימים בשוויון.

באופן דומה נוכל לטפל בנורמה אופרטורית הנובעת מהנורמה הווקטורית $\|\underline{v}\|_1 = \sum_{k=1}^{n} |v_k|$. מתקבל:

$$\begin{aligned} \|\mathbf{T}\|_1 &= \sum_{k=1}^{n} \left| \sum_{j=1}^{n} t_{kj} v_j \right| \le \sum_{k=1}^{n} \sum_{j=1}^{n} |t_{kj}| \cdot |v_j| = \sum_{j=1}^{n} \left[\sum_{k=1}^{n} |t_{kj}| \right] \cdot |v_j| \le \\ &\le \sum_{j=1}^{n} \max_j \left[\sum_{k=1}^{n} |t_{kj}| \right] \cdot |v_j| = \max_j \left[\sum_{k=1}^{n} |t_{kj}| \right] \cdot \sum_{j=1}^{n} |v_j| = \max_j \left[\sum_{k=1}^{n} |t_{kj}| \right] \end{aligned}.$$

קיבלנו כי גם גישה לפיה נחשב סכום ערכים מוחלטים בעמודות (ולא בשורות כמקודם!!) וניקח את המקסימלי שבסכומים אלה נקבל את הערכת הנורמה המושרית ע"י $\|\underline{v}\|_1$. גם כאן זהו חסם הדוק וההערכה אינה רק חסם עליון אלא חישוב הנורמה במדויק.

מסתבר כי באופן דומה, שימוש בנורמת ℓ^2 מוביל לביטוי סגור - $\|\mathbf{T}\|_2 = \sigma_1$, כלומר זהו הערך הסינגולארי הגדול ביותר. זאת כיוון שפירוק SVD על המטריצה **T** ייתן

$$\|\mathbf{T}\|_2 = \max_{\underline{v}} \frac{\|\mathbf{T}\underline{v}\|_2}{\|\underline{v}\|_2} = \max_{\underline{v}} \frac{\|\mathbf{VDU}^T\underline{v}\|_2}{\|\underline{v}\|_2} = \max_{\underline{v}} \frac{\|\mathbf{DU}^T\underline{v}\|_2}{\|\underline{v}\|_2}.$$

ההכפלה במטריצה היוניטרית **V** אינה משנה לעניין הנורמה במונה ולכן הושמטה. נגדיר $\underline{z} = \mathbf{U}^T\underline{v}$, ונקבל

$$\|\mathbf{T}\|_2 = \max_{\underline{v}} \frac{\|\mathbf{DU}^T\underline{v}\|_2}{\|\underline{v}\|_2} = \max_{\underline{z}} \frac{\|\mathbf{D}\underline{z}\|_2}{\|\mathbf{U}\underline{z}\|_2} = \max_{\underline{z}} \frac{\|\mathbf{D}\underline{z}\|_2}{\|\underline{z}\|_2} = \sigma_1.$$

בבואנו לבחון יציבות, די בנורמה אחת קטנה מ-1 על מנת להבטיח כי המטריצה בידינו מתכנסת. ישנו משפט (שלא נביאו בשלמותו כאן) האומר שאם מערכת מתכנסת (דהיינו רדיוסה הספקטראלי קטן מ-1), קיימת נורמה אופרטורית שתגלה זאת.

מעניין לציין כי נורמת Frobenius, על אף היותה נורמה לא אופרטורית, יכולה גם היא לשמש למבחן היציבות בתנאים מרעים. זאת כיוון ש:

$$\|\mathbf{T}\|_F = \sqrt{\sum_{k=1}^{n}\sum_{j=1}^{n}\left(t_{kj}\right)^2} = \sqrt{\sum_{k=1}^{n}\sigma_k^2} \geq \sigma_1.$$

הקשר בין נורמת Frobenius ובין הערכים הסינגולריים קלה לקבלה אם נשים לב לכך שסכום אברי האלכסון (מוכר כפעולת trace) של מטריצת ה-Gram, $\mathbf{T}^T\mathbf{T}$, אינו אלא שווה לנורמה זו, והרי ה-trace שווה לסכום הערכים העצמיים שהינם ריבוע הערכים הסינגולריים.

אם קיים ערך סינגולרי אחד מעל 1, הנורמה $\|\mathbf{T}\|_2$ תאבחן התבדרות, וכך גם "תחשוב" נורמת Frobenius. אם כל הערכים הסינגולאריים מתחת 1, נורמת $\|\mathbf{T}\|_2$ תאבחן התכנסות, ואילו נורמת Frobenius תוכל להיות מתחת 1 (ואז ניבויה נכון) או מעל 1 (ואז היא פסימית מידי). לעומת זאת, לא יקרה מצב בו נורמת Frobenius תנבא התכנסות ותטעה.

כנקודה אחרונה בסעיף זה, נציג את משפטו של גרשגורן (ללא הוכחה), אשר יכול לנבא אף הוא באמצעים פשוטים יציבות.

משפט 2: למטריצה **T** בגודל n-על-n, נגדיר את n העיגולים הבאות:

$$D_k = \left\{ z \in C \;\middle|\; |z - t_{kk}| \le \sum_{j \ne k} |t_{jk}| \right\}$$

אזי כל הערכים העצמיים של המטריצה **T** מצויים בשטח המאחד את עיגולים אלה.
משפט זה אומר כי יש לקחת את איבר האלכסון כנקודת מרכז של העיגול, והרדיוס הינו סכום יתר איברי השורה (או עמודה) בערך מוחלט.

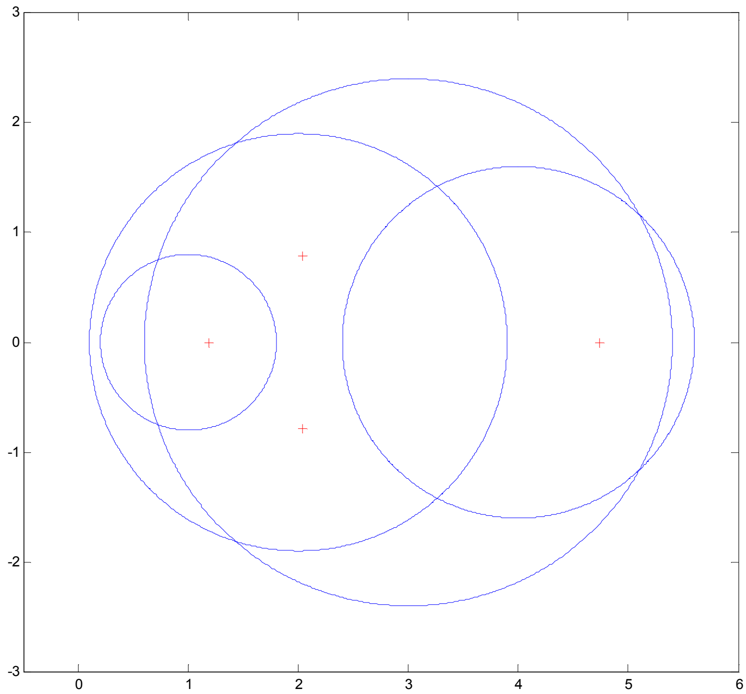
דוגמה: נתייחס למטריצה הבאה:

$$T = \begin{bmatrix} 4 & 0.4 & 0.2 & 1 \\ 0.9 & 2 & 0.3 & -0.7 \\ -0.2 & 0.1 & 1 & 0.5 \\ 1 & 1 & -0.4 & 3 \end{bmatrix}$$

העיגולים המתקבלים הם:

$$D_1 = \{ z \in C \mid |z - 4| \le 1.6 \} \quad D_2 = \{ z \in C \mid |z - 2| \le 1.9 \}$$
$$D_3 = \{ z \in C \mid |z - 1| \le 0.8 \} \quad D_4 = \{ z \in C \mid |z - 3| \le 2.4 \}$$

עיגולים אלה, ומיקום הערכים העצמיים האמיתיים מובאים בציור הבא:



והתוכנית שיצרה ציור זה הינה:

```
T=[4 0.4 0.2 1; 0.9 2 0.3 -0.7; -0.2 0.1 1 0.5; 1 1 -0.4 3]
lam=eig(T);
```

```
figure(1); clf;
plot(real(lam),imag(lam),'+r');
hold on;
% elementary circle
tt=0:2*pi/1000:2*pi; xx=sin(tt); yy=cos(tt);
plot(xx*1.6+4,yy*1.6);
plot(xx*1.9+2,yy*1.9);
plot(xx*0.8+1,yy*0.8);
plot(xx*2.4+3,yy*2.4);
axis image; axis([-0.5 6 -3 3]);
```

בעזרת משפט זה נוכל לקבוע באופן גס את מיקום הערכים העצמיים, ואם יש לנו מזל וכל עיגולים אלו יחדיו נופלים לתוך עיגול היחידה, ברור כי המטריצה יציבה. לדוגמה, אם נחזור על הדוגמה הקודמת, כשכעת המטריצה מוכפלת במקדם a, מה צריך להיות מקדם זה להבטחת יציבות? אם איננו יודעים את הערכים העצמיים ונפעל לפי משפט גרשגורן, מהציור ניכר שהנקודה הרחוקה ביותר מהראשית בשטח העיגולים היא זו המתייחסת לעיגול שמרכזו ב-4 ורדיוסו 1.6. לכן, עבור $a < 1/5.6$ נקבל כיווץ נאות שיבטיח יציבות. לעומת זאת, בהתבסס על הערכים העצמיים עצמם, די היה במקדם $a = 1/4.74$ להבטחת התכנסות.

נקודה מעניינת שנובעת ממשפט זה היא דרך לקביעה מיידית של היות המטריצה סינגולארית – אם האפס אינו כלול באזור העיגולים ברור כי המטריצה לא סינגולארית. זה מתקבל (כמו בדוגמה הנ"ל) כאשר סכום איברי כל שורה בערך מוחלט (וללא איבר האלכסון) קטן מאיבר האלכסון. מטריצה כזו קרויה Diagonally Dominant. אנו נייחס חשיבות רבה למטריצות אלה בהמשך.

4. פתרון איטרטיבי של מערכות משוואות ליניאריות

כעת נחזור ליעד המקורי – פתרון המערכת $\mathbf{A}\underline{x} = \underline{b}$. במקרים רבים פתרון בעזרת פירוק LU (או אלימיניציה גאוסית לעניין זה) אינם מעשיים. למשל עבור מטריצות גדולות (עשרות אלפי שורות ועמודות) ודלילות (שרוב איבריהן אפסים) שיטות אלו קורסות, ונדרשת אלטרנטיבה. כדוגמה אחרת, בשיטת ה-LU הפתרון מתקבל במדויק (אם נתעלם משגיאות העגלה במחשב) ורק בסוף התהליך. נרצה שיטה אחרת שכבר אחרי מעט חישובים מספת פתרון מקורב, כך שאם זה מספיק לנו, נעצור ונחסוך בחישובים מיותרים.

השיטה שנציע היא שיטה איטרטיבית, המשתמש בנוסחת העדכון הבאה לשם מציאת פתרון המערכת $\mathbf{A}\underline{x}=\underline{b}$:

$$\underline{x}_{k+1}=\mathbf{T}\underline{x}_k+\underline{c}\,.$$

ניכר כי כאן אימצנו משוואת עדכון כללית יותר, והסיבה לכך תובהר מיד. נצטרך לענות על שתי שאלות:

1. מה הקשר בין הצמד $\{\mathbf{A},\underline{b}\}$ לצמד $\{\mathbf{T},\underline{c}\}$ על מנת שאמנם תהליך איטרטיבי זה יביא לפתרונה של מערכת המשוואות הנתונה?
2. מה הדרישה על $\{\mathbf{T},\underline{c}\}$ על-מנת שתהליך זה יהיה מתכנס?

באשר לשאלה ראשונה, בהנחה שהתהליך מתכנס, עבור $k\to\infty$ נקבל את המשוואה

$$\underline{x}^*=\mathbf{T}\underline{x}^*+\underline{c}\,.$$

במשוואה זו, תוצאת המצב המתמיד קבועה ולכן זהה בשני אגפי המשוואה. לכן, אם נשתמש בנוסחת הפתרון הנכונה, נדרשו שיתקיים

$$(\mathbf{I}-\mathbf{T})\mathbf{A}^{-1}\underline{b}=\underline{c}\,.$$

למשל, פתרון אפשרי אחד הוא $\underline{b}=\underline{c}$ וכן $\mathbf{I}-\mathbf{T}=\mathbf{A}$. באופן כללי יותר נאמר כי פתרון קביל יהיה עם $\mathbf{T}$ שרירותית (כל עוד $\mathbf{I}-\mathbf{T}$ הפיכה), ו-$\underline{c}$ המחושב מהמשוואה הנ"ל.

באשר לשאלה השנייה, נציע להסתכל על השגיאה $\underline{e}_k=\underline{x}_k-\underline{x}^*$, ולראות כיצד זו מתקדמת כפונקציה של k. מתקבל

$$\underline{e}_{k+1}=\underline{x}_{k+1}-\underline{x}^*=\left(\mathbf{T}\underline{x}_k+\underline{c}\right)-\left(\mathbf{T}\underline{x}^*+\underline{c}\right)=\mathbf{T}\left(\underline{x}_k-\underline{x}^*\right)=\mathbf{T}\underline{e}_k\,.$$

אנו רואים כי למרות היות משוואת האיטרציה שונה מזו שנדונה בתחילת הפרק, הרי שלעניין התנהלות השגיאה, המשוואה התקפה היא זו אותה למדנו. ניכר כי לקבלת יציבות והתכנסות יהיה עלינו לדרוש שהרדיוס הספקטראלי של $\mathbf{T}$ קטן מ-1, או דרישה מחמירה יותר מבוססת משפט גרשגורן או נורמות של מטריצות.

כעת נעבור מתיאור עקרוני וכללי לתיאור של בחירות מסוימות של $\{\mathbf{T},\underline{c}\}$ אשר מבטיחות התכנסות (כי $\mathbf{T}^k$ מתכנס) ולפתרון הנכון. נניח כי נפרק את המטריצה $\mathbf{A}$ לסכום של 3 נתחים – משולש עליון $\mathbf{U}$, משולש תחתון $\mathbf{L}$, והאלכסון הראשי $\mathbf{D}$ (זה אינו פירוק LDU!). לכן,

$$\mathbf{A}=\mathbf{L}+\mathbf{D}+\mathbf{U}\,.$$

אזי המשוואה שעלינו לפתור היא

$$\mathbf{A}\underline{x}=(\mathbf{L}+\mathbf{D}+\mathbf{U})\underline{x}=\underline{b}\,.$$

ע"י העברת גורם האלכסון הראשי לאגף נפרד ומכפלה בהיפוך של **D** (בהנחה שמותר), נקבל

$$\mathbf{A}\underline{x} = (\mathbf{L}+\mathbf{D}+\mathbf{U})\underline{x} = \underline{b} \quad \Rightarrow \quad \begin{aligned} \underline{x} &= \mathbf{D}^{-1}\left[-(\mathbf{L}+\mathbf{U})\underline{x} + \underline{b}\right] = \\ &= -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\underline{x} + \mathbf{D}^{-1}\underline{b} = \mathbf{T}\underline{x} + \underline{c} \end{aligned}$$

הגענו לתבנית שחושפת את זהותם של **T** ו-$\underline{c}$. חשוב להבהיר כי צעד זה היה שרירותי, וניתן להציע רבים אחרים כמותו. מעצם בנייתו, מובטח כי אם תהליך זה יתכנס, הוא יתכנס לפתרון הנכון, וכעת לפנינו השאלה האם **T** שנבחר כאן מביא להתכנסות.

האלגוריתם הנ"ל מוכר בשם אלגוריתם Jacobi, והפעלתו קלה ומפתה. בהינתן מערכת משוואות $\mathbf{A}\underline{x} = (\mathbf{L}+\mathbf{D}+\mathbf{U})\underline{x} = \underline{b}$, אלגוריתם זה מציע איטרציות מהצורה

$$\underline{x}_{k+1} = -\mathbf{D}^{-1}\left[(\mathbf{L}+\mathbf{U})\underline{x}_k - \underline{b}\right].$$

יופיו של אלגוריתם זה בפשטותו. אם אנו מחזיקים פתרון מקורב כלשהו $\underline{x}_k$, תחילה עלינו לבצע את החישוב

$$(\mathbf{L}+\mathbf{U})\underline{x}_k - \underline{b} = \mathbf{A}\underline{x}_k - \mathbf{D}\underline{x}_k - \underline{b}$$

שהינו קל ודורש הכפלת הפתרון הנתון ב-**A** ובאלכסונה הראשי, והחסרת $\underline{b}$. לאחר מכן נלקח וקטור זה ומחולק איבר-איבר באלכסון של **D**, וזהו הפתרון העדכני.

ניגש אם כך לשאלה הנותרת – האם **T** הנבחר (המטריצה $-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$) מביא להתכנסות? מסתבר כי לא לכל מטריצה **A** תתקבל התכנסות! אם המטריצה **A** diagonally dominant נקבל כי סכום האיברים בערך מוחלט בכל שורה קטן מ-1, כיוון שבשורה ה-j האיברים יהיו

$$\left[-\frac{a_{j1}}{a_{jj}} \;\; -\frac{a_{j2}}{a_{jj}} \;\; -\frac{a_{j3}}{a_{jj}} \;\; \cdots \;\; -\frac{a_{j(j-1)}}{a_{jj}} \;\; 0 \;\; -\frac{a_{j(j+1)}}{a_{jj}} \;\; \cdots \;\; -\frac{a_{j(n-1)}}{a_{jj}} \;\; -\frac{a_{jn}}{a_{jj}}\right],$$

הסכום בשורה זו יהיה

$$\frac{\left(\sum_{k=1}^{n}\left|a_{jk}\right|\right) - a_{jj}}{a_{jj}}.$$

בשל היות המטריצה diagonally dominant, המונה בסכום הנ"ל קטן בערכו המוחלט מאיבר האלכסון, ולכן המנה קטנה מ-1. ולכן המקסימום של סכומים אלו יהיה אף הוא קטן מ-1, והרי זו הנורמה האופרטורית $\|\mathbf{T}\|_\infty$, ולכן יש לנו את הטענה הבאה:

משפט 3: אלגוריתם Jacobi לפתרון מערכת משוואות ליניארית $\mathbf{A}\underline{x} = \underline{b}$ מתכנס אם **A** הינה מטריצה diagonally dominant.

לדוגמה, נתייחס למטריצה הבאה אותה כבר פגשנו,

$$T = \begin{bmatrix} 4 & 0.4 & 0.2 & 1 \\ 0.9 & 2 & 0.3 & -0.7 \\ -0.2 & 0.1 & 1 & 0.5 \\ 1 & 1 & -0.4 & 3 \end{bmatrix}.$$

התוכנית הבאה מבצעת את אלגוריתם Jacobi. האלגוריתם מאותחל בווקטור אפסים, ומתבצעות 100 איטרציות. הגרף הנתון מתאר את המרחק בין הפתרון באיטרציה ה-k לפתרון הנכון (סכום ריבועי הפרשים) – גרף לוגריתמי להמחשת קצב ההתכנסות.

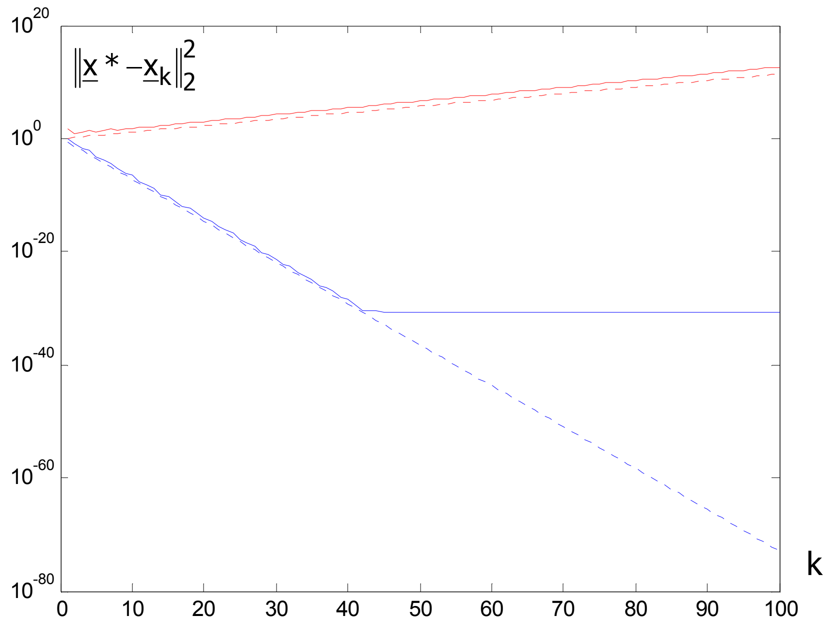
```
A=[4 0.4 0.2 1; 0.9 2 0.3 -0.7; -0.2 0.1 1 0.5; 1 1 -0.4 3];
D=diag(diag(A));
b=[1 2 3 4]';
Xs=inv(A)*b;
Di=inv(D);
X=zeros(4,1);
err=zeros(100,1);
for k=1:1:100,
    X=-Di*((A-D)*X-b);
    err(k)=(X-Xs)'*(X-Xs);
end;
semilogy(err); hold on;
KK=1:1:100;
semilogy(KK,0.432.^(2*KK),'r');
```

כפי שניתן לראות, כעבור 40 איטרציות מתקבלת התכנסות מושלמת עם שגיאה במסגרת שגיאות ההעגלה של המחשב. במקרה זה הרדיוס הספקטראלי של $-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$ הוא 0.432. הגרף הישר מתאר את ההתכנסות התיאורטית המנובאת 0.432^{2k} (החזקה 2 באה בשל שימוש בנורמת ℓ^2 בריבוע). כפי שרואים יש תיאום יפה בין הגרפים.

אותו ניסוי בוצע עבור המטריצה הבאה:

$$T = \begin{bmatrix} 1 & 0.4 & 0.2 & 1 \\ 0.9 & 1 & 0.3 & -0.7 \\ -0.2 & 0.1 & 1 & 0.5 \\ 1 & 1 & -0.4 & 1 \end{bmatrix}.$$

זו כבר לא מטריצה diagonally dominant, ובחינת הרדיוס הספקטראלי של $-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$ נותנת ערך של 1.14, דבר שמעיד על התבדרות וודאית. ואמנם, בגרף הנ"ל מתוארת גם השגיאה במקרה כזה, וניכר שישנה התבדרות. גם כאן יש תיאום בין ההתנהגות בפועל וזו המנובאת תיאורטית ע"י הרדיוס הספקטראלי.



ומה אם המטריצה לא diagonally dominant? מסתבר כי גם מקרה כזה ניתן לטיפול איטרטיבי, אם כי לא ע"י אלגוריתם Jacobi. אנו נכיר אלגוריתמים אלו כשנעסוק באופטימיזציה.

למעשה אלגוריתם Jacoby מאוד אינטואיטיבי – משמעותו המעשית היא כזו: בהנחה שיש בידינו פתרון, אנו לוקחים את המשוואה הראשונה, מקפיאים בה את כל הנעלמים לערכם הקיים בידינו למעט הראשון, ומחלצים את ערכו של המשתנה הראשון על מנת לקיים המשוואה. כך גם אנו נוהגים עם כל משוואה אחרת, כשכל משימות עדכון אלו נעשות **במקביל**.

באופן דומה מאוד לדרך בנייתו של אלגוריתם Jacobi, נוכל להציע אלגוריתם שונה ויעיל אף יותר, אם כי גם הוא מוגבל למטריצות diagonally dominant – זהו אלגוריתם Gauss-Seidel. כל ההבדל הוא במעבר מהמשוואה

$$\mathbf{A}\underline{x} = (\mathbf{L}+\mathbf{D}+\mathbf{U})\underline{x} = \underline{b}\,.$$

למשוואת האיטרציה. הפעם נעביר את גורם האלכסון הראשי והמשולש התחתון לאגף נפרד. מתקבל

$$\mathbf{A}\underline{x} = (\mathbf{L}+\mathbf{D}+\mathbf{U})\underline{x} = \underline{b} \quad \Rightarrow \quad \begin{aligned} \underline{x} &= (\mathbf{L}+\mathbf{D})^{-1}\left[-\mathbf{U}\underline{x}+\underline{b}\right] = \\ &= -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{U}\underline{x} + (\mathbf{L}+\mathbf{D})^{-1}\underline{b} = \mathbf{T}\underline{x}+\underline{c} \end{aligned}$$

ולכן משוואת האיטרציה תהיה

$$\underline{x}_{k+1} = (\mathbf{L}+\mathbf{D})^{-1}\left[-\mathbf{U}\underline{x}_k + \underline{b}\right].$$

כמקודם, חשיבות מרכזית ישנה ליכולת להפעיל את האיטרציה בפשטות. קודם זה היה כך כיוון שהכפלה בהיפוך של מטריצה אלכסונית קל למימוש. כאן ישנו אתגר גדול יותר אך אפשרי – היפוך של מטריצה משולשת תחתונה נותן פתרון בקלות ע"י שיטת ההחלפה – פתרון לאיבר הראשון, שימוש בו לפתרון השני, וכך הלאה.

ברוח ההסבר האינטואיטיבי שניתן ל-Jacoby, אלגוריתם ה-Gauss-Seidel בסך הכל מציע שינוי קטן מעל ה-Jacoby. אם כבר טרחנו ועדכננו את המשתנה הראשון בעזרת המשוואה הראשונה, מדוע שלא נשתמש בערכו החדש בבואנו לעדכן את המשתנה השני? ברור כי בכך יחול רווח כי אנו עושים שימוש בערך קרוב יותר לנכון. וכך בתהליך שאינו יכול להיות מקבילי כמקודם, אלא סדרתי בהכרח, אנו מעדכנים המשתנים ונשענים על ערכים מעודכנים כשאלה זמינים.

התוכנית הבאה פותרת את אותה משוואה, עליה המחשנו את ה-Jacobi קודם, בעזרת ה-Gauss-Seidel. הגרף המצורף מראה כיצד מואצת ההתכנסות ונדרש בפועל כחצי מכמות האיטרציות המקורית להגעה לשגיאה אפקטיבית אפס.

```
A=[4 0.4 0.2 1; 0.9 2 0.3 -0.7; -0.2 0.1 1 0.5; 1 1 -0.4 3];
b=[1 2 3 4]';
Xs=inv(A)*b;
D=diag(diag(A)); Di=inv(D); X=zeros(4,1);
err1=zeros(100,1);
for k=1:1:100,
    X=-Di*((A-D)*X-b);
    err1(k)=(X-Xs)'*(X-Xs);
end;
DL=tril(A); DLi=inv(DL); X=zeros(4,1);
err2=zeros(100,1);
for k=1:1:100,
    X=-DLi*((A-DL)*X-b);
```

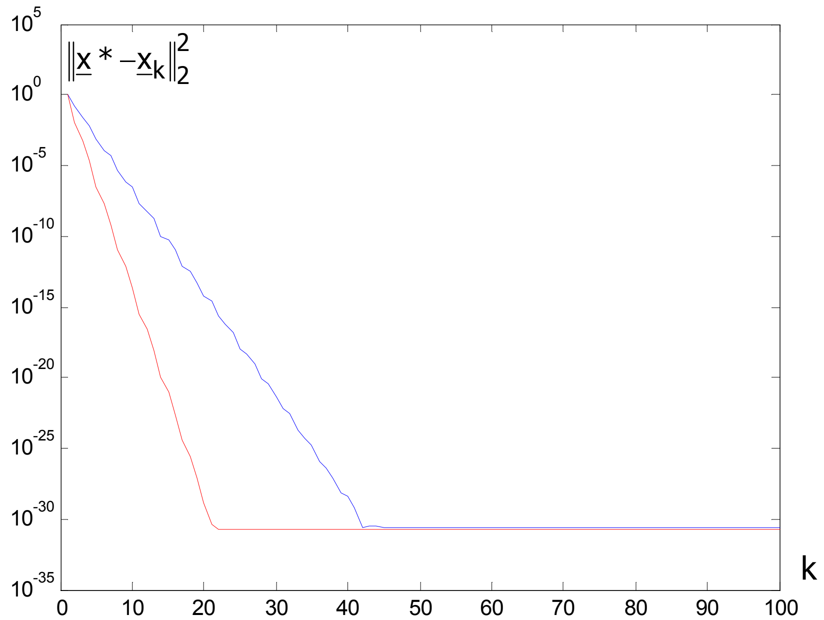
```
    err2(k)=(X-Xs)'*(X-Xs);
end;
figure(1); clf; semilogy(err1); hold on; semilogy(err2,'r');
```

משפט 4: (יובא ללא הוכחה) אלגוריתם Gauss-Seidel לפתרון מערכת משוואות ליניארית $\mathbf{A}\underline{x}=\underline{b}$ מתכנס אם **A** הינה מטריצה diagonally dominant.

וריאציה נוספת בהמשך טבעי לאלגוריתמים הקודמים היא שיטה הנקראת Successive Over Relaxation (SOR). בשיטה זו משתמשים בפירוק הבא:

$$\begin{aligned}\mathbf{A}\underline{x}&=(\mathbf{L}+\mathbf{D}+\mathbf{U})\underline{x}=\underline{b}=\\&=(\mathbf{L}+\alpha\mathbf{D}+(1-\alpha)\mathbf{D}+\mathbf{U})\underline{x}\end{aligned}\quad\Rightarrow\quad \underline{x}=-(\mathbf{L}+\alpha\mathbf{D})^{-1}\left[((1-\alpha)\mathbf{D}+\mathbf{U})\underline{x}-\underline{b}\right]$$

עבור $\alpha=1$ אלגוריתם זה יתלכד עם Gauss-Seidel. עבור $\alpha>1$ אנו מנרמלים בגורם גדול יותר ולכן נרכך את האלגוריתם (ועם זאת נייצב במצבים גבוליים בהם GS ישיר לא היה יציב). עבור $\alpha<1$ אנו מחלקים בגורם קטן ועם זאת מסכנים את יציבות האלגוריתם. אבל, אם פרמטר זה נבחר בחכמה, התוצאה תהיה התכנסות מהירה יותר.



5. פתרון איטרטיבי של בעיות אופטימיזציה ריבועיות

עד כה פעלנו איטרטיבית על מנת לפתור מערכות של משוואות. כעת נעבור לבעיות אופטימיזציה, ונתחיל בבעיות ריבועיות מהצורה הבאה:

$$p(\underline{x}) = \frac{1}{2}\underline{x}^T \mathbf{K}\underline{x} - \underline{x}^T \underline{f} + c \quad \text{או} \quad p(\underline{x}) = \frac{1}{2}\|\mathbf{A}\underline{x} - \underline{b}\|_2^2 .$$

כבר ראינו כי הבאת פונקציות אלו למינימום שקול לפתרון מערכות המשוואות

$$\mathbf{K}\underline{x} = \underline{f} \quad \text{או} \quad \mathbf{A}^T\mathbf{A}\underline{x} = \mathbf{A}^T\underline{b} ,$$

בהתאמה. לכן, נוכל להשתמש בשיטות Jacobi, Gauss-Seidel, ו-SOR לפתרון בעיות אלה, כשאנו מתייחסים להסיין ($\mathbf{K}$ או $\mathbf{A}^T\mathbf{A}$) במטריצה אותה אנו מפצלים למשולשת עליונה, תחתונה, ואלכסון ראשי. אולם ראינו כי גישות אלו יהיו טובות כאשר מטריצות אלה diagonally dominant וזו מגבלה. מה עוד ניתן לעשות?

אחת הגישות הפופולאריות ביותר למינימיזציה של פונקציה נקראת שיטת השיפוע המרבי (Steepest Descent - SD). אם אנו מחזיקים פתרון מקורב לבעיה, $\underline{x}_k$, חישוב וקטור הנגזרת של הפונקציה נותן לנו את וקטור הכיוון בו הפונקציה עושה את העלייה המרבית, ולכן הליכה בכיוון הפוך תביא לירידה בגובה הפונקציה. במקרה שלנו, וקטור הגרדיאנט נתון ע"י

$$\text{grad}\{p(\underline{x})\} = \mathbf{K}\underline{x} - \underline{f} \quad \text{או} \quad \text{grad}\{p(\underline{x})\} = \mathbf{A}^T(\mathbf{A}\underline{x} - \underline{b}) ,$$

ולכן, אנו מציעים בפועל אלגוריתם איטרטיבי מהצורה

$$\underline{x}_{k+1} = \underline{x}_k - \mu \cdot \text{grad}\{p(\underline{x}_k)\} ,$$

בו הפתרון מתעדכן ע"י הליכה בכיוון הפוך לכיוון העלייה המרבית (וזהו הכיוון לירידה המרבית), כשגודל הצעד הינו תלוי פרמטר μ. עבור הפונקציה הריבועית גישה זו נותנת

$$\underline{x}_{k+1} = \underline{x}_k - \mu(\mathbf{K}\underline{x}_k - \underline{f}) \quad \text{או} \quad \underline{x}_{k+1} = \underline{x}_k - \mu\mathbf{A}^T(\mathbf{A}\underline{x}_k - \underline{b}) .$$

שתי שאלות מתעוררות:

1. מהו גודל הצעד μ שיבטיח התכנסות? ו-
2. מה הקשר (אם קיים) לגישות קודמות?

ניקח את המשוואה $\underline{x}_{k+1} = \underline{x}_k - \mu(\mathbf{K}\underline{x}_k - \underline{f})$ נמבחן אותה כביטוי איטרטיבי מהסוג שנלמד קודם. ברור כי פתרון שיקיים את המשוואה $\mathbf{K}\underline{x} = \underline{f}$ יביא ביטוי זה לשוויון של שני האגפים. האם ביטוי זה מייצג תהליך מתכנס? לצורך כך נכתוב את הביטוי הנ"ל אחרת כ:

$$\underline{x}_{k+1} = \underline{x}_k - \mu(\mathbf{K}\underline{x}_k - \underline{f}) = (\mathrm{I} - \mu\mathbf{K})\underline{x}_k + \mu\underline{f} = \mathbf{T}\underline{x}_k + \underline{c}$$

לכן ברור כי אם המטריצה $(\mathrm{I} - \mu\mathbf{K})$ בעלת רדיוס ספקטראלי קטן מ-1, תחול התכנסות. כזכור, אנו מניחים שהמטריצה $\mathbf{K}$ חיובית מוגדרת, ולכן כל ערכיה העצמיים, λ_k, ממשיים וחיוביים. אם $\underline{v}$ הוא וקטור עצמי של $\mathbf{K}$, אזי מתקבל כי

$$\mathbf{K}\underline{v} = \lambda\underline{v} \;\Rightarrow\; (\mathbf{I} - \mu\mathbf{K})\underline{v} = (1 - \mu\lambda)\underline{v}$$

ולכן הערכים העצמיים של המטריצה $(I-\mu\mathbf{K})$ הם $(1-\mu\lambda_k)$. לכן עלינו לקבוע את כך שכל גורמים אלה יהיו בתוך האינטרוול [-1,+1]. נסתכל גראפית על הגורמים $|1-\mu\lambda_k|$ כפונקציה של μ, וננסה להבין כיצד נראית הפונקציה

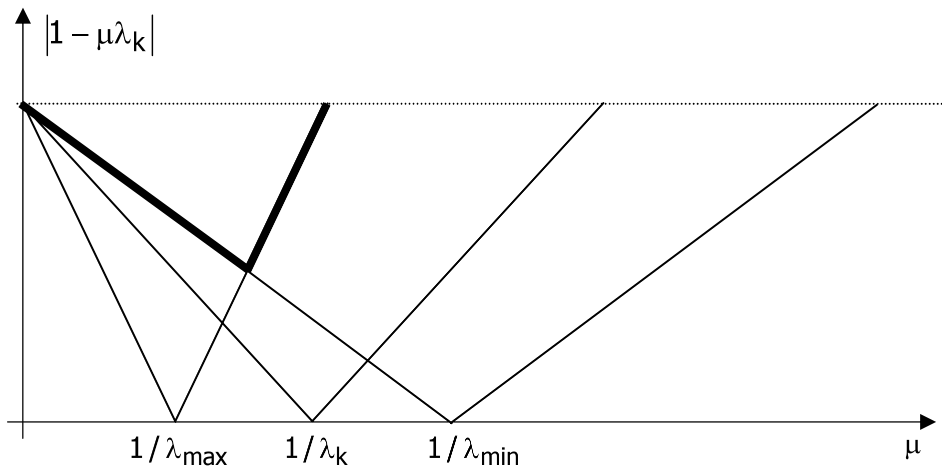
$$g(\mu)=\max_{1\le k\le n}|1-\mu\lambda_k| .$$

הקו המעובה הוא תיאור הפונקציה הנ"ל. אנו רואים כי עבור הבחירה

$$0<\mu<\frac{2}{\lambda_{max}}$$

מובטח רדיוס ספקטראלי מתחת 1 ולכן התכנסות. כמו כן, הרדיוס הספקטראלי הקטן ביותר מושג עבור המפגש בין שני הישרים המעובים

$$\lambda_{max}\cdot\mu-1=1-\lambda_{min}\cdot\mu \quad\Rightarrow\quad \mu_{opt}=\frac{2}{\lambda_{max}+\lambda_{min}} ,$$



ואז רדיוס זה הינו

$$\rho\{\mathbf{T}\}=\frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}=\frac{\frac{\lambda_{max}}{\lambda_{min}}-1}{\frac{\lambda_{max}}{\lambda_{min}}+1}=\frac{\kappa\{\mathbf{T}\}-1}{\kappa\{\mathbf{T}\}+1}$$

כזכור, ה- condition number מוגדר כיחס בין הערכים הסינגולאריים הקיצוניים. במקרה של מטריצה חיובית מוגדרת, אלה מתלכדים עם הערכים העצמיים, ולכן התוצאה הנ"ל נכונה. קיבלנו לכן כי ה-condition number של המטריצה **K** שולט בקצב ההתכנסות, ואם הוא גדול מידי זו תהיה איטית להחריד.

עד כה עסקנו בשאלה מה צריך להיות μ להשגת התכנסות, כשאנו מתמקדים בבחירת ערך קבוע. האם נרוויח משהו אם נרשה לו להשתנות מאיטרציה לאיטרציה? מסתבר כי

הדבר אפשרי והגישה לניתוח במקרה כזה שונה, וקשורה לנושא הקרוי "חיפוש על ישר". אנו מביאים למינימום את הפונקציה

$$p(\underline{x}) = \frac{1}{2}\underline{x}^T \mathbf{K}\underline{x} - \underline{x}^T \underline{f} + c$$

ובידינו הצעה לפתרון

$$\underline{x}_{k+1} = \underline{x}_k - \mu(\mathbf{K}\underline{x}_k - \underline{f}) = \underline{x}_k - \mu\underline{e}_k$$

השתמשנו בסימון $\mathbf{K}\underline{x}_k - \underline{f} = \underline{e}_k$ כי זהו וקטור שגיאה של מערכת משוואות זו. הנקודה המעניינת היא שבידינו הוקטורים $\underline{x}_k, \underline{e}_k$, ולכן הצבת ביטוי זה בפונקציה המקורית תיתן לנו פרבולה חד-ממדית,

$$\begin{aligned} p(\underline{x}_{k+1}) &= \frac{1}{2}\underline{x}_{k+1}^T \mathbf{K}\underline{x}_{k+1} - \underline{x}_{k+1}^T \underline{f} + c = \\ &= \frac{1}{2}(\underline{x}_k - \mu\underline{e}_k)^T \mathbf{K}(\underline{x}_k - \mu\underline{e}_k) - (\underline{x}_k - \mu\underline{e}_k)^T \underline{f} + c \end{aligned}$$

הנקודה המעניינת כאן היא שפונקציה ריבועית מהסוג בו אנו עוסקים תיתן כי כל חתך חד-ממדי שלה הינו פרבולה. ככזו, ברור שיש לה נקודת מינימום, ואנו נחפשו. נגזור הפונקציה הנ"ל ביחס לפרמטר μ ונקבל

$$\frac{dp(\underline{x}_{k+1})}{d\mu} = -\underline{e}_k^T \mathbf{K}(\underline{x}_k - \mu\underline{e}_k) + \underline{e}_k^T \underline{f} = 0 \quad \Rightarrow \quad \mu_k^{opt} = \frac{\underline{e}_k^T \mathbf{K}\underline{x}_k - \underline{e}_k^T \underline{f}}{\underline{e}_k^T \mathbf{K}\underline{e}_k} = \frac{\underline{e}_k^T \underline{e}_k}{\underline{e}_k^T \mathbf{K}\underline{e}_k}.$$

קיבלנו אלגוריתם המוכר בשם Normalized Steepest Descent (NSD) בו גודל הצעד משתנה מאיטרציה לאיטרציה.

מה הקשר בין כל זה ובין שיטות ה-Jacobi, Gauss-Seidel, וה-SOR? במונחים של מערכת משוואות, כל שרצינו הוא פתרון של מערכת המשוואות

$$\mathbf{K}\underline{x} = \underline{f}.$$

במערכת זו המטריצה $\mathbf{K}$ חיובית מוגדרת, אך לא בהכרח diagonally dominant. לכן יישום שיטת Jacobi והרחבותיה לא מבטיחים התכנסות. הדרך שלנו להתמודד עם זאת היא הבאה – נכפול משוואה זו במקדם μ, ונוסיף ונחסיר את $\underline{x}$. לכן המשוואה שלפנינו הינה

$$\underline{x} - \underline{x} + \mu\mathbf{K}\underline{x} = \mu\underline{f}.$$

מערכת משוואות זו נכונה כמו המקורית ופתרון שלה יהיה הפתרון הרצוי (והוא יחיד!). כעת נבצע הפרדה מהסוג שכיכב בבניית האלגוריתמים דוגמת Jacobi:

$$\underline{x}_{k+1} - \underline{x}_k + \mu\mathbf{K}\underline{x}_k = \mu\underline{f} \quad \Rightarrow \quad \underline{x}_{k+1} = \underline{x}_k - \mu\mathbf{K}\underline{x}_k + \mu\underline{f}.$$

וזו בדיוק משוואת ה-SD. אבל, כמו קודם, אין זו הדרך היחידה האפשרית. נראה כעת כיצד ניתן ליצור אלטרנטיבה כללית יותר. במקום לכפול במקדם סקלר μ, נכפול במטריצה חיובית מוגדרת וסקלר, $\mu\mathbf{P}$, ונקבל:

$$\underline{x}_{k+1} - \underline{x}_k + \mu\mathbf{PK}\underline{x}_k = \mu\mathbf{P}\underline{f} \quad \Rightarrow \quad \underline{x}_{k+1} = \underline{x}_k - \mu\mathbf{PK}\underline{x}_k + \mu\mathbf{P}\underline{f} = \underline{x}_k - \mu\mathbf{P}\left(\mathbf{K}\underline{x}_k - \underline{f}\right).$$

ניתן להראות כי לכל בחירה של $\mathbf{P}$ חיובית מוגדרת, קיים μ שייתן התכנסות. זאת כיוון שבמקרה זה המטריצה השולטת בהתכנסות היא

$$\mathbf{T} = \mathbf{I} - \mu\mathbf{PK},$$

ומסתבר כי ערכיה העצמיים של מטריצה זו יכולים להיות מובאים לתוך מעגל היחידה.

לכן יש לנו עניין בבחירת מטריצה $\mathbf{P}$ פשוטה להשגה אשר תיתן התכנסות מהירה. לדוגמה, הבחירה $\mathbf{P} = \mathbf{K}^{-1}$ ו-$\mu = 1$ תיתן התכנסות באיטרציה אחת כי

$$\underline{x}_{k+1} = \underline{x}_k - \mathbf{K}^{-1}\left(\mathbf{K}\underline{x}_k - \underline{f}\right) = \mathbf{K}^{-1}\underline{f}.$$

זה גם ניכר מהקביעה $\mathbf{T} = \mathbf{I} - \mu\mathbf{PK} = 0$. אך זאת לא הצלחה כבירה כי בניית מטריצה זו כמוה כמו פתרון מלוא הבעיה, ולו זה היה אפשרי היינו עושים זאת מלכתחילה. כשנציג בעיות שאינן ריבועיות נחזור לגישה זו, המוכרת כשיטת ניוטון, ונראה כי היא מאוד יעילה בהקשרים אחרים, אך כאן כמובן היא ריקה מתוכן.

נוכל לבחור $\mathbf{P} = \left((1-\alpha)\mathbf{I} + \alpha\mathbf{D}\right)^{-1}$ כש-$\mathbf{D}$ היא האלכסון הראשי של $\mathbf{K}$. תכונה של מטריצות חיוביות מוגדרות היא שכל אברי אלכסונן חיוביים ושונים מאפס, ולכן $\mathbf{P}$ שנבחרה חיובית מוגדרת. עבור $\alpha = 0$ נקבל את אלגוריתם ה-SD. עבור $\alpha = 1$ נקבל את אלגוריתם Jacobi, ולערכי ביניים אלגוריתם מעורב. אגב, בחירה כזו של $\mathbf{P}$ מוצלחת כיוון שקל להופכה בהיותה מטריצה אלכסונית – זאת בניגוד לבחירה הקודמת $\mathbf{P} = \mathbf{K}^{-1}$ שאינה קלה להפעלה.

נחזור כעת לאלגוריתם ה- NSD: נראה כי הוא מאוד יעיל כיוון שבכל איטרציה הוא "סוחט" היטב את פוטנציאל הצעד. עם זאת, מתקיימת התכונה הבעייתית הבאה, אשר מעידה כי ה-NSD לא טוב כל כך כפי שנדמה. כזכור, האיטרציה ב- NSD נעשית ע"י המשוואה

$$\underline{x}_{k+1} = \underline{x}_k - \mu\left(\mathbf{K}\underline{x}_k - \underline{f}\right) = \underline{x}_k - \mu\underline{e}_k.$$

ולכן כיוון התנועה נתון ע"י

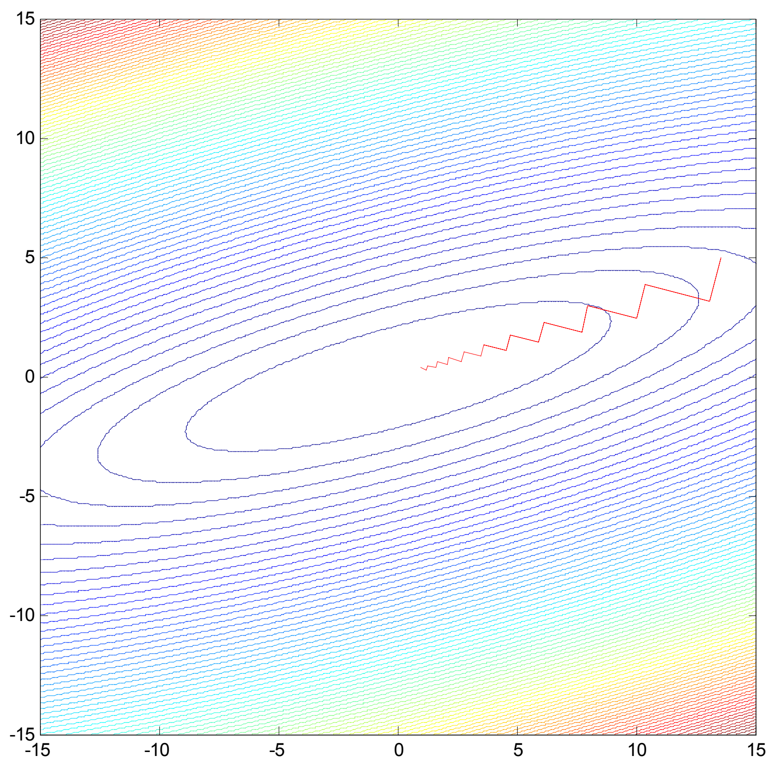
$$\underline{x}_{k+1} - \underline{x}_k = -\mu\underline{e}_k.$$

את גודל הצעד המיטבי מצאנו לפי הקשר:

$$\frac{dp(\underline{x}_{k+1})}{d\mu} = -\underline{e}_k^T\mathbf{K}(\underline{x}_k - \mu\underline{e}_k) + \underline{e}_k^T\underline{f} = -\underline{e}_k^T(\mathbf{K}\underline{x}_{k+1} - \underline{f}) = -\underline{e}_k^T\underline{e}_{k+1} = 0$$

מתקבל כי הכיוונים העוקבים באלגוריתם ה-NSD בפועל מזגזגים בהיותם אנכיים זה לזה. הציור הבא ממחיש זאת, והתוכנית שיצרה גרף זה היא הבאה:

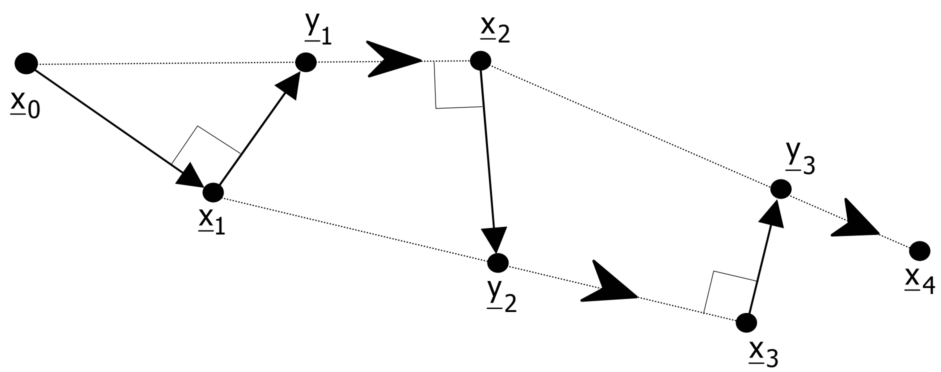
```
A=randn(2,2); A=orth(A);
K=A'*[1 0 ; 0 0.05]*A; f=[0 0]';
Xs=inv(K)*f;
[xx,yy]=meshgrid(-15:0.1:15,-15:0.1:15);
p=K(1,1)*xx.^2+2*K(1,2)*xx.*yy+K(2,2)*yy.^2-f(1)*xx-f(2)*yy;
figure(1); clf; contour(xx,yy,p,80); hold on; axis image;
X=[13.5; 5];
for k=1:1:20,
    Er=K*X-f;
    mu=(Er'*Er)/(Er'*K*Er);
    Xold=X;
    X=X-mu*Er;
    plot([X(1) Xold(1)],[X(2) Xold(2)],'r');
end;
```



מה הפתרון לבעיה זו? מסתבר כי ישנו מגוון עשיר של דרכים להאיץ את ה- NSD, והמעניין בסיפור זה הוא שחלק לא קטן משיטות אלה מגיע בסופו של דבר לאותו מקום – אלגוריתם הכיוונים הצמודים – Conjugate Directions algorithm. אלגוריתם ה-CG מאוד פופולארי וזאת בשל ביצועיו הטובים. תכונה מעניינת שלו שלא ראינו קודם היא ההבטחה שדי ב-n איטרציות להגעה לפתרון המדויק. אנו נציג כאן גרסה מסוימת ונוחה להבנה שלו – אלגוריתם ה-PARTAN.

ראינו קודם כי אלגוריתם ה-NSD עבור הבעיה הריבועית יוצר תופעת זיג-זג בהתקדמותו, כיוון שכיווני החיפוש הם הגרדיאנטים, ואלו אמורים להיות אנכיים זה לזה בצעדים עוקבים בשל החיפוש המדויק על הישר. הצעת שיפור הנקראת שיטת ה- PARTAN (Parallel Tangent) מנצלת תכונה זו. שיטה זו הוצעה לראשונה בשנת 1951 אך נותחה ונוסחה במלואה בשנת 1964. ה-PARTAN מציעה לקחת תוצאות של איטרציות עוקבות של ה- NSD ולבצע חיפוש על ישר אשר מאחד אותן.

באופן מפורט יותר, האלגוריתם נראה כך: נתחיל בנקודת אתחול שרירותית $\underline{x}_0$. ממנה באיטרציה של ה-NSD נקבל את $\underline{x}_1$. באופן דומה נבצע צעד NSD נוסף ונקבל את $\underline{y}_1$. כעת בא ההבדל - במקום לקבוע $\underline{x}_2 = \underline{y}_1$ כפי שעושה זאת ה- NSD, נקבע את $\underline{x}_2$ כנקודה המשיגה את המינימום של הפונקציה הריבועית על הישר אשר נמתח בין $\underline{x}_0$ ל- $\underline{y}_1$. באופן דומה, אם נניח כי אנו באיטרציה ה- k - ובידנו הנקודה $\underline{x}_k$, צעד NSD יוצר את הנקודה $\underline{y}_k$, והנקודה $\underline{x}_{k+1}$ תתקבל על הישר הקושר את $\underline{x}_{k-1}$ ל- $\underline{y}_k$. תיאור גרפי של הרעיון מתואר בציור הבא.



נכתוב את משוואות האלגוריתם אשר תואר קודם מילולית – אנו בנקודה $\underline{x}_k$, ותחילה נחשב את השגיאה עבור הפתרון הנוכחי,

$$\underline{e}_k = \mathbf{K}\underline{x}_k - \underline{f}.$$

עם שגיאה זו מוגדרת האיטרציה הבאה של ה-NSD,

$$\underline{y}_k = \underline{x}_k - \frac{\underline{e}_k^T \underline{e}_k}{\underline{e}_k^T \mathbf{K} \underline{e}_k} \underline{e}_k.$$

כעת נחפש על הישר המתוח בין $\underline{x}_{k-1}$ ל- $\underline{y}_k$:

$$\theta_k = \arg\min_{\theta} \; p\left(\theta \underline{y}_k + (1-\theta)\underline{x}_{k-1}\right).$$

פרמטר זה ניתן לחישוב (דומה לדרך בה חישבנו את μ ב-NSD):

$$\begin{aligned} \theta_k &= \arg\min_{\theta} \; p\left(\theta \underline{y}_k + (1-\theta)\underline{x}_{k-1}\right) = \\ &= \arg\min_{\theta} \frac{1}{2}\left(\theta \underline{y}_k + (1-\theta)\underline{x}_{k-1}\right)^T \mathbf{K}\left(\theta \underline{y}_k + (1-\theta)\underline{x}_{k-1}\right) - \left(\theta \underline{y}_k + (1-\theta)\underline{x}_{k-1}\right)^T \underline{f} \end{aligned}.$$

גזירה והשוואה לאפס נותנים

$$\begin{aligned} 0 &= \theta\left(\underline{y}_k - \underline{x}_{k-1}\right)^T \mathbf{K}\left(\underline{y}_k - \underline{x}_{k-1}\right) + \left(\underline{y}_k - \underline{x}_{k-1}\right)^T \mathbf{K}\underline{x}_{k-1} - \left(\underline{y}_k - \underline{x}_{k-1}\right)^T \underline{f} = \\ &= \theta\left(\underline{y}_k - \underline{x}_{k-1}\right)^T \mathbf{K}\left(\underline{y}_k - \underline{x}_{k-1}\right) + \left(\underline{y}_k - \underline{x}_{k-1}\right)^T \left(\mathbf{K}\underline{x}_{k-1} - \underline{f}\right) \end{aligned},$$

ולכן הפתרון הוא

$$\theta_k = \frac{\left(\underline{y}_k - \underline{x}_{k-1}\right)^T \underline{e}_{k-1}}{\left(\underline{y}_k - \underline{x}_{k-1}\right)^T \mathbf{K}\left(\underline{y}_k - \underline{x}_{k-1}\right)}.$$

כשאנו משתמשים בערך זה, נקבע:

$$\underline{x}_{k+1} = \theta_k \underline{y}_k + (1-\theta_k)\underline{x}_{k-1}$$

אחד מיתרונותיו הבולטים של אלגוריתם זה הוא שמכל נקודה נתונה $\underline{x}_k$, שיטה זו אינה יכול להיות גרועה יותר מה- NSD - כיוון שמבוצע חיפוש על ישר הכולל את נקודת ה-NSD, ואם היא הכי נמוכה, נקבל אותה ונתלכד עם ה- NSD. בנוסף, אלגוריתם זה פשוט ליישום, ומורכבותו החישובית גבוהה רק במקצת בהשוואה ל- NSD.

להלן תוצאה אופיינית של אלגוריתם זה, המתקבלת מהקוד הבא. בדוגמה זו נבנתה בעיה ריבועית במימד 5, ונפתרה ע"ה ה-NSD וה-PARTAN.

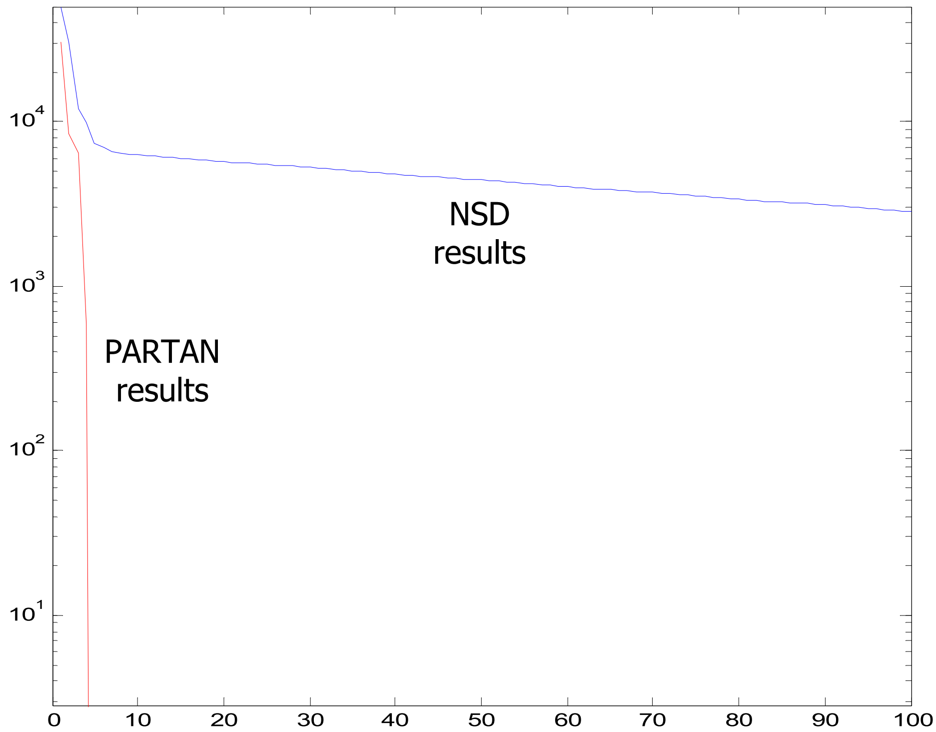
```
A=randn(5,5); A=orth(A);
K=A'*diag([1 0.03 2 7 0.01])*A; f=zeros(5,1);
Xs=inv(K)*f;
% NSD
errVEC=zeros(100,1);
X=100*ones(5,1);
for k=1:1:100,
    errVEC(k)=X'*X;
```

```
    disp(X'*X);
    Er=K*X-f;
    mu=(Er'*Er)/(Er'*K*Er);
    X=X-mu*Er;
end;
% PARTAN
errVECn=zeros(40,1);
Xold=100*ones(5,1);
Er=K*Xold-f;
mu=(Er'*Er)/(Er'*K*Er);
X=Xold-mu*Er;
for k=1:1:40,
    errVECn(k)=X'*X;
    Er=K*X-f;
    mu=(Er'*Er)/(Er'*K*Er);
    Y=X-mu*Er;
    tt=-((Y-Xold)'*(K*Xold-f))/((Y-Xold)'*K*(Y-Xold));
    Xnew=tt*Y+(1-tt)*Xold;
    Xold=X;
    X=Xnew;
end;
figure(1); clf; semilogy(errVEC); hold on;
semilogy(errVECn,'r');
```

כאמור, אלגוריתם הגרדיאנטים הצמודים אשר נראה שונה מלכתחילה, אינו אלא וריאציה על אלגוריתם זה. נשים לב כי האלגוריתם שהוצע הינו אלגוריתם בעל משוואת עדכון מסדר שני, כיוון שבעיקרה היא מציעה את דרך העדכון הבאה:

$$\begin{aligned} \underline{x}_{k+1} &= \theta_k\left(\underline{x}_k + \mu_k\left(\underline{f} - \mathbf{K}\underline{x}_k\right)\right) + \left(1-\theta_k\right)\underline{x}_{k-1} = \\ &= \theta_k\left(\mathbf{I} - \mu_k\mathbf{K}\right)\underline{x}_k + \left(1-\theta_k\right)\underline{x}_{k-1} + \mu_k\theta_k\underline{f} \end{aligned}$$

השימוש בשתי נקודות קודמות בתהליך לקביעת הנקודה החדשה הוא רעיון המכליל את תפיסת האלגוריתמים שהוצגו קודם, וברור כי בוא יכול רק להועיל אם ייושם נכון.



6. פתרון איטרטיבי של בעיות ערכים עצמיים

מסתבר כי גם מציאת ערכים עצמיים יכולה להיעשות בשיטות איטרטיביות. כמקודם, יופיין של שיטות אלה ביכולתן להתמודד עם מטריצות גדולות מאוד, ובכך שגם אם לא ימצו את החיפוש, התוצאות החלקיות בעלות משמעות. מעבר לכך, לעתים יש עניין במציאת הערך העצמי הגדול או הקטן ביותר, ואז שיטות אלו קלות לאין ערוך מגישות מלאות בכל מקרה חשוב לומר כי במימדים סבירים ומעלה של מטריצות לא מקובל ולא נכון להשתמש בפולינום האופייני למציאת הערכים העצמיים. מסתבר כי גישה זו מפוקפקת בדיוקה ורגישה מאוד, וערכה לכך הוא תיאורטי.

נתחיל עם שיטה המוכרת בשמה – The Power Method (PM) – אשר נועדה למצוא את הערך העצמי הגדול ביותר של מטריצה, בהנחה שהוא מבודד. שיטה זו בנויה על רעיון אותו כבר פגשנו בתחילת פרק זה.

כשדנו בתהליך ליניארי סדרתי מהצורה $\underline{u}_{k+1} = \mathbf{T}\underline{u}_k$ אם אתחול $\underline{u}_0$, ראינו כי קיים הקשר

$$\underline{u}_k = \mathbf{T}^k \underline{u}_0 .$$

עבור מטריצה $\mathbf{T}$ הניתנת ללכסון, עם n ערכים עצמיים λ_j ו-n וקטורים עצמיים תואמים להם $\underline{v}_j$, מתקבל כי בשלב ה-k מוצא התהליך הוא

$$\underline{u}_k = \mathbf{V}\Lambda^k \underline{x}_0 = \sum_{j=1}^{n} x_0(j)\lambda_j^k \underline{v}_j$$

כשאנו מסמנים $\mathbf{V}\underline{x}_0 = \underline{u}_0$. אם $|\lambda_1| < |\lambda_k|$, $k = 2,3,\ldots,n$, יתקבל עבור k מספיק גדול כי המוצא הינו בקירוב

$$\underline{u}_k = \mathbf{V}\Lambda^k \underline{x}_0 = \sum_{j=1}^{n} x_0(j)\lambda_j^k \underline{v}_j \cong x_0(1)\lambda_1^k \underline{v}_1 ,$$

כלומר נקבל את הווקטור העצמי הראשון המוכפל בקבוע. ממנו נוכל להקיש על הערך העצמי התואם ע"י

$$\lambda_1 \cong \frac{\underline{u}_k}{\underline{u}_{k-1}}$$

(חלוקה איבר-איבר, ומיצוע או משהו דומה), או יותר טוב, ע"י

$$\lambda_1 \cong \frac{\underline{u}_k^T \mathbf{T} \underline{u}_k}{\underline{u}_k^T \underline{u}_k} .$$

זהו כמעט ה-Power method במדויק – שינוי חשוב המיושם בדרך כלל הוא שילוב של נרמול המוצא לאחר כל איטרציה, כך שהמוצא לא יגדל לאינסוף, וברור כי נרמול כזה לא משנה לעניין ההתכנסות לווקטור העצמי הנכון.

בעזרת אלגוריתם זה ניתן גם למצוא ערכים עצמיים אחרים. למשל, לשם מציאת הערך העצמי הקטן ביותר, נעבוד אם אותו אלגוריתם בו האיטרציות מבוצעות ע"י

$$\underline{u}_{k+1} = \mathbf{T}^{-1}\underline{u}_k \quad \Rightarrow \quad \mathbf{T}\underline{u}_{k+1} = \underline{u}_k$$

כאן כל איטרציה דורשת פתרון של מערכת משוואות, ולכן אלגוריתם זה מורכב יותר.

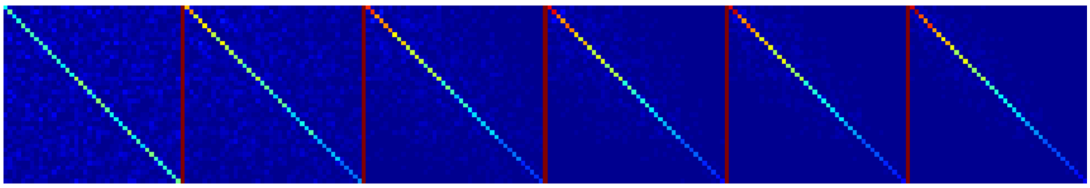
מה כדאי לעשות כאשר נדרשים כל הערכים העצמיים והוקטורים התואמים להם? הנה אלגוריתם המכליל את שיטת ה-PM למשימה זו – גישה זו קרויה Block-PM. נניח לשם פשטות שבידינו מטריצה חיובית מוגדרת, לה יש וקטורים עצמיים המהווים בסיס אורתונורמלי. בשיטת ה-PM אנו מקדמים וקטור אחד מאיטרציה לאיטרציה ע"י הכפלתו ב-**T** ונרמולו. כעבור מספיק איטרציות וקטור זה מתכנס לווקטור עצמי המתאים לערך העצמי הגדול ביותר. אם נרצה לקדם N וקטורים בזמן מתוך תקווה לקבל N וקטורים עצמיים, נפעיל את התהליך המשולב הבא:

$$\left\{\underline{u}_{k+1}^j = \mathbf{T}\underline{u}_k^j\right\}_{j=1}^N .$$

אם נאתחל אפילו עם N וקטורים שונים, תהליכים אלו יגיעו כולם לאותו וקטור עצמי, ולא השגנו דבר. ברור גם שהפעלת N התהליכים חסרת כל צימוד. אם לעומת זאת

אחרי כל איטרציה נבצע אורתוגונליזציה של N הוקטורים המתקבל, נבטיח צימוד מוצלח יותר של הבעיות, והתכנסות לוקטורים העצמיים. ההסבר האינטואיטיבי (אנו ננפנף ידיים כאן ולא נביא הוכחה מסודרת להתנהגות זו) לכך הוא שהוקטורים העצמיים המקוריים צפויים להיות אנכיים זה לזה, וכאשר אנו מבצעים את תהליך Gram-Schmidt, אנו מתקרבים אליהם. מעבר לכך, בהפעלת האיטרציה, ההשפעה ההדדית של וקטור אחד על האחר יורדת. התוכנית הבאה ממחישה אלגוריתם זה. נשים לב לכך שהאורתוגונליזציה נעשית על-ידי פירוק QR.

```
n=40;
A=randn(3*n,n);
A=A'*A;
LL=eig(A);
Q=eye(n);
RES=[abs(Q'*A*Q)];
for k=1:1:5,
    Q=A*Q;
    [Q,R]=qr(Q);
    RES=[RES,max(LL)*ones(n,1),abs(Q'*A*Q)];
    imagesc(RES); axis image; pause;
end;
axis off;
```



רואים כי עם התקדמות האיטרציות מתקרבת המטריצה $\mathbf{Q}^T\mathbf{AQ}$ למטריצה אלכסונית, דבר המעיד על היות **Q** מטריצה המכילה את הוקטורים העצמיים.

אנו נעצור את הדיון כאן – יש עוד גישות רבות וטובות למציאת ערכים ווקטורים עצמיים, אך אלה בדרך כלל מורכבים יותר לתיאור ולהוכחה.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 7 – אופטימיזציה ללא אילוצים

נושאים שייסקרו:

1. מינימיזציה של פונקציה?
2. תפקיד הגרדיאנט וההסיין
3. קבוצות קמורות ופונקציות קמורות
4. מינימיזציה חד-ממדית – שיטות חיפוש על ישר
5. שיטות בסיסיות למינימיזציה איטרטיבית (SD וניוטון)

1. מינימיזציה של פונקציה?

נתחיל בשאלה המתבקשת - מהי בעיית אופטימיזציה? נגדיר את הבעיה הבאה - נתונה פונקציה $f(\underline{x}):\mathbf{R}^n \rightarrow \mathbf{R}^1$, כלומר פונקציה המקבלת וקטור של n ערכים המסומן $\underline{x}$, ומניבה ערך סקלרי. בעיית אופטימיזציה קלאסית תהיה - מצא את הווקטור $\underline{x}$ אשר מגשים את המינימום של הפונקציה f, דהיינו - $\underline{x}^* = \text{ArgMin}\, f(\underline{x})$. יש לשים לב לעובדה שאין צורך לדון בבעיות מינימום ובעיות מקסימום ביחד - זוהי כפילות מיותרת כיוון שכל בעיית מקסימום ניתנת לייצוג כבעיית מינימום ע"י הכפלת הפונקציה המטופלת ב- (1-). לכן, רוב הדיון בקורס זה יתמקד בבעיות מינימום.

בעיה זו שהגדרנו אינה הכללית ביותר, כיוון שלא שילבנו בה אילוצים. במקרה הכללי עשויה הבעיה לכלול את חיפוש ה- $\underline{x}$ אשר מביא למינימום את הפונקציה בכפוף לשלושה סוגים אפשריים של אילוצים (ואפשר כמובן מספר אילוצים מכל סוג ושילובים שונים שלהם):

א. אילוצי שייכות לקבוצה - $\underline{x} \in \Omega$.

ב. אילוצי שוויון - $h(\underline{x}) = 0$.

ג. אילוצי אי-שוויון - $g(\underline{x}) \geq 0$.

במסגרת פרק זה נעסוק בהיבטים מתמטיים של בעיות אופטימיזציה, ובאלגוריתמים הנובעים מניתוח אנליטי זה. תחום תורת האופטימיזציה הינו רחב ועם פעילות מחקרית ענפה. המדעים המדויקים, ההנדסה, כלכלה, ניהול ועוד תחומים מספקים אינספור סוגיות בהן בעיות מגוונות הופכות לבעיות אופטימיזציה בהן יש למצוא ערכים שיביאו פונקציה מסוימת למינימום או מקסימום. במהלך מאה ויותר השנים האחרונות הועלו מגוון רחב מאוד של גישות לפתרון בעיות אופטימיזציה. בתחילה עסקו בגישות אנליטיות (כדוגמה - גישת כופלי לגרנז' לפתרון בעיות עם אילוצים), אך עם התקדמות מקבילה בעולם המחשוב גבר העניין בדרכים נומריות ובתחום זה מרוכז הפרק. המעבר

מפתרון אנליטי לגישות נומריות אינו נובע רק מהתפתחות עולם המחשוב. במקרים מסוימים (והם רבים והולכים), מעבר זה הכרחי. סיבות אופייניות לכך יכולות להיות:

א. אי-ליניאריות מורכבת של הפונקציה - במקרה כזה טיפול אנליטי במקרה הטוב מסורבל, ובמקרה הפחות טוב בלתי אפשרי.
ב. כמות נעלמים גדולה - במדעים המדויקים אין זה נדיר למצוא בעיות אופטימיזציה עם אלפי, עשרות אלפי, ואף מיליוני נעלמים. מצבים אלו כמובן מונעים כל אפשרות מעשית לטיפול אנליטי.
ג. פונקציות מערכתיות - במקרים מסוימים, במקום שתהיה נתונה פונקציה כביטוי מתמטי, היא נתונה כקופסה שחורה בה יש לטפל. במצב כזה רק טיפול נומרי ניתן ליישום.

ואמנם, רבים והולכים היישומים בהם נדרש טיפול נומרי לבעיית אופטימיזציה. לטיפול כזה מספר מאפיינים מובהקים - בראש ובראשונה - מעורבותו של מחשב כזה או אחר בביצוע מלאכת האופטימיזציה. כפועל יוצא מכך, אין מקום לדבר על רצף - כל הטיפול הוא דיסקרטי, ולכן וקטור הנעלמים יהיה פשוט סידרה של מספרים אשר את ערכיהם אנו מחפשים. הקו המנחה של גישות נומריות לבעיות אופטימיזציה הוא גישה איטרטיבית, בה מוצע פתרון ראשוני, וממנו בעדכונים לפי כללים שונים משופר הפתרון עד להתכנסות. וישנם מגוון רחב מאוד של אלגוריתמים בעלי מבנה כללי איטרטיבי המסוגלים לפתור בעיות אופטימיזציה, ביניהם יש לבחור את השיטה המתאימה ביותר למקרה המטופל. למעשה, כבר נתקלנו בבעיות אופטימיזציה כשדנו בבעיית ה-LS, ואז ראינו את המוטיבים הנ"ל בפעולה.

סוגיה דומה לאופטימיזציה, אשר זכתה לטיפול נרחב, היא פתרון נומרי של מערכות של משוואות. יש קשר ואף חפיפה בין מגוון התחומים הללו. כך למשל - ידיעת הפתרון לבעיית אופטימיזציה שקול לידיעת פתרון מערכת משוואות ולהיפך, כיוון שבהינתן מערכת משוואות מהצורה:

$$h_k(x_1, x_2, \ldots, x_n) = 0 \qquad k = 1,2,\ldots,n,$$

אנו יכולים להגדיר בעיית מינימיזציה לפונקציה הבאה:

$$f(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} h_k^2(x_1, x_2, \ldots, x_n).$$

ברור כי פתרונה שקול לפתרון מערכת המשוואות. מהכיוון ההפוך, ידוע (ואנו נראה זאת במפורט בהמשך) כי תנאי הכרחי לנקודת מינימום היא איפוס הגרדיאנט. לכן, בהינתן $f(x_1, x_2, \ldots, x_n)$, גזירתה לפי כל אחד מהמשתנים והשוואה לאפס מניבה מערכת

משוואות. באופן דומה ניתן להראות כי ידיעת פתרון מערכת משוואות שקול לידיעת פתרון נומרי של משוואות דיפרנציאליות חלקיות, וזאת אולי נציג בהמשך הקורס.

בבואנו לתקוף בעיות אופטימיזציה, אין ספק כי חלוקת אוסף כלל בעיות האופטימיזציה למשפחות מקילה את הדרך לפתרונן, משום שלכל משפחה של בעיות ניתן ל-"תפור" גישות מתאימות שימצו את כל הפוטנציאל האפשרי הטמון במשפחה, ויביאו לפתרון מהיר ופשוט ככל האפשר. חלוקה אפשרית אחת לקטגוריות היא לפי גודל הבעיה - לפי כמות הנעלמים, כמות האילוצים (אם יש), מידת האי-ליניאריות של הבעיה ועוד. ברור כי יש הבדל מהותי בין בעיה בת שני נעלמים ובין בעיה עם אלפי נעלמים, וסביר כי גישות יעילות לפתרון האחת לא יתאימו או יהיו מאוד לא יעילות לאחרת.

החלוקה לפיה נפעל במסגרת פרק זה היא אחרת. בתחילה נתייחס לטיפול בבעיות אופטימיזציה נטולות, ולאחר מכן נכליל לבעיות אופטימיזציה עם אילוצים. למעשה יכולנו מיד להתחיל בבעיות עם אילוצים - וכמקרה פרטי שלהן לדבר על בעיות נטולות אילוצים. גישה זו לא נבחרה משום שהטיפול בבעיות עם אילוצים מורכב למדי, ורצוי להכיר תחילה מסגרת של בעיות פשוטה יותר לשם הבנת העקרונות בתחום.

מהו העניין שלנו כמהנדסים בפרק זה של אלגוריתמי אופטימיזציה נומריים? מסתבר כי אין כמעט תחום במדעי המחשב, הנדסת חשמל, תעשיה וניהול, מכונות, הנדסה אזרחית, כימיה ועוד, אשר אינו נדרש לפרק זה. מהנדסים במדעי המחשב ימצאו צורך לתחום זה בעיקר עיבוד אותות ותמונות, ראיה ממוחשבת, תקשורת, מבנה מחשבים, גרפיקה, ויזואליזציה, גישות חיפוש במאגרי נתונים, ועוד.

לפני שנתחיל בדיון המפורט, ניתן תזכורת קצרה לקירוב טיילור של פונקציות - לפי משפט טיילור, בהינתן פונקציה ממשית סקלרית $f(\underline{x})\in C^1$, ובהינתן שתי נקודות $\underline{x}$ ו- $\underline{y}$ בתחום הגדרת הפונקציה, אזי ניתן לכתוב את שני הקשרים הבאים:

$$f(\underline{y}) = f(\underline{x}) + \nabla f(\underline{x})(\underline{y}-\underline{x}) + o(\|\underline{x}-\underline{y}\|)$$
$$f(y) = f(\underline{x}) + \nabla f(\theta\underline{x} + (1-\theta)\underline{y})(\underline{y}-\underline{x}) \quad \theta\in[0,1]$$

שפירושם שהערך בנקודה $\underline{y}$ ניתן לקירוב ע"י הערך בנקודה $\underline{x}$, עם גורם עדכון הקשור לנגזרת הפונקציה בנקודה $\underline{x}$ (או בנקודה בין $\underline{x}$ ל- $\underline{y}$ אם אין רצוננו בקירוב), וגורם שגיאה שעבור מרחק בין הנקודות הדועך לאפס, דועכת לאפס מהר יותר, דהיינו:

$$\lim_{\alpha\to 0}\frac{o\{\alpha\}}{\alpha} = 0$$

אם הפונקציה ממשית סקלרית ושייכת ל- C^2 אז ניתן לכתוב קירוב מסדר שני,

$$f(\underline{y}) = f(\underline{x}) + \nabla f(\underline{x})(\underline{y}-\underline{x}) + \frac{1}{2}(\underline{x}-\underline{y})^T \nabla^2 f(\underline{x})(\underline{x}-\underline{y}) + o\left(\|\underline{x}-\underline{y}\|^2\right)$$

.

$$f(\underline{y}) = f(\underline{x}) + \nabla f(\underline{x})(\underline{y}-\underline{x}) + \frac{1}{2}(\underline{x}-\underline{y})^T \nabla^2 f(\theta\underline{x} + (1-\theta)\underline{y})(\underline{x}-\underline{y})$$

אנו רואים כי כל פונקציה "חלקה דייה" (בשל היכולת שלה להיגזר) ניתנת לקירוב לא רע (לפחות מקומית) כפונקציה ריבועית – דבר השופך אור על העניין שהיה לנו כבר בפונקציה זו בעבר.

2. תפקיד הגרדיאנט וההסיין

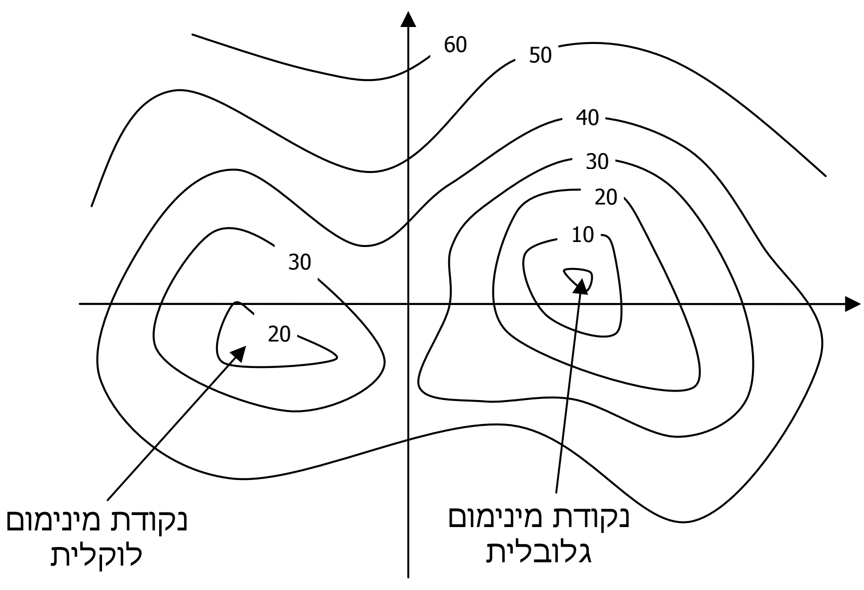
נתחיל בהגדרתן של נקודת מינימום לוקלית וגלובלית: נקודה כלשהי $\underline{x}^*$ תיקרא נקודת מינימום לוקלית אם

$$\exists \varepsilon > 0 \mid \quad \forall \|\underline{x} - \underline{x}^*\| < \varepsilon \quad f(\underline{x}) \geq f(\underline{x}^*).$$

אם סימן הגדול-שווה מוחלף בסימן גדול, תיקרא נקודה זו נקודת מינימום לוקלית ממש. באופן דומה, נקודה $\underline{x}^*$ תיקרא נקודת מינימום גלובלית אם קיים

$$\forall \underline{x} \quad f(\underline{x}) \geq f(\underline{x}^*).$$

אם סימן הגדול-שווה מוחלף בסימן גדול, תיקרא נקודה זו נקודת מינימום גלובלית ממש.



למרות שרצוננו בדרך כלל במינימום גלובלי, אנו נעסוק במשך פרק זה בנקודות המינימום הלוקליות. הסיבה לכך היא שמשיקולים מעשיים לא ניתן לבצע חיפוש של נקודות מינימום גלובליות ללא חיפוש מייגע שעל-פי רוב גם אינו מעשי. לעומת זאת, אם נציע אלגוריתמים איטרטיביים אשר מתכנסים לנקודת מינימום לוקלי, ונאתחל את האלגוריתם באופן נאות (קרוב לנקודת המינימום הגלובלית), הרי שממילא נקבל את

הפתרון הרצוי. מעבר לכך, ישנן פונקציות מיוחדות (קמורות) המבטיחות כי כל נקודת מינימום לוקלי היא מינימה גלובלית – נכיר אותן בהמשך.

בהימצאנו בנקודה כלשהי $\underline{x}$, ורצוננו לנוע בכיוון $\underline{d}$ ממנה, אנו מגדירים ישר פרמטרי מהצורה $\underline{x}(\alpha)=\underline{x}^*+\alpha\underline{d}$. הפונקציה $f(\underline{x}(\alpha))=g(\alpha)$ היא פונקציה בת משתנה יחיד ומתארת חתך של f בכיוון שנבחר. עבור $\alpha=0$ וסביבתה הקרובה, פונקציה זו ניתנת לתיאור ע"י קירוב טיילור (תחת הנחת קיומה של נגזרת ראשונה) כ:

$$g(\alpha)-g(0)=g'(0)\alpha+o(\alpha).$$

עבור $\alpha\to 0$, הגודל $o(\alpha)$ דועך לאפס יותר מהר מאשר α (תכונות של טור טיילור).
לכן, אם ידוע לנו כי $\underline{x}^*$ היא נקודת מינימום לוקלית, אזי בהכרח

$$g(\alpha)-g(0)\geq 0 \quad\Rightarrow\quad g'(0)=0.$$

משמעות קביעה זו היא

$$g'(0)=\left.\frac{dg(\alpha)}{d\alpha}\right|_{\alpha=0}=\left.\frac{df(\underline{x}^*+\alpha\underline{d})}{d\alpha}\right|_{\alpha=0}=\underline{d}^T\nabla f(\underline{x}^*)\geq 0,$$

(כאן נעשה שימוש בכלל השרשרת) ומכיוון שהקשר הנ"ל נכון לכל כיוון $\underline{d}$, מתקבל כי יש לדרוש $\nabla f(\underline{x})\equiv 0$.

עד כה עסקנו בהתנהגות הנגזרת הראשונה. מטריצת ההסיין של פונקציה סקלרית $f(\underline{x})$ מוגדרת כמטריצת הנגזרות השניות בנקודה $\underline{x}$

$$\mathbf{H}(\underline{x})=\nabla^2 f(\underline{x})\overset{\Delta}{=}\begin{bmatrix}\frac{\partial^2 f(x)}{\partial x_1\partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1\partial x_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 f(x)}{\partial x_n\partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n\partial x_n}\end{bmatrix}.$$

אם נשתמש בקירוב טיילור מדויק יותר סביב $\alpha=0$, נכתוב

$$g(\alpha)-g(0)=g'(0)\alpha+\frac{1}{2}g''(0)\alpha^2+o(\alpha^2)=\frac{1}{2}g''(0)\alpha^2+o(\alpha^2)$$

הגודל $g'(0)\alpha$ הושמט כיוון שהוא צריך להתאפס לפי האמור קודם. לכן, אם ידוע לנו כי $\underline{x}^*$ היא נקודת מינימום לוקלית, אזי בהכרח $g''(0)\geq 0$ (שוב ניתן להזניח את השפעת הגורם $o(\alpha^2)$ בשל דעיכתו המהירה לאפס ליד הראשית). לכן נדרוש

$$g''(0)=\left.\frac{d^2 g(\alpha)}{(d\alpha)^2}\right|_{\alpha=0}=\left.\frac{d^2 f(\underline{x}^*+\alpha\underline{d})}{(d\alpha)^2}\right|_{\alpha=0}=\underline{d}^T\nabla^2 f(\underline{x}^*)\cdot\underline{d}=\underline{d}^T\mathbf{H}(\underline{x}^*)\cdot\underline{d}\geq 0$$

ומכיוון שהקשר הנ"ל נכון לכל $\underline{d}$, מתקבל כי יש לדרוש כי ההסיין יהיה מטריצה חיובית חצי מוגדרת, $\mathbf{H}(\underline{x}^*) \geq \mathbf{0}$. אם נדרוש חיוביות מוגדרת ממש בנוסף לאיפוס הגרדיאנט, נבטיח כי זו נקודת מינימום לוקלית ממש. לכן, יש לנו את התוצאה הבאה:

משפט 1: עבור פונקציה f גזירה ברציפות פעמיים ($f \in C^2$), תהי הנקודה $\underline{x}^*$ נקודה כלשהי בתחום ההגדרה של הפונקציה f. הנקודה $\underline{x}^*$ היא נקודת מינימום לוקלית ממש של הפונקציה f אם ורק אם מתקיים

$\nabla f(\underline{x}^*) \equiv 0$, וכן כי $\nabla^2 f(\underline{x}^*) > 0$.

ממשפט זה נובע כי בנקודת המינימום $\underline{x}^*$, מתקיימת מערכת של n משוואות הנתונות ע"י $\nabla f(\underline{x}^*) \equiv 0$. עם זאת, בד"כ מערכת המשוואות המתקבלת קשה לפתרון ישיר, ואנו נתעניין בדרכים עוקפות לפתרון ישיר מעין זה. כמו כן, כאשר תנאי זה מתקיים הניתוח של אופי הנקודה לא מסתיים כיוון שיש לבדוק גם את הנגזרות השניות.

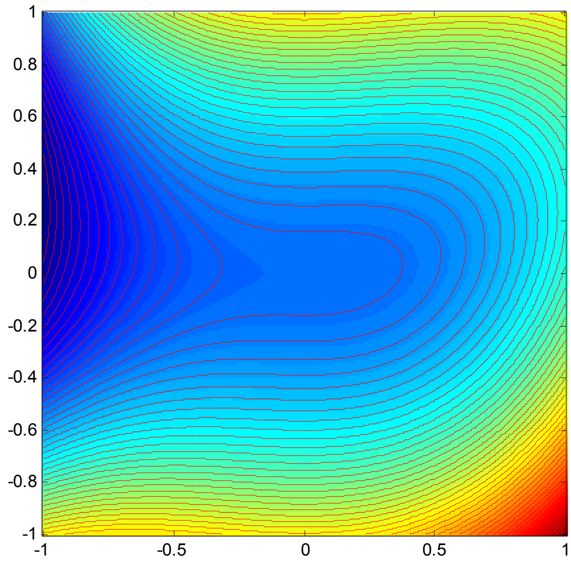
כדוגמה, נתונה הפונקציה $f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2$, ורצוננו למצוא לה נקודת מינימום. התנאי על הנגזרות הראשונות ייתן:

$$\nabla f(x_1, x_2) = \begin{bmatrix} 3x_1^2 - 2x_1 x_2 \\ -x_1^2 + 4x_2 \end{bmatrix} = 0 .$$

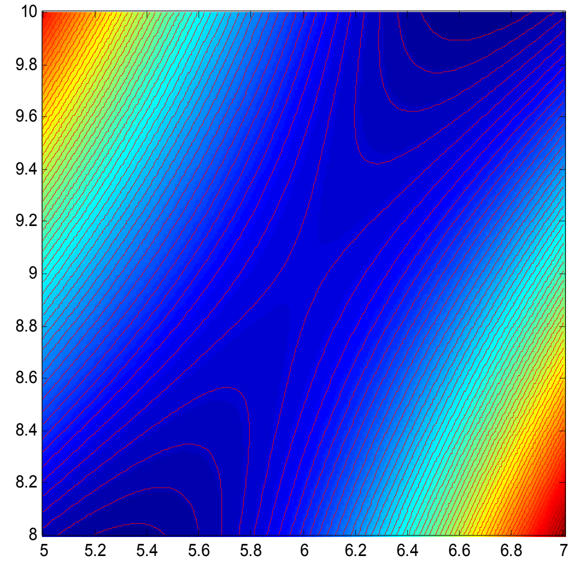
נקודות אפשריות למינימום הינן לפיכך: $x_1 = 0 , x_2 = 0$ וכן - $x_1 = 6 , x_2 = 9$. ההסיין בנקודות אלו נתון ע"י:

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} 6x_1 - 2x_2 & -2x_1 \\ -2x_1 & 4 \end{bmatrix} \Rightarrow \begin{cases} x_1 = 0 , x_2 = 0 \quad \Rightarrow \quad \mathbf{H} = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} \geq 0 \\ \\ x_1 = 6 , x_2 = 9 \quad \Rightarrow \quad \mathbf{H} = \begin{bmatrix} 18 & -12 \\ -12 & 4 \end{bmatrix} \end{cases}$$

הנקודה הראשונה, $x_1 = 0 , x_2 = 0$ <u>יכולה</u> להיות נקודת מינימום כיוון שהגרדיאנט בה מתאפס, וההסיין חיובי חצי מוגדר. נשים לב שאיננו יכולים לקבוע בוודאות אם אמנם זו נקודת מינימום לוקלית, ומי שיקבע בפועל הן הנגזרות המתקדמות יותר של הפונקציה. אזור זה של הפונקציה מתואר בציור הבא:



הנקודה $x_1 = 6 , x_2 = 9$ היא בודאי לא נקודת מינימום כיוון שערכיה העצמיים הם 2- ו-24. נקודה זו גם לא נקודת מקסימה כיוון שלצורך זה נדרש שההסיין יהיה שלילי חצי מוגדר (כלומר שמינוס מטריצת ההסיין תהיה חיובית חצי מוגדרת). אזור זה מתואר בציור הבא:



תופעה מעניינת עבור הפונקציה הנ"ל היא שבנקודה $x_1 = 6 , x_2 = 9$, קביעת אחד המשתנים ומינימיזציה עפ"י השני נותן כי בשני הצירים ישנה נקודת מינימום בנקודה זו. ועם זאת - אין זאת נקודת מינימום של הפונקציה הדו-ממדית. עובדות אלו ניכרות מהנגזרות שחושבו עד כה - הגרדיאנט מתאפס בנקודה, ונגזרות שניות בכוונים האנכיים הם 18 ו- 4 בהתאמה - מה שמבטיח מינימום. קווי הגובה של הפונקציה בציור הנ"ל ממחישים כיצד הדבר יכול לקרות.

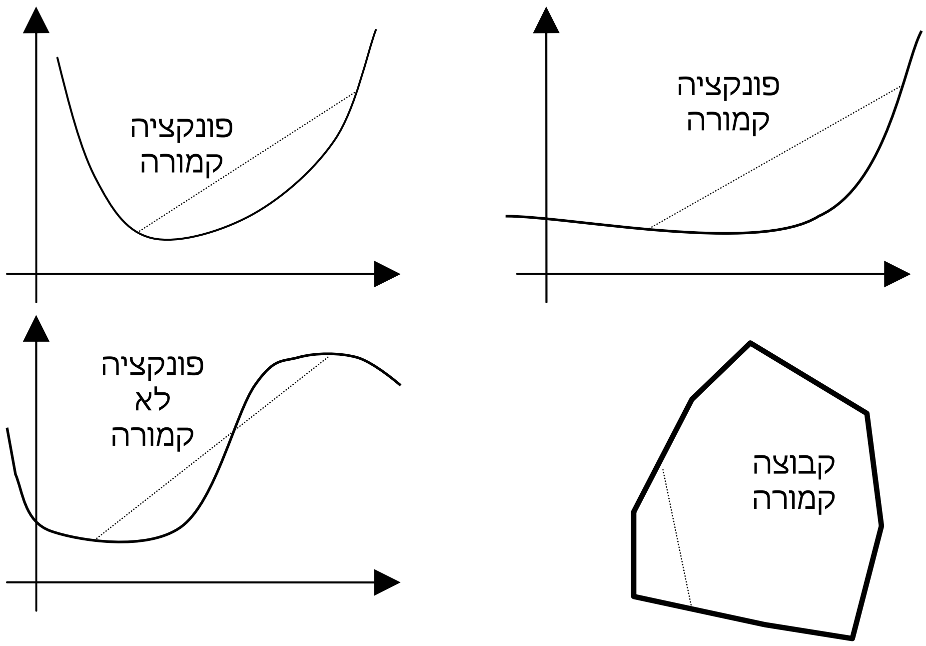
3. קבוצות קמורות ופונקציות קמורות

מבין כלל הפונקציות בהן נתעניין למינימיזציה, יש לנו עניין עצום בפונקציות קמורות (convex). נתחיל באפיונן, ואז נראה מדוע הן כה מעניינות. נתחיל בהגדרה של קבוצה קמורה: קבוצה Ω מוגדרת כקמורה אם לכל שתי נקודות $\underline{x}_1, \underline{x}_2 \in \Omega$ מתקבל כי הקו המתוח ביניהן אף הוא ב-Ω, כלומר, לכל $0 \le \theta \le 1$ מתקיים $\theta\underline{x}_1 + (1-\theta)\underline{x}_2 \in \Omega$.

כשעוברים לפונקציות, פונקציה $f(\underline{x})$ מעל קבוצה קמורה Ω מוגדרת כפונקציה קמורה אם לכל שתי נקודות $\underline{x}_1, \underline{x}_2 \in \Omega$ מתקבל כי:

$$f(\theta\underline{x}_1 + (1-\theta)\underline{x}_2) \le \theta f(\underline{x}_1) + (1-\theta) f(\underline{x}_2).$$

הנה מספר דוגמאות לפונקציות ולשאלת קמירותן



לעיתים קרובות נקבל צירופים ליניאריים של פונקציות קמורות. מסתבר כי קיימת התוצאה הבאה (קל להוכיחה מתוך ההגדרה):

משפט 2: תהיינה הפונקציות $f_1(\underline{x})$ ו- $f_2(\underline{x})$ פונקציות קמורות מעל תחום הגדרה קמור Ω, ויהיו הסקלרים a_1, a_2 אי-שליליים כלשהם. אזי, הפונקציה $a_1 f_1(x) + a_2 f_2(x)$ גם היא קמורה מעל Ω.

מה הקשר בין קבוצות ופונקציות קמורות? מסתבר שקשר כזה קיים ודי פשוט להוכיחו (שוב - מתוך ההגדרה):

משפט 3: תהי נתונה פונקציה קמורה $f(\underline{x})$ מעל קבוצה קמורה Ω. אזי הקבוצה הנתונה ע"י הביטוי $\Gamma = \{\underline{x} | f(\underline{x}) \le c\}$ היא קבוצה קמורה.

כעת נדון בהתנהגותן של נגזרות של פונקציות קמורות. עפ"י תכונות אלה נוכל לדון לאחר מכן במיקומן של נקודות מינימום ומקסימום בפונקציות כאלה.

משפט 4: תהי $f(\underline{x}) \in C^1$. אזי פונקציה זו קמורה מעל תחום הגדרה קמור Ω אם ורק אם מתקיים כי לכל $\underline{x}, \underline{y} \in \Omega$:

$$f(\underline{y}) \geq f(\underline{x}) + \nabla f(\underline{x})(\underline{y} - \underline{x}).$$

הוכחת משפט זה מחייבת טיפול בשני כיוונים – יש להראות כי אם אמנם הפונקציה קמורה, אזי הדבר גורר התכונה הנ"ל, ולהיפך – אם התכונה מתקיימת, קמירות נובעת ישירות. בכיוון הראשון, נניח שהפונקציה קמורה, ולכן ניתן לכתוב עבור כל $0 < \alpha \leq 1$

$$f(\alpha \underline{y} + (1-\alpha)\underline{x}) \leq \alpha f(\underline{y}) + (1-\alpha) f(\underline{x}).$$

סידור איברים בצורה שונה מוביל ל:

$$\frac{f(\alpha \underline{y} + (1-\alpha)\underline{x}) - f(\underline{x})}{\alpha} = \frac{f(\underline{x} + \alpha(\underline{y} - \underline{x})) - f(\underline{x})}{\alpha} \leq f(\underline{y}) - f(\underline{x}).$$

עם הבאת $\alpha \to 0$ מתקבלת טענת הלמה כיוון ש:

$$\lim_{\alpha \to 0} \frac{f(\underline{x} + \alpha(\underline{y} - \underline{x})) - f(\underline{x})}{\alpha} = \nabla f^T(\underline{x})(\underline{y} - \underline{x}) \leq f(\underline{y}) - f(\underline{x}).$$

באופן זה הוכחנו כיוון אחד. כדי להוכיח את הכיוון ההפוך נניח כי הקשר $f(\underline{y}) \geq f(\underline{x}) + \nabla f(\underline{x})(\underline{y} - \underline{x})$ נכון, ונראה כי נובע ממנו שהפונקציה קמורה. תהיינה $\underline{x}_1, \underline{x}_2 \in \Omega$ שתי נקודות כלשהן בתחום ההגדרה, ויהיה α סקלר בעל ערך שרירותי באינטרוול [0,1]. קביעת הנקודה $x = \alpha \underline{x}_1 + (1-\alpha)\underline{x}_2$ ושימוש באי-השוויון אשר הונח כנכון נותן כי:

$$\begin{aligned} f(\underline{x}_1) &\geq f(\underline{x}) + \nabla f(\underline{x})(\underline{x}_1 - \underline{x}) \\ f(\underline{x}_2) &\geq f(\underline{x}) + \nabla f(\underline{x})(\underline{x}_2 - \underline{x}) \end{aligned}$$

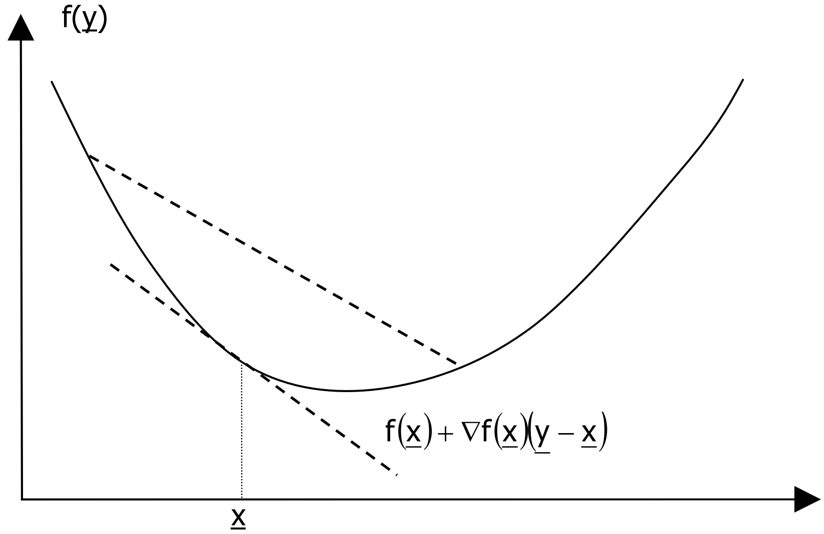
הכפלת המשוואה העליונה ב- α והשנייה ב- $(1-\alpha)$ וסיכומן ייתן את התוצאה הרצויה:

$$\begin{aligned} &\alpha f(x_1) + (1-\alpha) f(x_2) \geq f(x) + \alpha \nabla f(x)(x_1 - x) + (1-\alpha)\nabla f(x)(x_2 - x) \\ &\Rightarrow \alpha f(x_1) + (1-\alpha) f(x_2) \geq f(x) \end{aligned}$$

ולכן המשפט הוכח. בציור הבא מתוארת טענת משפט זה. אם קודם קבענו כי חיבור שתי נקודות מניב ישר המצוי מעל הפונקציה, כאן אנו רואים כי מתיחת קו המקרב מקומית את הפונקציה ע"י ישר משיק מניבה קו אשר הינו תמיד מתחת לפונקציה.

מסתבר כי בעזרת התכונה הקודמת ניתן להציע תכונה אחרת חשובה יותר. הנוגעת בהתנהגות ההסיין של פונקציות קמורות. ע"י קירוב טיילור נוכל לכתוב כי לכל $\underline{y}$ מתקיים:

$$f(\underline{y}) = f(\underline{x}) + \nabla f(\underline{x})(\underline{y}-\underline{x}) + \frac{1}{2}(\underline{y}-\underline{x})^T \mathbf{H}(x)(\underline{y}-\underline{x}),$$



כשאנו בכוונה עוסקים בנקודות y כה קרובות ל-x כך שאיבר השגיאה בטור טיילור הופך זניח. אם ההסיין חיובי מוגדר בכל נקודה נקבל

$$\frac{1}{2}(\underline{y}-\underline{x})^T \mathbf{H}(x)(\underline{y}-\underline{x}) > 0 \;\Rightarrow\; f(\underline{y}) < f(\underline{x}) + \nabla f(\underline{x})(\underline{y}-\underline{x}),$$

ופירוש הדבר שהפונקציה קמורה ממש. את אותו רציונל ניתן ליישם גם בכיוון ההפוך ולהראות שאם הפונקציה קמורה ממש, בהכרח ההסיין חיובי מוגדר. לכן יש לנו את המשפט הבא:

משפט 5: תהי $f(\underline{x}) \in C^2$. אזי פונקציה זו קמורה ממש מעל תחום הגדרה קמור Ω אם ורק אם מתקיים כי ההסיין של פונקציה זו - $\mathbf{H}(\underline{x})$ - חיובי מוגדר בכל נקודה בתחום ההגדרה.

כל זה טוב ויפה, אך כעת נגיע לעניין העיקרי הוא מינימיזציה של פונקציות קמורות:

משפט 6: תהי נתונה פונקציה קמורה $f(\underline{x})$ מעל קבוצה קמורה Ω. אזי, תת-הקבוצה $\Gamma \subseteq \Omega$ בה הפונקציה מקבלת את ערכי המינימום היא קבוצה קמורה, וכל נקודת מינימום מקומית היא גם נקודת מינימום גלובלית של הפונקציה.

קל להוכיח משפט זה – אם נניח כי ערך המינימום של הפונקציה f הוא c_0, אז נוכל לכתוב את תת-קבוצת נקודות המינימום ע"י הביטוי $\Gamma = \{\underline{y} \mid f(\underline{y}) \le c_0\}$, וכבר ראינו קודם

כי זוהי קבוצה קמורה. מעבר לכך, אם נניח כי הנקודה $\underline{x}_1$ היא נקודת מינימום לוקלית, וקיימת נקודה אחרת - $\underline{x}_2$ - בה הפונקציה משיגה את המינימום הגלובלי שלה, פירוש הנחה זו הוא ש - $f(\underline{x}_2) < f(\underline{x}_1)$. בשל היות הפונקציה קמורה נוכל לכתוב עבור כל $\alpha \in (0,1)$ כי:

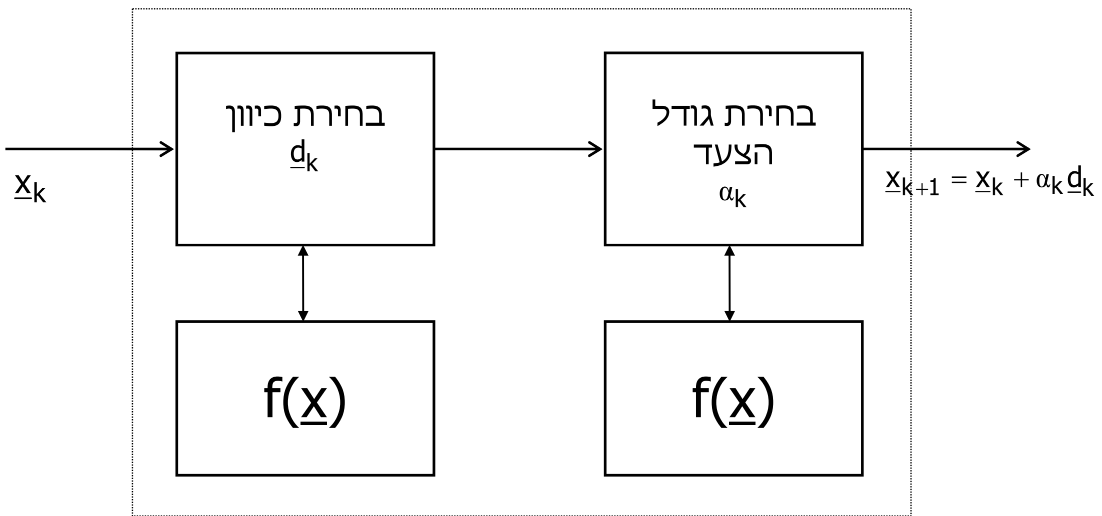
$$\begin{aligned} & f(\underline{x}_1 + \alpha(\underline{x}_2 - \underline{x}_1)) \le (1-\alpha)f(\underline{x}_1) + \alpha f(\underline{x}_2) \le \\ & \le f(\underline{x}_1) + \alpha[f(\underline{x}_2) - f(\underline{x}_1)] < \max\{f(\underline{x}_1), f(\underline{x}_2)\} \le f(\underline{x}_1) \end{aligned}.$$

עם $\alpha \in (0,1)$ קרוב מאוד לאפס נתקרב לנקודה $\underline{x}_1$ כרצוננו, ועבור נקודה זו נשיג ערך נמוך יותר לפונקציה. זה סותר את ההנחה המקורית, ולכן ברור כי נקודה $\underline{x}_1$ היא גם נקודת מינימום גלובלי.

לסיכום חלק זה, אם התמזל מזלנו ואנו עוסקים במינימיזציה של פונקציה קמורה, אין לנו חששות ביחס לנקודות מינימום לוקליות. כך היה המקרה עם הפונקציה הריבועית בה ההסיין לא רק חיובי מוגדר אלא קבוע במקום. מסתבר כי במקרים רבים בעיות הנדסיות ניתנות לתיאור כבעיות מינימיזציה קמורות, ואז חיינו קלים לאין ערוך.

4. מינימיזציה חד-ממדית – שיטות חיפוש על ישר

נניח כי אנו נמצאים בנקודה כלשהי $\underline{x}$ המוצעת כפתרון, ורצונו לעדכנה. כמעט כל אלגוריתם באופטימיזציה זאת באיטרציות, כשכל איטרציה בת שני שלבים - תחילה ייבחר כיוון $\underline{d}$ (עם סימן) בו האלגוריתם מוצא כי כדאי לנוע, ובשלב שני ייקבע המרחק האופטימלי (או תת-אופטימלי) בו יש ללכת כדי להשיג ירידה מרבית. הן חישוב הכיוון והן מציאת המרחק יבוצעו ע"י פניה לפונקציה ל-"התייעצות". רעיון זה מתואר בציור הבא. כל אלגוריתם בעל מבנה כזה ישתמש שוב ושוב (בכל איטרציה פעם אחת) ברוטינה המבצעת חיפוש מינימום על ישר. כעת נדון במרכיב זה.



למעשה, אינטרפרטציה אחרת לתיאור הנ"ל היא הבאה: בהתקבל בעיית מינימיזציה בה יש נעלמים רבים, אנו מציעים את המרתה בפתרון בעיות רבות של מינימיזציה חד-ממדית, וכעת אנו מתמקדים בשאלה כיצד מבצעים מינימיזציה כזו. אנו נתמקד בשתי דרכים, ויש רבות אחרות. נתחיל עם שיטת ניוטון אשר הינה פופולארית.

נתונה פונקציה f חד ממדית אותה רצוננו להביא למינימום, ונניח כי נתונה נקודה בה אנו מצויים כעת - x_k - עבורה ידועים ערכה של הפונקציה ונגזרותיה מסדר ראשון ושני - $f(x_k), f'(x_k), f''(x_k)$. נקרב את הפונקציה במיקום זה לפונקציה ריבועית ונקבל את הקירוב:

$$q(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2} f''(x_k)(x - x_k)^2 .$$

נקבע את נקודת המינימום המשוערת כנקודת המינימום של q, דהיינו,

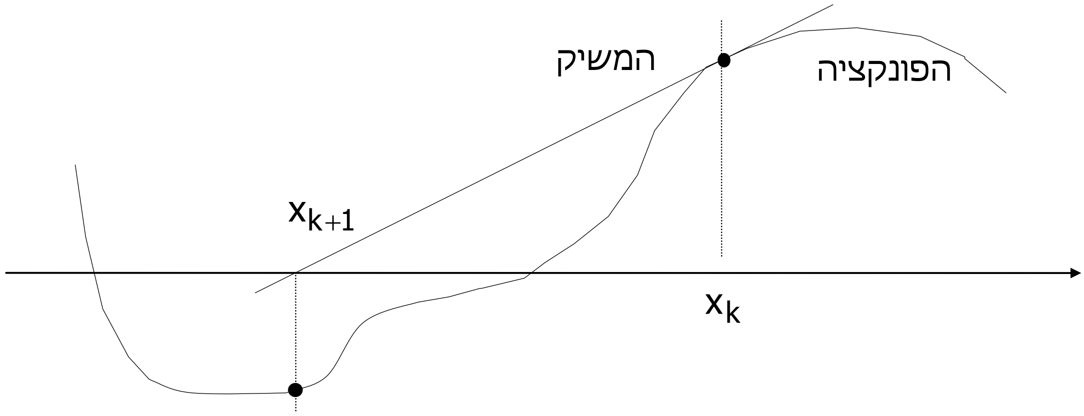
$$q'(x_{k+1}) = 0 = f'(x_k) + f''(x_k)(x_{k+1} - x_k) \Rightarrow x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

יש לשים לב לעובדה שלערך הפונקציה במיקום x_k אין השפעה על קביעת המיקום הבא. כן יש לשים לב לעובדה שאלגוריתם זה אשר במקורו נועד לאיתור נקודת מינימום של פונקציה f(x), יכול להתפרש כאלגוריתם לחיפוש פתרון של המשוואה $q'(x) = 0$. נראה זאת באופן הבא. בהינתן פונקציה g(x) רצוננו למצוא את נקודת התאפסותה - g(x)=0. נניח כי הקירוב הנתון בידינו הוא x_k. נעביר משיק לפונקציה ב-x_k, ונקבע את הנקודה הבאה - x_{k+1} - להיות הנקודה בה המשיק חוצה את האפס (משוואה ליניארית). משוואת המשיק נתונה ע"י

$$L(x) = f'(x_k)(x - x_k) + f(x_k) \quad \Rightarrow \quad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

אלגוריתם זה מוכר בשם "אלגוריתם ניוטון-רפסון", והוא זהה לאלגוריתם ניוטון המוצג כאן. השוני הוא התפיסה - בעוד שאלגוריתם ניוטון חושב במונחים של מינימיזציה, הרי שאלגוריתם ניוטון-רפסון חושב במונחים של פתרון משוואה (מתואר בציור) - שתי בעיות שקולות, כיוון שנקודת המינימום של הבעיה הראשונה היא נקודת ההתאפסות של נגזרת הפונקציה.

אין למעשה כל הבטחה שאלגוריתם ניוטון יתכנס! אם הנגזרת השנייה שלילית בנקודה אליה הגענו - אנו בצרות! עם זאת, אם אנו מצויים בקרבת נקודת המינימום האמיתית, והפונקציה חלקה דיה, האלגוריתם יתכנס ודי מהר. ניתן לשפר את אלגוריתם ניוטון במגוון דרכים, אך אנו לא נמשיך להעמיק בכך.



באופן דומה לאלגוריתם ניוטון נציע אלגוריתם אשר מתבסס על ידיעת ערך הפונקציה בנקודה x_k, וידיעת ערכי הנגזרות הראשונות בשתי נקודות אחרונות - $f'(x_k), f'(x_{k-1})$.

הפונקציה הריבועית אשר מקיימת נתונים אלו תהיה:

$$q(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2} \cdot \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}}(x - x_k)^2$$

והדמיון לשיטה הקודמת ניכר. מציאת נקודת המינימום של $q(x)$ נותנת:

$$x_{k+1} = x_k - \frac{(x_k - x_{k-1}) \cdot f'(x_k)}{f'(x_k) - f'(x_{k-1})}$$

בדומה לאלגוריתם ניוטון, העדכון של הפתרון אינו מסתמך על ערכה של הפונקציה בנקודה, אלא בנגזרותיה בלבד.

ניתן להוכיח כי קצב ההתכנסות של אלגוריתם זה איטי יותר מהקודם. בדומה לאלגוריתם ניוטון, אלגוריתם זה עלול להתבדר, לשוטט ללא התכנסות או להתנדנד - תלוי בבחירת נקודת ההתחלה, והפונקציה עליה אנו עובדים. בדומה לאלגוריתם ניוטון, גם כאן ניתן להציע וריאציות יציבות יותר.

גישה אחרת לגמרי לחיפוש על ישר היא גישה מקורבת שאינה מציעה הנחת מודל פרמטרי על הפונקציה לאורך הישר. אחת השיטות הפופולאריות ביותר היא אלגוריתם Armijo הפועלת באופן הבא – כשאנו בנקודה x_k, קירוב טיילור נותן לנו את הקשר

$$f(x_k + \delta) \approx f(x_k) + f'(x_k)\delta,$$

אשר אמור להיות קירוב טוב בסביבת הנקודה x_k. שיטת Armijo דורשת לפחות חלק α משיעור הירידה המנובא ע"י הנגזרת. לכן, נבחר $0 < \alpha < 1$ (בד"כ 0.5-0.8), ונחפש גודל צעד δ גדול ככל האפשר שיקיים

$$f(x_k + \delta) \le f(x_k) + \alpha \cdot f'(x_k)\delta.$$

החיפוש נעשה באופן איטרטיבי – מתחילים ב- δ_0 ובוחנים האם צעד זה מקיים את אי-השוויון הנ"ל. אם לאו, נקטין את גודל הצעד, $\delta_1 = \alpha\delta_0$, ונחזור על הבדיקה. כך ימשיך התהליך כשגודל הצעד קטן לפי $\delta_j = \alpha^j\delta_0$. כשאנו מתקרבים ל- $\delta = 0$ אנו חייבים למצוא נקודה ראויה כיוון שבקרבת x_k הנגזרת מכתיבה ירידה גדולה יותר.

5. שיטות בסיסיות למינימיזציה איטרטיבית (SD וניוטון)

אם אנו יודעים לבצע חיפוש על ישר, כל שנותר לשם השלמת אלגוריתם האופטימיזציה היא קביעת הכיוון בו יש ללכת. כבר פגשנו בפרק קודם את שיטת השיפוע המרבי (SD). שיטה זו מציעה ללכת בכיוון הפוך לגרדיאנט הפונקציה בנקודה, דהיינו

$$\underline{x}_{k+1} = \underline{x}_k - \mu_k \nabla f(\underline{x}_k).$$

עבור הבעיה הריבועית ראינו כי קביעת μ_k קלה בשל היות החתך של הפונקציה פרבולה מדויקת. עבור פונקציות מורכבות יותר מרכיב זה יוחלף באחת משיטות החיפוש על ישר שהוזכרו קודם. לדוגמה, נוכל לחשב את ההסיין בנקודה $\underline{x}_k$, לאמוד את μ_k כאילו לפנינו בעיה ריבועית,

$$\mu_k = \frac{\nabla f(\underline{x}_k)^T \nabla f(\underline{x}_k)}{\nabla f(\underline{x}_k)^T \mathbf{H}(\underline{x}_k) \nabla f(\underline{x}_k)},$$

ולהשתמש בערך המתקבל כאתחול לשיטת Armijo.

ראינו כי מיצוי תהליך החיפוש על ישר בבעיה הריבועית מוביל לאלגוריתם ה-Normalized SD, וזה יוצר תנועת זיג-זג שאינה כה יעילה. מסתבר כי תכונה זו נכונה גם לפונקציות כלליות. נניח כי מהנקודה $\underline{x}_k$ התקדמנו בצעד

$$\underline{x}_{k+1} = \underline{x}_k - \mu_k \nabla f(\underline{x}_k)$$

כך ש- μ_k נותן מינימום על הישר. פירוש הדבר ש-

$$0 = \frac{d}{d\mu} f(\underline{x}_k - \mu \nabla f(\underline{x}_k)) = \nabla f(\underline{x}_k)^T \nabla f(\underline{x}_k - \mu \nabla f(\underline{x}_k)) = \nabla f(\underline{x}_k)^T \nabla f(\underline{x}_{k+1}),$$

כלומר, גרדיאנטים עוקבים שהינם כיווני התנועה של האלגוריתם יהיו אנכיים זה לזה, ושוב תתקבל תופעת הזיג-זג. לכן, גם כאן ניתן להציע את שיטת ה- PARTAN או את שיטות ה- Conjugate Gradient הנובעת ממנה, כשנוסחאות סגורות של גדלי צעד מומרים בחיפוש על ישר.

האם הגרדיאנט הינו הכיוון היחיד הבא בחשבון בבואנו לקבוע את תנועת האלגוריתם? נניח כי במקום, האלגוריתם למינימיזציה יבחר את כיוון התנועה $\underline{d}$ להיות מכפלה של מטריצה חיובית מוגדרת כלשהי במינוס וקטור הגרדיאנט, דהיינו,

$$\underline{d}_k = -\mathbf{M}\nabla f(\underline{x}_k) \quad \mathbf{M} > 0 \;\Rightarrow\; \underline{x}_{k+1} = \underline{x}_k - \alpha_k \mathbf{M}\nabla f(\underline{x}_k).$$

האם כיוון זה ייתן הבטחה לירידה? האינטואיציה אומרת שכן כיוון שהכפלה במטריצה חיובית מוגדרת כמו הכפלה בסקלר חיובי. אם נכתוב את **M** בתלות בוקטורים ובערכים העצמיים שלה נקבל כי הכיוון שייקבע ע"י הכפלה ב- **M** הוא:

$$\mathbf{M} = \sum_{k=1}^{n} \lambda_k \underline{u}_k \underline{u}_k^T \;\Rightarrow\; \mathbf{M}\underline{v} = \sum_{k=1}^{n} \lambda_k \left(\underline{v}^T \underline{u}_k\right) \underline{u}_k$$

בעוד שהכיוון ללא הכפלה ב- **M** ניתן לתיאור כ:

$$\underline{v} = \sum_{k=1}^{n} \left(\underline{v}^T \underline{u}_k\right) \underline{u}_k,$$

כשאנו עושים שימוש בעובדה שהוקטורים העצמיים $\underline{u}_k$ אורתונורמליים. לכן, משמעות הכפלה ב- **M** היא משקול שונה של מרכיבי הכיוון ללא פגיעה בכיוון הכללי. נטען זאת באופן יותר פורמאלי.

משפט 7: בחירת הכיוון $\underline{d}_k = -\mathbf{M}\nabla f(\underline{x}_k)$ עם מטריצה **M** חיובית מוגדרת מבטיחה עבור $\alpha_k > 0$ קטן דיו ירידה בפונקציה f.

למעשה, שימוש בקירוב טיילור מסדר ראשון יאפשר לנו לראות מדוע זה נכון. פיתוח הפונקציה לטור טיילור סביב $\underline{x}_k$ והצבת הכיוון המוצע נותנים

$$\begin{aligned} f(\underline{x}_{k+1}) &= f(\underline{x}_k) + \nabla f(\underline{x}_k)^T (\underline{x}_{k+1} - \underline{x}_k) + o\{\|\underline{x}_{k+1} - \underline{x}_k\|\} = \\ &= f(\underline{x}_k) + \alpha_k \nabla f(\underline{x}_k)^T \mathbf{M} \nabla f(\underline{x}_k) + \alpha_k \cdot o\{\|\mathbf{M}\nabla f(\underline{x}_k)\|\} \end{aligned}.$$

עבור $\alpha_k \to 0$ ברור כי חלה כאן ירידה כיוון ש - $\nabla f(\underline{x}_k)^T \mathbf{M} \nabla f(\underline{x}_k) > 0$ בשל היות **M** חיובית מוגדרת.

בשל האמור לעיל, יש לנו עניין להציע הכפלה של הגרדיאנט במטריצה חיובית מוגדרת על מנת להשיג מהאלגוריתם יותר. אלגוריתם מסוג זה אשר הינו מאוד פופולארי הוא אלגוריתם ניוטון, בו הבחירה היא $\alpha_k = 1$ וקביעת **M** להיות היפוך ההסיין,

$$\mathbf{M}_k = \mathbf{H}^{-1}(\underline{x}_k).$$

עבור בעיה ריבועית מהצורה

$$p(\underline{x}) = \underline{x}^T \mathbf{K} \underline{x} - \underline{f}^T \underline{x} + c$$

אלגוריתם ניוטון מתכנס באיטרציה אחת, וללא תלות באתחול. זאת כיוון ש-

$$\underline{x}_1 = \underline{x}_0 - \mathbf{K}^{-1}(\mathbf{K}\underline{x}_0 - \underline{f}) = \mathbf{K}^{-1}\underline{f}.$$

אין כאן כל הפתעה – אם אנו יכולים להפוך את המטריצה **K**, ברור שבעיה ריבועית תיפתר בחישוב ישיר. המסר כאן הוא אחר – בבעיות אופטימיזציה שאינן ריבועיות ועם כמות משתנים שאינה גדולה, שימוש באלגוריתם ניוטון יהיה מאוד יעיל. אלגוריתם זה יתייחס לפונקציה בכל נקודה כפונקציה ריבועית, ויביא לעדכון מיטבי ביחס לכך. אם הפונקציה בסביבת העבודה קרובה בהתנהגותה לפונקציה ריבועית, יושג הפתרון במעט איטרציות – הרבה פחות מהכמות הנדרשת ל-SD. אגב – אלגוריתם זה עבור פונקציה חד-ממדית יוביל אותנו לשיטת החיפוש על הישר הראשונה בא דננו.

לשיטת ניוטון חסרון מרכזי אחד - בשל היותה חזקה מאוד בהתכנסותה, היא גם עלולה להניב התנהגות פרועה למדי אם היא אינה מאותחלת בנקודה הקרובה ליעד. לפיכך נדרש ריסון לשיטה אשר יפגע קלות בקצב ההתכנסות על חשבון יציבות גבוהה יותר של התהליך האיטרטיבי. המקורות לאי יציבותה של שיטת ניוטון המקורית הם הבאים:

- מטריצת ההסיין לבעיה הכללית עלולה לא להיות חיובית מוגדרת. במקרה כזה כבר ראינו בתחילת פרק זה כי הכיוון המבוצע אינו מחייב כלל ירידה.
- מטריצת ההסיין עלולה להיות סינגולארית - עובדה שפירושה שלא קיימת מטריצה הפוכה לה.
- קצב התכנסות מהיר יושג עבור קירוב ריבועי. כאשר הפונקציה שונה באופן מהותי מקירוב זה החיפוש האיטרטיבי עלול ליצור שיטוט של החיפוש ללא התכנסות.

כדי להתגבר על בעיות אלו הוצעו מגוון שינויים. אנו נציג במסגרת זו שני אלגוריתמים אלטרנטיביים שהינם וריאציות מרוסנות של גישת ניוטון. האלגוריתם הראשון נקרא השיטה המשולבת (היברידית) בשל היותו שילוב של ה-NSD והגישה הניוטונית. השינויים המהותיים בו הם (א) שינוי מטריצת ההסיין ע"י תוספת של מטריצת יחידה מוכפלת בקבוע

$$\mathbf{M}_k = \left[\lambda\mathbf{I} + \mathbf{H}(\underline{x}_k)\right]^{-1}.$$

את λ נקבע כך שיבטיח כי התוצאה היא מטריצה חיובית מוגדרת. שינוי זה נועד להבטיח שתתקבל מטריצה חיובית מוגדרת (ולכן לא סינגולארית) **M**. בנוסף, (ב) מבוצע חיפוש על ישר ולא נקבע כבאלגוריתם ניוטון כי $\alpha_k = 1$, דהיינו

$$\underline{x}_{k+1} = \underline{x}_k - \alpha_k\left[\lambda\mathbf{I} + \mathbf{H}(\underline{x}_k)\right]^{-1}\nabla f(\underline{x}_k).$$

האלגוריתם אחר הנובע מהשיטה הניוטונית קרוי אלגוריתם Levenberg-Marquardt. אלגוריתם זה דומה מאוד, מוסיף ניואנס מעניין – קביעת λ נעשית בד-בבד עם פירוק Cholesky למטריצה $\lambda\mathbf{I} + \mathbf{H}(\underline{x}_k) = \mathbf{L}\mathbf{L}^T$ (**L** משולשת תחתונה). כל פעם שהאלגוריתם

נתקע יוגדל ה- λ. בסיום התהליך הרווח כפול – מעבר לקביעת λ ראוי, פירוק זה מקל על חישוב גורם העדכון

$$\underline{v} = \left(\mathbf{LL}^T\right)^{-1}\nabla f\left(\underline{x}_k\right) \;\Rightarrow\; \mathbf{LL}^T\underline{v} = \nabla f\left(\underline{x}_k\right),$$

ולכן בשני תהליכי החלפה (אחורי וקדמי), ייקבע כוון העדכון.

אלגוריתם אחרון אותו נציג במסגרת זו הוא אלגוריתם ה- Diagonal Normalized SD. אלגוריתם ה - DNSD הינו הכללה של אלגוריתם ה- NSD, ובו במקום פרמטר יחיד לקביעת גודל צעד העדכון, ישנם n פרמטרים - אחד כנגד כל מרכיב של וקטור העדכון. במקרה הפרטי בו כל ערכים אלו שווים, אנו מקבלים כמקרה פרטי את ה-NSD. נכתוב את משוואת האלגוריתם הכללית עבור הבעיה הריבועית:

$$\underline{x}_{k+1} = \underline{x}_k - \mathbf{D}_k\nabla f\left(\underline{x}_k\right) = \underline{x}_k + \mathbf{D}_k\underline{e}_k = \underline{x}_k + \begin{bmatrix}\alpha_k^1 & & 0\\ & \ddots & \\ 0 & & \alpha_k^n\end{bmatrix}\underline{e}_k.$$

גישה מקובלת לקביעת מטריצת המשל האלכסונית **D** היא שימוש באלכסון הראשי של ההסיין,

$$\mathbf{D}_k = \begin{bmatrix}\left[\dfrac{\partial^2 f}{\left(\partial x_k^1\right)^2}\right]^{-1} & & 0\\ & \ddots & \\ 0 & & \left[\dfrac{\partial^2 f}{\left(\partial x_k^1\right)^2}\right]^{-1}\end{bmatrix}$$

שימוש במטריצה זו יחד עם חיפוש על ישר יביאו לאלגוריתם Jacobi כאשר מדובר בבעיה ריבועית.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 8 – אופטימיזציה עם אילוצים

נושאים שייסקרו:

1. בעיות עם אילוצי שוויון – כופלי לגרנז'
2. בעיות עם אילוצי אי-שוויון - תנאי KKT
3. שיטת המחיר לטיפול באילוצים
4. שיטת המחסום לטיפול באילוצים
5. שיטת ההטלה לטיפול באילוצים
6. שיטות מבוססות Lagrange לטיפול באילוצים
7. קמירות בבעיות עם אילוצים
8. בעיות תכנות ליניארי

---

1. בעיות עם אילוצי שוויון – כופלי לגרנז'

כשעסקנו בבעיות נטולות אילוצים, ראינו כי גרדיאנט הפונקציה ומטריצת ההסיין שלה בנקודה נתונה משמשות ככלים חשובים בקביעת אופייה של הנקודה, ובדרך ההתקדמות ממנה עבור אלגוריתמים איטרטיביים שונים. בפרק זה נכליל תוצאות אלה עבור בעיות עם אילוצים. הבעיה הכללית ביותר בה נטפל כעת היא

$$\min_{\underline{x}} \; f(\underline{x}) \quad \text{S.T.} \quad \{h_k(\underline{x}) = 0\,, \quad 1 \le k \le m\}$$

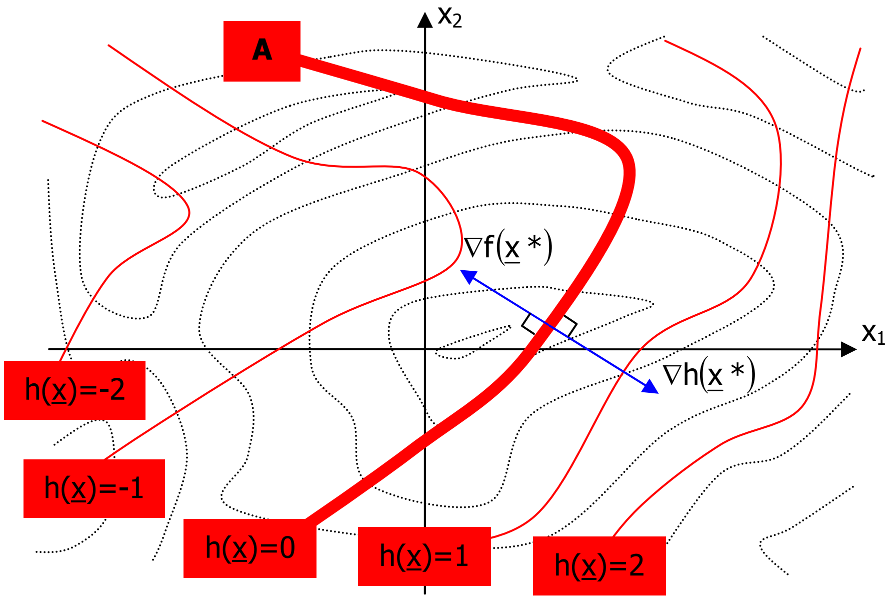
כאשר $\underline{x}$ הוא וקטור של n נעלמים, ולבעיה יש m אילוצי שוויון – ברור כי חייב להתקיים m<n (אחרת האילוצים לבדם עשויים להכתיב את התוצאה ללא מינימיזציה). לעתים נשתמש ברישום מקוצר:

$$H(\underline{x}) = \begin{bmatrix} h_1(\underline{x}) \\ h_2(\underline{x}) \\ \vdots \\ h_m(\underline{x}) \end{bmatrix} = \underline{0}\,.$$

במהלך הדיון הקרוב נציג את "כופלי לגרנז' ", ונראה כיצד אלה משמשים ביעילות לפתרון הבעיה הנ"ל. כמקודם, נניח כי הפונקציה $f(\underline{x})$ וכן פונקציות האילוצים רציפות וגזירות ברציפות פעמיים לפחות, $f(\underline{x}), \{h_k(\underline{x})\}_{k=1}^{m} \in C^2$, ואז נוכל להכליל תוצאות קודמות המתייחסות לגרדיאנט ולהסיין של הפונקציה והקשר שלהם לפתרון הבעיה.

נציג את הדברים באופן אינטואיטיבי עבור בעיית אופטימיזציה בדו-ממד: בציור הבא מתוארים קוים שווי גובה של הפונקציה $f(\underline{x})$ ופונקצית אילוץ $h(\underline{x})$ למקרה דו-מימדי - $\underline{x} \in \Re^2$. קווי הגובה של פונקצית האילוץ מתוארים באדום, והקו בו האילוץ מתקיים, $h(\underline{x})=0$, מתואר כקו עבה A. על קו זה ורק עליו אנו מחפשים את נקודת הפתרון $\underline{x}^*$.

כיוון ש- A הוא קו שווה גובה של פונקציית האילוץ, בכל נקודה בו יתקיים כי הגרדיאנט $\nabla h(\underline{x})$ ניצב לקו. מאידך, אנו מסתכלים על ערכי הפונקציה $f(\underline{x})$ לאורך A ומחפשים מינימום. נניח שמצאנו את הנקודה $\underline{x}^*$ בה מתקבל המינימום של $f(\underline{x})$. אם נסתכל בסביבה מאוד קרובה לנקודה זו לאורכו של A (מעט קדימה ואחורה מ- $\underline{x}^*$), נוכל להתייחס לנתח זה של A כקו ישר. בכיוון ישר זה הנגזרת של $f(\underline{x})$ מתאפסת, עובדה שפירושה שגרדיאנט הפונקציה $\nabla f(\underline{x})$ בנקודה זו אף הוא ניצב ל-A (למעשה, נתקלנו כבר בתופעה זו קודם כשדנו ב-NSD, וראינו שחיפוש על ישר ומיצוי האופטימיזציה עליו מבטיחה אנכיות של הישר והגרדיאנט בנקודת העצירה).



לכן, קיבלנו כי שני הגרדיאנטים הנ"ל מקבילים בנקודה $\underline{x}^*$. לכן, קיים סקלר כלשהו λ שייתן כי

$$\nabla f(\underline{x}^*) + \lambda \nabla h(\underline{x}^*) = 0 .$$

משמעות משוואה זו, באמירה מוכללת, היא ששני וקטורי גרדיאנט אלו תלויים ליניארית. כעת נטען זאת למקרה הכללי, אך לצורך כך נזדקק להגדרה של מטריצת היעקוביאן של האילוצים, שהיה בגודל של n שורות ו-m עמודות,

$$\nabla H(\underline{x}) = \begin{bmatrix} \frac{\partial h_1(\underline{x})}{\partial x_1} & \frac{\partial h_2(\underline{x})}{\partial x_1} & \cdots & \frac{\partial h_m(\underline{x})}{\partial x_1} \\ \frac{\partial h_1(\underline{x})}{\partial x_2} & \frac{\partial h_2(\underline{x})}{\partial x_2} & \cdots & \frac{\partial h_m(\underline{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_1(\underline{x})}{\partial x_n} & \frac{\partial h_2(\underline{x})}{\partial x_n} & \cdots & \frac{\partial h_m(\underline{x})}{\partial x_n} \end{bmatrix} .$$

אנו נניח כי דרגתה של מטריצה זו מלאה (כלומר עמודותיה בלתי תלויות ליניארית) בנקודת הפתרון $\underline{x} = \underline{x}^*$. אם אמנם זה מתקיים, הנקודה $\underline{x}^*$ נראית רגולארית.

נניח, אם כן, כי בידינו נקודה $\underline{x}^*$ רגולארית, ונניח כי היא פתרון בעיית האופטימיזציה שבידנו

$$\min_{\underline{x}} \; f(\underline{x}) \quad \text{S.T. } H(\underline{x}).$$

מטריצת היעקוביאן מגדירה סט של m כיוונים אפשריים במרחב n ממדי, אשר צעידה בהם לפי

$$\underline{x}_{new} = \underline{x}^* + \mu \nabla H(\underline{x}^*) \underline{v}$$

תוביל מידית להפרת אחד האילוצים או יותר (לכל בחירה של הסקלר μ ווקטור יחידה $\underline{v}$ שממשקל את עמודות היעקוביאן). זאת כיוון שלפי הגדרתם, גרדיאנטים אלו מצביעים לכיוון השינוי המרבי בערכי האילוצים. לעומת זאת, תנועה בכיוון אנכי לכל אלו (תת-מרחב בממד n-m) תבטיח מקומית המשך קיום האילוצים.

מה נוכל לומר על הגרדיאנט $\nabla f(\underline{x}^*)$? נניח כי הוא אנך לחלוטין לכל עמודות היעקוביאן $\nabla H(\underline{x}^*)$. פירוש הדבר שמשיקולי מינימיזצית הפונקציה, נוכל להמשיך ולנוע מ-$\underline{x}^*$ בתוך תת-מרחב בו האילוצים נשמרים, וזה סותר את קביעתנו המקורית ש-$\underline{x}^*$ הוא פתרון בעיית האופטימיזציה. באופן דומה, אם ל-$\nabla f(\underline{x}^*)$ מרכיב (לא כל הגרדיאנט, אלא חלקו) האנך לעמודות היעקוביאן, אותו שיקול יביא ליכולת להמשיך להוריד מגובה הפונקציה ללא חריגה מהאילוצים. לן, המסקנה היא ש-$\nabla f(\underline{x}^*)$ נפרש ליניארית ע"י עמודות אלה.

מדוע נחוצה רגולאריות? למשל, מה יקרה עם שתי עמודות של היעקוביאן זהות? כל שישתנה הוא שיש אילוץ מיותר וניתן להשמיטו – עניין טכני שלא מהווה מכשלה מהותית. נציג כעת את משפט Lagrange הנובע מהדיון הנ"ל:

משפט 1: (משפט כופלי לגרנז' – חלק 1). תהי $\underline{x}^*$ נקודת מינימום רגולארית של $f(\underline{x})$ המקיימת את האילוץ $H(\underline{x})=0$. אזי קיים וקטור אחד ויחיד $\underline{\lambda}^* = [\lambda_1^*, \; \cdots \;, \lambda_m^*]$ הנקרא וקטור כופלי לגרנז' כך שמתקיים:

$$\nabla f(\underline{x}^*) + \sum_{k=1}^{m} \lambda_k^* \nabla h_k(\underline{x}^*) = \nabla f(\underline{x}^*) + \nabla H(\underline{x}^*) \underline{\lambda}^* = \underline{0}.$$

הדיון שקדם למשפט נותן למעשה את ההוכחה. משמעות המשפט הנ"ל יכולה להתפרש באופן הבא - בנקודת המינימום הלוקלית מתקבל כי וקטור הגרדיאנט של פונקצית המחיר $f(\underline{x})$ נפרס ליניארית ע"י וקטורי גרדיאנט פונקציות האילוצים. חשוב לציין שכאשר אין אילוצים, גרדיאנט האילוצים הוא וקטור אפסים, ולכן תוצאת פרק קודם מתקבלת כמקרה פרטי של הטענה כאן.

למעשה קיבלנו כי בבעיה עם m אילוצי שוויון עלינו לפתור מערכת של m+n משוואות ב- m+n נעלמים, כיוון שהמערכת לפתרון כעת היא:

$$\begin{aligned} 0 &= \nabla f(\underline{x}) + \nabla H(\underline{x}) \cdot \underline{\lambda} \\ 0 &= H(\underline{x}) \end{aligned}$$

נוח להגדיר פונקציה הקרויה לנגרז'יאן - $L(\underline{x},\underline{\lambda}) = f(\underline{x}) + \underline{\lambda}^T H(\underline{x})$. פונקציה זו מהווה ריכוז של בעיית האופטימיזציה לפתרון, והיא כוללת בתוכה את האילוצים. במקרה זה התנאים ההכרחיים לקיומה של נקודת מינימום מתקבלים ע"י גזירתה,

$$\begin{aligned} \frac{\partial L(\underline{x},\underline{\lambda})}{\partial \underline{x}} &= 0 = \nabla f(\underline{x}) + \nabla H(\underline{x}) \cdot \underline{\lambda} \\ \frac{\partial L(\underline{x},\underline{\lambda})}{\partial \underline{\lambda}} &= 0 = H(\underline{x}) \end{aligned}$$

כדוגמה, נפתור את הבעיה הבאה, המנסה להביא למקסימום שטח פניה של תיבה בהינתן כי סכום אורכי צלעותיה ידוע,

$$\begin{aligned} &\text{Max} \quad x_1x_2 + x_1x_3 + x_2x_3 \\ &\text{S.T.} \quad x_1 + x_2 + x_3 = 3 \end{aligned}$$

נגדיר את הלנגרז'יאן:

$$L(x_1,x_2,x_3,\lambda) = x_1x_2 + x_1x_3 + x_2x_3 + \lambda[x_1 + x_2 + x_3 - 3].$$

גזירתו כפי שתואר קודם תיתן:

$$\begin{aligned} &\frac{\partial L(x_1,x_2,x_3,\lambda)}{\partial x_1} = x_2 + x_3 + \lambda = 0 \ ; \quad \frac{\partial L(x_1,x_2,x_3,\lambda)}{\partial x_2} = x_1 + x_3 + \lambda = 0 \\ &\frac{\partial L(x_1,x_2,x_3,\lambda)}{\partial x_3} = x_1 + x_2 + \lambda = 0 \ ; \quad \frac{\partial L(x_1,x_2,x_3,\lambda)}{\partial \lambda} = x_1 + x_2 + x_3 - 3 = 0 \end{aligned}$$

כך שהתקבלה מערכת משוואות ליניארית:

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 3 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -2 \end{bmatrix}$$

כדוגמה אחרת, נפתור את הבעיה הבאה אשר מתייחסת לאנטרופיה מקסימאלית כפי שזו מוגדרת בתורת האינפורמציה. בהינתן חוק הסתברות בדיד $\{p_k\}_{k=1}^{N}$ לסט ערכים אפשריים $\{x_k\}_{k=1}^{N}$, (דהיינו - $\mathrm{Prob}\{x_k\} = p_k$) מוגדרת האנטרופיה ע"י:

$$\varepsilon = -\sum_{k=1}^{N} p_k \log_2(p_k)$$

רצוננו להביא למקסימום את האנטרופיה תחת שני האילוצים הבאים - חוק ההסתברות חוקי ולכן סכום ההסתברויות חייב להיות 1, ועל כל ההסתברויות להיות אי-שליליות. נגדיר את הלנגרז'יאן באופן הבא:

$$L\left[\{p_k\}_{k=1}^{N}, \lambda_1, \lambda_2\right] = -\sum_{k=1}^{N} p_k \log_2(p_k) + \lambda_1\left[-1 + \sum_{k=1}^{N} p_k\right] = \sum_{k=1}^{N} p_k\left[-\log_2(p_k) + \lambda_1\right] - \lambda_1$$

בחרנו להתעלם מאילוץ החיוביות בשל היותו אילוצי אי-שיווין. גזירה לפי אוסף ההסתברויות נותנת:

$$.k = 1,2,\ldots,N: \quad -\log_2(p_k) + \lambda_1 - 1 = 0 \quad \Rightarrow \quad p_k = 2^{-1+\lambda_1}$$

כך שהתקבל כי אמנם ההסתברויות חיוביות, ואיננו צריכים לדאוג לכך באופן מפורש. את כופל הלגרנז' קובעים ע"י קיום האילוצים,

$$,1 = \sum_{k=1}^{N} p_k = N \cdot 2^{-1+\lambda_1} \quad \Rightarrow \quad \lambda_1 = \log_2\left(\frac{1}{N}\right) + 1$$

וברור אם כך שאנטרופיה מרבית תתקבל עבור הסתברויות שוות, $p_k = 1/N$, אז האנטרופיה היא $\log_2 N$.

כעת נציג את המשכו של המשפט הקודם, הנוגע לנגזרות מסדר שני. חלק זה מורכב יותר ויובא ללא הוכחה:

משפט 2: בנוסף לאמור במשפט 1, מתקבל

$$\begin{aligned} &\forall \underline{y} \in V(\underline{x}^*) \qquad \underline{y}^T\left[\nabla^2 f(\underline{x}^*) + \sum_{k=1}^{m} \lambda_k^* \nabla^2 h_k(\underline{x}^*)\right]\underline{y} \ge 0 \\ &\text{where}: \qquad V(\underline{x}^*) = \left\{\underline{y} \,\middle|\, \nabla H^T(\underline{x}^*)\underline{y} = 0\right\} \end{aligned}$$

אם הדרישה מוחלפת מאי-שליליות לחיוביות ממש, תנאים אלו גם מספיקים לקביעה כי הנקודה $\underline{x}^*$ אופטימאלית מקומית.

משמעות הטענה הנ"ל היא שגם כאן נדרשת חיוביות (חצי) מוגדרת על מנת שהנקודה אמנם תהווה פתרון מקומי. הפעם חיוביות (חצי) מוגדרת זו מנוסחת ביחס למטריצת

ההסיין של פונקצית הלנגרנז'יאן ולא רק זו של פונקצית המחיר, וכמו כן מוגבלת רק לווקטורים אנכיים לעמודות היעקוביאן של האילוצים. אנכיות זו פירושה כל הווקטורים אשר הליכה בכיוונם לא מפרה את האילוצים.

נמשיך את הדוגמה הקודמת שעסקה בשטח פיה של תיבה. עבור הבעיה

$$\begin{aligned} &\text{Max} \quad x_1x_2 + x_1x_3 + x_2x_3 \\ &\text{S.T.} \quad x_1 + x_2 + x_3 = 3 \end{aligned}$$

הגדרנו את הלנגרז'יאן כ:

$$L(x_1, x_2, x_3, \lambda) = x_1x_2 + x_1x_3 + x_2x_3 + \lambda[x_1 + x_2 + x_3 - 3].$$

גזירתו כפי שתואר קודם נתנה מערכת משוואות ליניארית שפתרונה קל. נחשב את מטריצת ההסיין המוכללת:

$$\nabla^2 f(\underline{x}^*) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad \nabla^2 h(\underline{x}^*) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow \nabla^2 f(\underline{x}^*) + \lambda^* \cdot \nabla^2 h(\underline{x}^*) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

מטריצה זו בבירור אינה חיובית מוגדרת. עם זאת, רצוננו למצוא את התנהגותה עבור וקטורים המקיימים

$$\nabla h^T(\underline{x}^*)\underline{y} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = 0$$

נגדיר אם כך את הווקטור הכי כללי המקיים זאת ע"י - $y^T = \begin{bmatrix} y_1 & y_2 & -(y_1 + y_2) \end{bmatrix}$. שימוש בווקטור זה נותן

$$\begin{bmatrix} y_1 & y_2 & -(y_1 + y_2) \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ -(y_1 + y_2) \end{bmatrix} =$$

$$= \begin{bmatrix} y_1 & y_2 & -(y_1 + y_2) \end{bmatrix} \begin{bmatrix} -y_1 \\ -y_2 \\ (y_1 + y_2) \end{bmatrix} = -y_1^2 - y_2^2 - (y_1 + y_2)^2$$

וכך קיבלנו כי לתת-מרחב של $\underline{y}$ אלו מתקבל כי ההסיין המוכלל הינו שלילי מוגדר - עובדה המאפשרת כי מדובר בנקודת מקסימום לוקלית.

2. בעיות עם אילוצי אי-שוויון

כעת נדון בבעיית האופטימיזציה הכללית ביותר אשר תכלול אילוצי אי-שוויון. אנו נניח כי הבעיה הינה בעלת המבנה הכללי הבא:

$$\begin{aligned} \min\ f(\underline{x}) \quad \text{S.T.}\quad & h_1(\underline{x}) = 0,\ \dots\ ,h_m(\underline{x}) = 0 \quad \{H(\underline{x}) = 0\} \\ & g_1(\underline{x}) \le 0,\ \dots\ ,g_p(\underline{x}) \le 0 \quad \{G(\underline{x}) \le 0\} \end{aligned}$$

כאשר $f, h_k, g_j \in C^2$. מסתבר כי לשם ניתוח בעיות אלו יש לעשות אבחנה בין אילוצים אקטיביים (המתקבלים בשוויון) ובין פסיביים (בהם התוצאה נותנת אי-שוויון ממש, ובשל כך השמטתם לא הייתה משפיעה על הבעיה). בדומה לנעשה קודם, נגדרי את היעקוביאן של האילוצים

$$J(\underline{x}) = \left[\nabla H(\underline{x}), \nabla\tilde{G}(\underline{x})\right] = \left[\begin{array}{cccc|cccc} \frac{\partial h_1(\underline{x})}{\partial x_1} & \frac{\partial h_2(\underline{x})}{\partial x_1} & \cdots & \frac{\partial h_m(\underline{x})}{\partial x_1} & \frac{\partial \tilde{g}_1(\underline{x})}{\partial x_1} & \frac{\partial \tilde{g}_2(\underline{x})}{\partial x_1} & \cdots & \frac{\partial \tilde{g}_{p_a}(\underline{x})}{\partial x_1} \\ \frac{\partial h_1(\underline{x})}{\partial x_2} & \frac{\partial h_2(\underline{x})}{\partial x_2} & \cdots & \frac{\partial h_m(\underline{x})}{\partial x_2} & \frac{\partial \tilde{g}_1(\underline{x})}{\partial x_2} & \frac{\partial \tilde{g}_2(\underline{x})}{\partial x_2} & \cdots & \frac{\partial \tilde{g}_{p_a}(\underline{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_1(\underline{x})}{\partial x_n} & \frac{\partial h_2(\underline{x})}{\partial x_n} & \cdots & \frac{\partial h_m(\underline{x})}{\partial x_n} & \frac{\partial \tilde{g}_1(\underline{x})}{\partial x_n} & \frac{\partial \tilde{g}_2(\underline{x})}{\partial x_n} & \cdots & \frac{\partial \tilde{g}_{p_a}(\underline{x})}{\partial x_n} \end{array}\right].$$

במטריצה זו נעשה שימוש רק ב-P_a אילוצי האי-שוויון המתקבלים בשוויון (לו רק ידענו אותם), והאחרים מושמטים. אנו נניח כי מטריצה זו בעלת דרגה מלאה.

משפט 3: (תנאים הכרחיים עפ"י Karush-Kuhn-Tucker). תהי $\underline{x}^*$ נקודת מינימום לוקלית של $f(\underline{x})$ המקיימת $H(\underline{x}^*)=0$, וכן - $G(\underline{x}^*) \le 0$, ונניח כי הנקודה $\underline{x}^*$ הינה רגולארית. אזי קיימים שני וקטורי כופלי לגרנז' יחידים $\underline{\lambda}^* = [\lambda_1, \;\ldots\;, \lambda_m]$, $\underline{\mu}^* = [\mu_1, \;\ldots\;, \mu_p]$ המקיימים כי עבור הגדרת הלנגרז'יאן הבאה:

$$L(\underline{x}, \underline{\lambda}, \underline{\mu}) = f(\underline{x}) + \sum_{k=1}^{m} \lambda_k h_k(\underline{x}) + \sum_{j=1}^{p} \mu_j g_j(\underline{x})$$

את הקשרים הבאים:

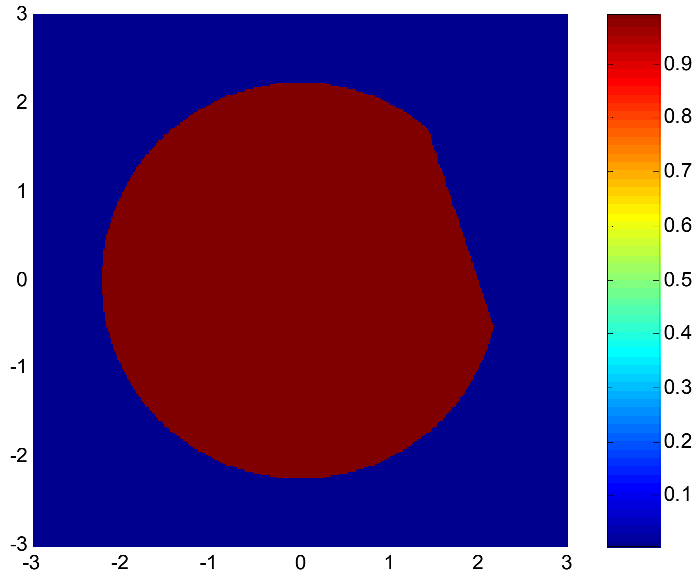
1. $\nabla_x L(\underline{x}^*, \underline{\lambda}^*, \underline{\mu}^*) = 0$
2. $\mu_j \ge 0$ לכל האילוצים האקטיביים, ואפס אחרת.
3. $\underline{y}^T \nabla^2_{xx} L(\underline{x}^*, \underline{\lambda}^*, \underline{\mu}^*)\underline{y} \ge 0$ לוקטורים $\underline{y}$ הניצבים ל-$J(\underline{x})$.

למעשה, עם שינוי דרישת אי-השליליות בחיוביות ממש הן בערכי כופלי הלגרנז' של האילוצים אקטיביים ובתכונת ההסיין, מומר משפט זה לספק תנאים הכרחיים ומספיקים לכך שנקודה כלשהי תהיה פתרון מקומי של בעיית האופטימיזציה. אנו נטען זאת ללא הוכחה.

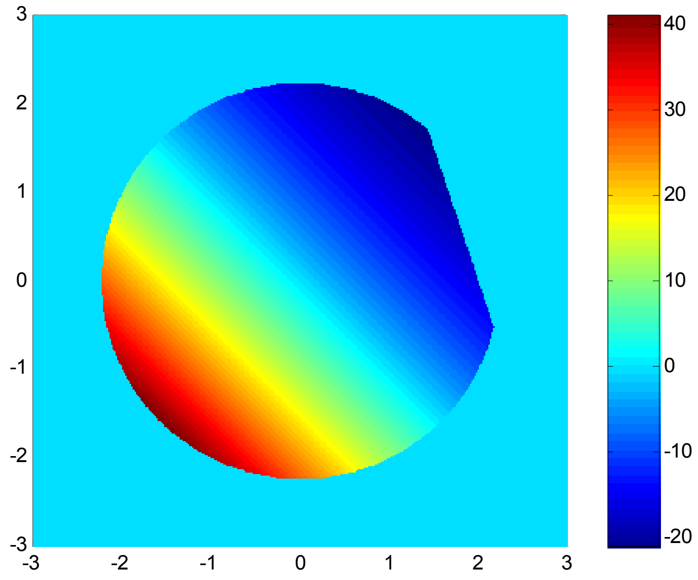
נתייחס לדוגמה הבאה:

$$\begin{array}{ll} \text{Min.} & (x+y)^2 - 10(x+y) \\ \text{S.T.} & x^2 + y^2 \le 5 \quad ; \quad 3x + y \le 6 \end{array}.$$

התחום בו האילוצים מתקיימים מתואר בציור הבא:



הפונקציה בתחום זה נראית כך:



ניכר כי הפתרון נמצא בפינה המחברת בין הישר למעגל – עובדה שאומרת ששני האילוצים יהיו פעילים. נתעלם מידיעה זו ונקרה אם נמצא אותה.

הלנגרז'יאן יהיה:

$$L(x,y,\mu_1,\mu_2)=(x+y)^2-10(x+y)+\mu_1\left(x^2+y^2-5\right)+\mu_2(3x+y-6)$$

תנאים מסדר ראשון מניבים את המשוואות הבאות:

$$0 = 2(x+y) - 10 + 2x\mu_1 + 3\mu_2$$
$$0 = 2(x+y) - 10 + 2y\mu_1 + \mu_2$$

כעת עלינו לעבור על ארבעת האפשרויות של אופיים (הפעיל\סביל) של אילוצי האי-שוויון בבעיה. נניח כי שני האילוצים אינם אקטיביים. במקרה כזה נקבל כי $\mu_1 = \mu_2 = 0$ ולכן:

$$\begin{bmatrix} 10 \\ 10 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow \forall\alpha, \quad \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha \\ 5-\alpha \end{bmatrix}$$

הצבת תוצאה זו באילוצים נותנת:

$$\begin{aligned} x^2 + y^2 &= \alpha^2 + (5-\alpha)^2 = 25 - 10\alpha \le 5 &&\Rightarrow \alpha \ge 2 \\ 3x + y &= 3\alpha + 5 - \alpha = 5 + 2\alpha \le 6 &&\Rightarrow \alpha \le 0.5 \end{aligned}$$

וקבוצת פתרונות זו ריקה - משמע - ההנחה כי שני האילוצים פסיביים אינה נכונה. נניח כי האילוץ הראשון בלבד אקטיבי. נפתור את מערכת המשוואות הבאה:

$$\left.\begin{aligned} 5 &= x + y + x\mu_1 \\ 5 &= x + y + y\mu_1 \\ x^2 + y^2 &= 5 \end{aligned}\right\} x = y \Rightarrow x = y = \sqrt{\frac{5}{2}} \Rightarrow \mu_1 = \frac{5 - 2\sqrt{\frac{5}{2}}}{\sqrt{\frac{5}{2}}} = \frac{5\sqrt{2} - 2\sqrt{5}}{\sqrt{5}} > 0$$

כך שלא התקבלה כל סתירה בינתיים. הצבת פתרון זה באילוץ השני נותנת:

$$3x + y = 4 \cdot \sqrt{\frac{5}{2}} = 6.32 > 6$$

כך שאילוץ זה כלל לא מתקיים, לכן גם הנחה זו אינה נכונה. נניח כי האילוץ השני הוא האקטיבי. מתקבל:

$$\left.\begin{aligned} 10 &= 2(x+y) + 3\mu_2 \\ 10 &= 2(x+y) + \mu_2 \\ 3x + y &= 6 \end{aligned}\right\} \mu_2 = 0 \Rightarrow \left.\begin{aligned} x + y &= 5 \\ 3x + y &= 6 \end{aligned}\right\} x = 0.5 \quad y = 4.5$$

וקיבלנו כי מצב זה חוקי (אין איסור שכופל הלגרנז' יהיה אפס עבור אילוץ אי-שיוויון אקטיבי). הצבת פתרון זה לאילוץ הראשון נותנת:

$$x^2 + y^2 = 0.5^2 + 4.5^2 = 20.5 > 5$$

כך שגם פתרון זה אינו נכון. נבדוק כעת את האפשרות ששני האילוצים אקטיביים. מתקבל:

$$\left.\begin{aligned} x^2 + y^2 &= 5 \\ 3x + y &= 6 \end{aligned}\right\} \begin{aligned} &x^2 + 36 - 36x + 9x^2 = 5 \Rightarrow 10x^2 - 36x + 31 = 0 \\ &x_1 = 2.17 \quad y_1 = -0.53 \quad \text{or} \quad x_2 = 1.425 \quad y_2 = 1.723 \end{aligned}$$

הצבת פתרונות אלו במשוואות:

$$0 = 2(x+y) - 10 + 2x\mu_1 + 3\mu_2$$
$$0 = 2(x+y) - 10 + 2y\mu_1 + \mu_2$$

נותנת לשני הפתרונות האפשריים:

$$\left.\begin{array}{l} 6.72 = 4.34\mu_1 + 3\mu_2 \\ 6.72 = -1.06\mu_1 + \mu_2 \end{array}\right\} \quad \mu_1 = -1.78 \quad \mu_2 = 4.825$$

$$\left.\begin{array}{l} 10 = 7.36 + 2.32\mu_1 + 3\mu_2 \\ 10 = 7.36 + 5.04\mu_1 + \mu_2 \end{array}\right\} \quad \mu_1 = 0.9882 \quad \mu_2 = 0.2945$$

כך שהפתרון הראשון נופל, והשני נראה חוקי - הקביעה אם אמנם הפתרון המוצע הוא המינימום יכול מעשית להיקבע ע"י חישוב מטריצת ההסיין המוכללת ובדיקתה - האם היא חיובית מוגדרת\שלילית מוגדרת תחת תת-המרחב $V(X^*)$ של הפתרון. לכן, בידינו הפתרון:

$$x^* = 1.425 \quad y^* = 1.723 .$$

עד כה התמקדנו בתנאים הכרחיים ומספיקים לנקודות מינימום (ומקסימום) לבעיות אופטימיזציה הכוללות אילוצי שוויון ואי-שוויון. במסגרת זו נפגשנו באופן מסודר עם המושגים פונקציית הלנגרז'איין, כופלי לגרנז', מטריצת ההסיין המוכללת ועוד. כעת נעבור לדון בפתרון נומרי של בעיות עם אילוצים.

3. שיטת המחיר לטיפול באילוצים

השאלה עליה נענה כעת היא: כיצד ניתן לרתום את תוצאות הדיון הקודם לשם בניית שיטות נומריות לפתרון בעיות אופטימיזציה כלליות הכוללות אילוצים. נרצה כמובן להכליל את האלגוריתמים אותם פגשנו כבר אשר התייחסו לבעיות ללא אילוצים, ולהרחיבם לטיפול בבעיות החדשות. מסתבר כי יש מספר אפשרויות, ואנו נתחיל את הדיון בגישות מחיר (Penalty Methods) אשר מציעות כאלטרנטיבה בעיה חדשה ללא אילוצים בה יש n נעלמים למציאה, וזאת ע"י ספיגת האילוצים לתוך פונקצית המחיר.

נניח אם כך כי לפנינו הבעיה הכללית הבאה:

$$\min_{\underline{x}} \; f(\underline{x}) \quad \text{S.T.} \; \{h_k(\underline{x}) = 0 \;\; 1 \le k \le m\} \,\&\, \{g_j(\underline{x}) \le 0 \;\; 1 \le j \le p\}$$

אזי, נוכל להגדיר באופן הבא את משפחה של פונקציות מחיר:

$$f(\underline{x}) + c_k \cdot P(\underline{x}) = f(\underline{x}) + c_k \cdot \left[\sum_{j=1}^{p} \rho_g \{\max[0, g_j(\underline{x})]\} + \sum_{j=1}^{m} \rho_h \{h_j(\underline{x})\} \right]$$

כאשר הפונקציות $\rho_g(\alpha), \rho_h(\alpha)$ חיוביות לכל α, שוות לאפס רק עבור $\alpha=0$ ומונוטוניות עולות כשמתרחקים מהאפס (לשני הצדדים). פונקצית מחיר כזו תגבה מחיר אפס לפתרון חוקי (העונה על כל האילוצים במדויק) ומחיר הפרופורציוני ל-c ולמידת החריגה אם אינו כזה. נתחיל עם c נמוך יחסית ונפתור את הבעיה שלנו באופן איטרטיבי. לאחר

מספר איטרציות וקבלת קרבה להתכנסות, נגדיל את c ונחזור על התהליך. כש-c גדל וגדל, מחיר החריגה מהאילוצים עולה וכופה עלינו את קיומם. הסיבה לזחילה למעלה עם ערכו של c במקום קביעתו כערך גבוה מלכתחילה נובעת מכך שעבור ערך גבוה מאוד, לפונקציה הכוללת הסיין חולני – כזה שה-condition-number שלו גבוה מאוד. אנו ראינו כי אלגוריתם ה-NSD קובע את קצב התכנסותו לפי יחס זה, ומסתבר שגם אלגוריתמים רבים אחרים תלויים בו ישירות. לכן, עם התחלה בערכים נמוכים והסיין בריא, נתקרב לפתרון, ובסביבתו נרשה הקצנה של ההסיין.

נמחיש את רעיון פונקצית המחיר דרך דוגמה: עבור הבעיה הבאה:

$$\begin{aligned}&\text{Min. } f(x,y)=x^2+xy+y^2-2y\\&\text{S.T. } h(x,y)=x+y-2=0\end{aligned}$$

נציע את פתרונה בגישת המחיר. נגדיר את הבעיה החדשה:

$$\text{Min. } q(x,y,c)=x^2+xy+y^2-2y+c(x+y-2)^2$$

גזירה לפי הנעלמים x ו- y מניבה:

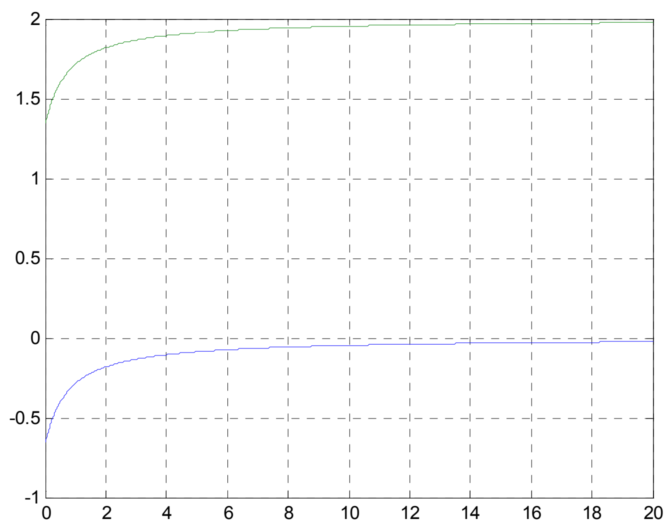
$$\frac{\partial q(x,y,c)}{\partial x}=2x+y+2c(x+y-2)=(2+2c)x+(1+2c)y-4c=0$$

$$\frac{\partial q(x,y,c)}{\partial y}=x+2y-2+2c(x+y-2)=(1+2c)x+(2+2c)y-2-4c=0$$

ופתרון מערכת משוואות זו נתון ע"י:

$$\begin{bmatrix}2+2c & 1+2c\\1+2c & 2+2c\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}=\begin{bmatrix}4c\\2+4c\end{bmatrix}\Rightarrow\begin{bmatrix}x\\y\end{bmatrix}=\frac{1}{3+4c}\begin{bmatrix}2+2c & -1-2c\\-1-2c & 2+2c\end{bmatrix}\begin{bmatrix}4c\\2+4c\end{bmatrix}$$

הגרף הבא מתאר את הפתרונות כפונקציה של c:

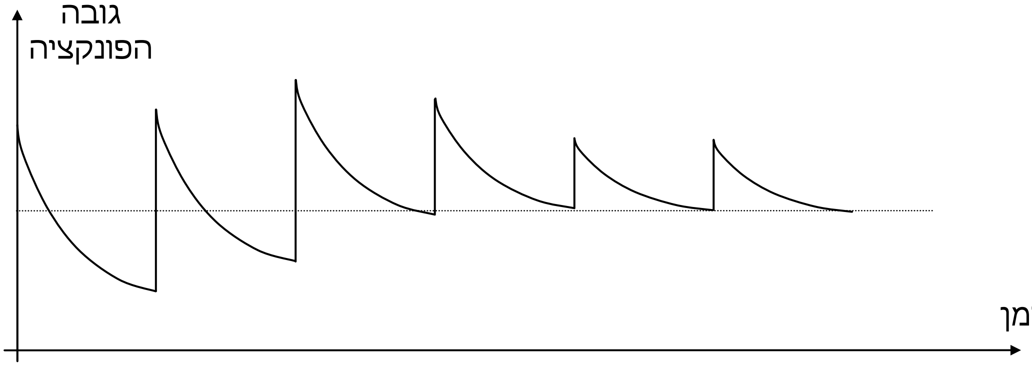


עבור הבעיה הנ"ל קל גם לקבל פתרון ישיר, ע"י הצבת האילוץ בפונקצית המחיר, דהיינו - $x = y - 2$. הבעיה תהיה שקולה לבעיה הבאה, נטולת האילוצים:

$$\min \quad (2-y)^2 + (2-y)y + y^2 - 2y = y^2 - 4y + 4 = (y-2)^2$$

ולכן ברור כי הפתרון נתון ע"י $y = 2, x = 0$, כפי שקיבלנו משיטת המחיר.

בחזרה למקרה הכללי: אם נסתכל על גובה הפונקציה $f(\underline{x}) + cP(\underline{x})$ כפונקציה של הזמן, נקבל מראה של פונקציה משוננת, כמתואר בציור הבא. כשציר הזמן מתקדם אנו מבצעים עוד ועוד איטרציות ואז חלה ירידה מתונה. מפעם לפעם אנו מגדילים את הערך c ואז חלה קפיצה.



בהנחה שקבוצת הפתרונות החוקיים אינה ריקה, נגיע לערכי c בהם המחיר $P(\underline{x})$ יהיה כבר קרוב מאוד לאפס או אפס ממש, וכך נקבל את התופעה המתוארת בציור - של נקודות התכנסות שוות גובה. זהו כמובן גם קריטריון מוצלח לקביעת עצירה לאלגוריתם.

מהאמור לעיל ברור שעבור $c \to \infty$ יתקבל הפתרון הנכון. מסתבר כי בחירה מסוימת מאוד של $\rho_g(\alpha), \rho_h(\alpha)$ יכולה להוביל למצב בו פתרון הבעיה החדשה יתלכד במדויק עם פתרונה של הבעיה המקורית עבור c סופי. גישה זו קרויה פונקצית המחיר המדויקת, וזו נתונה ע"י

$$P(\underline{x}) = \sum_{j=1}^{p} \max\left[0, g_j(\underline{x})\right] + \sum_{j=1}^{p} \left|h_k(\underline{x})\right|$$

(כלומר, הפונקציות $\rho_{g,h}$ הן פונקציות ערך מוחלט). עבור כל בחירה של c מעל סף מסוים מתקבל כי פתרון בעיה זו זהה לפתרון הבעיה המקורית עם אילוצים (מובא ללא הוכחה). יתרה מזו – הסף הנדרש אינו אלא כופל הלגרנז' הגדול ביותר (בערך מוחלט).

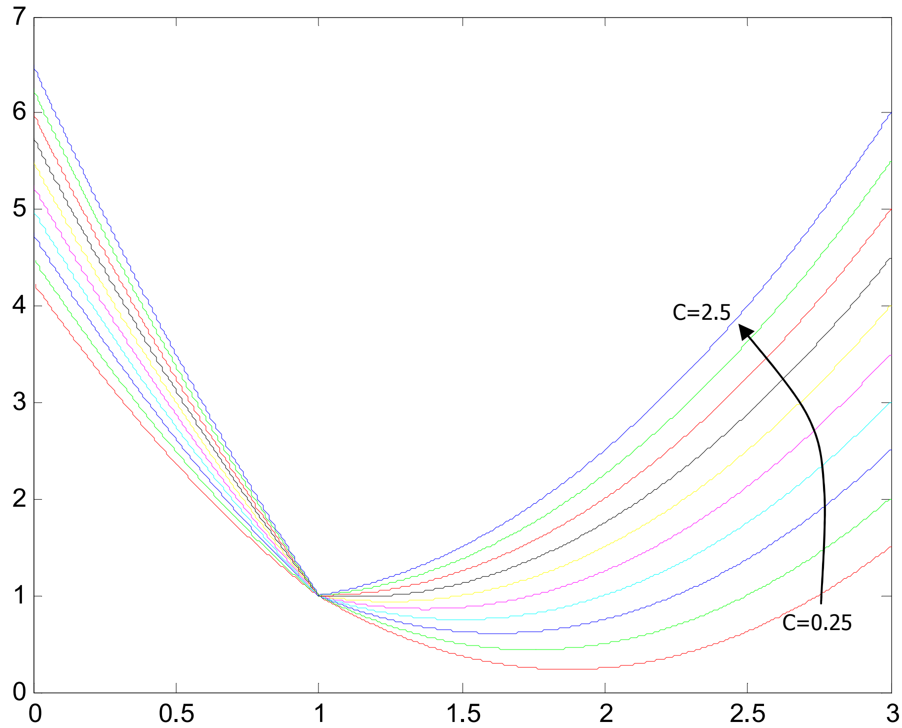
נמחיש זאת דרך דוגמה אשר נועדה להמחיש אינטואיטיבית את משמעות התוצאה הנ"ל. נתייחס לבעיית האופטימיזציה הבאה אשר פתרונה פשוט:

$$\min_x \; f(x) = (x-2)^2 \quad \text{S.T.} \;\; x = 1$$

על פי גישת המחיר המדויקת נבצע מינימיזציה של הפונקציה

$$. \, q(x,c) = (x-2)^2 + c|x-1|$$

הנה תיאור גרפי של פונקציות אלה עבור ערכי c הולכים וגדלים בקפיצות של 0.25 מ- 0.25 ועד 2.5.



אנו רואים כי עבור c גדול דיו המינימום מתקבל בערך x=1. פונקציית הלגרנז' של בעיה זו נתונה ע"י - $L(x,\lambda) = (x-2)^2 + \lambda(x-1)$, ופתרון התנאים ההכרחיים מניב:

$$. \, 2(x-2) + \lambda = 0, \;\; x - 1 = 0 \quad \Rightarrow \quad x = 1, \lambda = 2$$

לכן התיאוריה אומרת לנו כי נדרש $c \ge 2$ להתלכדות התוצאה. בדוגמה זו, לסף זה יש הסבר אינטואיטיבי פשוט: כאשר c>2 מתקבל כי מימין ומשמאל לנקודה x=1 בפונקציה q(x,c) תחול עליה כיוון שמתקיים:

$$\begin{aligned} &\nabla f(1^+) = -2 \quad \text{and} \quad \nabla h(1^+) = 1 \\ &\Rightarrow \forall c > 2, \; \nabla q(1^+, c) = \nabla f(1^+) + c \cdot \nabla h(1^+) > 0 \end{aligned}$$

(העלייה משמאל ברורה לחלוטין כיוון שחיברנו שתי פונקציות עולות). לכן, ברור כי בחירה זו של c תיתן כי פתרון המינימום מתלכד עם x=1, כדרוש.

4. שיטת המחסום לטיפול באילוצים

גישת המחסום דומה לגישת המחיר כפי שנראה מיד, אך מתאימה לבעיות בהן יש אילוצי אי-שוויון בלבד. אנו נתייחס לבעיה הכללית - $\min\ f(\underline{x})\ \text{S.T.}\ \underline{x}\in S$, כאשר האילוץ $\underline{x}\in S$ מתייחס לאוסף אילוצים פונקציונאליים אשר מגדירים מרחב פתרונות חוקיים S.

הרעיון בגישת המחסום הוא להציע בעיה תחליפית בה נבנית פונקצית מחסום $B(\underline{x})$ אשר מונעת מתהליך החיפוש לצאת את תחום הקבוצה S. האלגוריתם יאותחל בנקודה המקיימת את האילוצים ומצויה <u>בתוך S</u>, וייושם תהליך חיפוש כלשהו על הפונקציה המשולבת הכוללת את פונקצית המחסום. הפונקציה $B(x)$ בעלת התכונות הבאות:

1. B תהיה פונקציה רציפה.
2. אי-שליליות - $B(\underline{x})\geq 0\quad \forall\underline{x}$.
3. עבור נקודות השפה של S מתקיים - $B(\underline{x})\to\infty$

בהינתן פונקצית מחסום כזו, נגדיר משפחה של בעיות אופטימיזציה פרמטריות נטולות אילוצים:

$$\text{Min.}\ q(\underline{x},c_k)=f(\underline{x})+\frac{1}{c_k}B(\underline{x})$$

ניתן להראות כי עבור ערכים גבוהים מאוד של c_k יתקבל כי פתרון בעיה זו שקול לפתרון הבעיה המקורית. הרעיון לפתרון בעיית אופטימיזציה עם אילוצים בגישה זו דומה לנעשה קודם – לערך קבוע של c_k מובאת הפונקציה $q(\underline{x},c_k)$ למינימום, ואז מוגדל c_k וחוזר חלילה. גם כאן, לכאורה ניתן לגשת לפתור ישירות את הבעיה עם ערך גבוה מאוד של c_k, בידיעה שאז אנו מקרבים בצורה נאותה את הבעיה המקורית. אך ניתן להראות כי מאותם שיקולים אשר ניתנו בגישת המחיר, גישה זו אינה כדאית.

באשר לדרך בה יש לבנות את פונקצית המחסום, נניח אם כך כי לפנינו הבעיה הכללית הבאה (אילוצי אי-שוויון בלבד):

$$\min\ f(\underline{x})\quad \text{S.T.}\ g_j(\underline{x})\leq 0\ \ 1\leq j\leq p$$

אזי, נוכל להגדיר באופן הבא את פונקצית המחסום, כאשר הפונקציות $\rho(\alpha)$ חיוביות לכל α, שווה לאפס רק עבור $\alpha=0$, מונוטונית עולה, וכן - $\rho(\alpha)\xrightarrow[\alpha\to\infty]{}\infty$:

$$B(\underline{x})=\sum_{j=1}^{p}\rho\left\{\frac{-1}{g_j(\underline{x})}\right\}.$$

דרך אחרת לקביעת פונקצית המחסום היא:

$$B(\underline{x}) = -\sum_{j=1}^{p} \log\{-g_j(\underline{x})\}$$

בחירה זו לא מקיימת את התכונה $B(\underline{x}) \ge 0 \quad \forall \underline{x}$, אך אין בכך נזק לענייננו. החשיבות היא להציע פונקציה שתהיה חסומה מלמטה לכל הנקודות בתוך התחום החוקי. אי-שליליות היא דרך אחת לדרוש זאת. עבור קבוצה קומפקטית, הצעת ה- LOG חוקית באותה מידה.

עבור הבעיה הפשוטה הבאה נמחיש את משמעות פונקצית המחסום, לעומת פונקצית המחיר. נניח כי לפנינו הבעיה הבאה:

$$\text{Min. } f(x) = (x-3)^4 \quad \text{S.T. } g_1(x) = x-4 < 0 \ \ \& \ \ g_2(x) = 2-x < 0$$

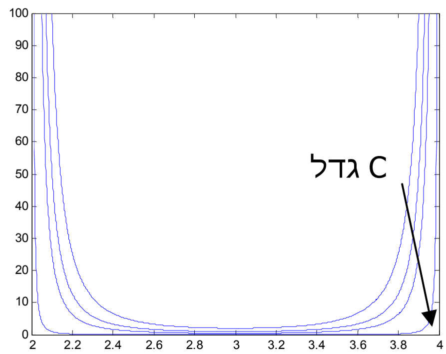
שיטת המחיר מציעה כאפשרות את הפונקציה הבאה:

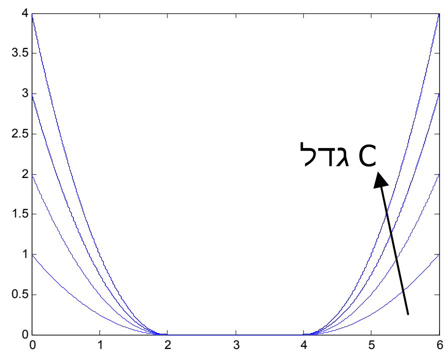
$$P(x) = \max\{0, x-4\}^2 + \max\{0, 2-x\}^2 = \begin{cases} (x-2)^2 & x < 2 \\ 0 & 2 \le x \le 4 \\ (x-4)^2 & 4 < x \end{cases}$$

שיטת המחסום מציעה כאפשרות את הפונקציה –

$$.B(x) = \left(\frac{1}{4-x}\right)^2 + \left(\frac{1}{x-2}\right)^2$$

גראפית שתי גישות אלה נראות כבציור הבא.





כפי שאנו רואים - בעוד שפונקצית המחיר מתירה פלישה מתחום האילוצים עם מחיר מתאים אשר הולך לאינסוף, גישת המחסום פועלת מתוך התחום בו מתקיימים האילוצים. כאשר הפרמטר c שואף לאינסוף מתקבל כי פונקצית המחסום היא בעלת מחיר אפס לכל נקודה בתוך S, וקירות בגובה אינסוף על שפת S.

באופן דומה למתואר בגישה הקודמת, גם כאן ניתן להסתכל בזמן על גובה הפונקציה המובאת למינימום. פונקציה זו היא $f(\underline{x}) + c^{-1}B(\underline{x})$, ובניגוד לגישת המחיר, בכל פעם שמגדילים את הקבוע c נקבל כי הפונקציה מונמכת, והתהליך האיטרטיבי מקטין אף הוא את גובה הפונקציה. לכן, בגישה זו יש הקטנה מתמשכת של גובה הפונקציה.

5. שיטת ההטלה לטיפול באילוצים

אפשרות אחרת לטיפול באילוצים היא גישת ההטלה: הרעיון המרכזי כאן הוא לפתור את בעיית האופטימיזציה ע"י אלגוריתם מבוסס גרדיאנט כדוגמת ה- SD, ולשלב בתוך אלגוריתם זה פעולות הטלה לתוך תת-המרחב של הפתרונות אשר מקיימים את האילוצים, וזאת לאחר כל איטרציה.

לפני שניגש לבעיית האופטימיזציה שעל הפרק, נראה תחילה כיצד מטילים על אילוצים, כשאנו מניחים תחילה שאילוצים אלה ליניאריים, מהצורה $\mathbf{A}\underline{x} = \underline{b}$ כשהמטריצה $\mathbf{A}$ בגודל של q שורות (כל שורה מייצגת אילוץ) על n עמודות (ממד הנעלם). נתונה נקודה $\underline{x}$ שאינה מקיימת את האילוצים, ורצוננו למצוא את הנקודה הקרובה ביותר אליה $\underline{y}$ אשר תקיימם. בניסוח מתמטי, זה נראה כך:

$$\min_{\underline{y}} \ \frac{1}{2}\left\|\underline{y} - \underline{x}\right\|^2 \text{ S.T. } \mathbf{A}\underline{y} = \underline{b}$$

פתרון בעיה זו יכול להיעשות ע"י q כופלי לגרנז' (אחד כנגד כל אילוץ שוויון):

$$L(\underline{y},\underline{\lambda}) = \frac{1}{2}\left\|\underline{y} - \underline{x}\right\|^2 + \underline{\lambda}^T(\mathbf{A}\underline{y} - \underline{b})$$
$$\frac{\partial L(\underline{y},\underline{\lambda})}{\partial \underline{y}} = \underline{y} - \underline{x} + \mathbf{A}^T\underline{\lambda} = 0 \quad \Rightarrow \quad \underline{y} = \underline{x} - \mathbf{A}^T\underline{\lambda}$$

הצבת הפתרון באילוץ נותנת:

$$\mathbf{A}\left(\underline{x} - \mathbf{A}^T\underline{\lambda}\right) = \underline{b} \quad \Rightarrow \quad \underline{\lambda} = \left(\mathbf{A}\mathbf{A}^T\right)^{-1}(\mathbf{A}\underline{x} - \underline{b}),$$

ולכן

$$\underline{y} = \underline{x} - \mathbf{A}^T\underline{\lambda} = \underline{x} - \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}(\mathbf{A}\underline{x} - \underline{b}) =$$
$$= \left[\mathbf{I} - \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{A}^T\right]\underline{x} + \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\underline{b}$$

המטריצה המכפילה את $\underline{x}$ מוכרת כמטריצת הטלה. אגב, קל לראות שכאשר $\underline{x}$ מקיים את האילוץ נקבל כי $\underline{y} = \underline{x}$.

נחזור כעת לבעיית האופטימיזציה שעל הפרק. נתייחס לבעיה הכוללת n נעלמים, m אילוצי שוויון, ו- p אילוצי אי-שוויון:

$$\min\ f(\underline{x})\quad S.T.\quad H(\underline{x})=0,\ G(\underline{x})\le 0$$

שיטת ה- SD מציעה את משוואת האיטרציה

$$\underline{x}_{k+1}=\underline{x}_k-\mu\nabla f(\underline{x}_k).$$

נניח כי הפתרון הנתון באיטרציה ה- k - $\underline{x}_k$ - מקיים את האילוצים. נניח כי בנקודה זו q אילוצים פעילים (כל אילוצי השוויון, וחלק מאילוצי האי-שוויון), והגרדיאנט עבורם ידוע לנו ונתון כמטריצה $A(\underline{x}_k)$ בגודל $[q\times n]$ (הפעם שורות מייצגות וקטורי גרדיאנט, ולא עמודות כמקודם, וזאת על מנת להתאים לדרך ההטלה שפותחה קודם). הסימון H(x) יכלול את כל המשוואות הללו יחדיו, ו- $A(\underline{x}_k)$ היא מטריצת היעקוביאן של סט משוואות אלה.

סביב הנקודה $\underline{x}_k$ באופן מקומי נוכל לקבוע את הקשר הבא המהווה קירוב טיילור מסדר ראשון של משוואת אלה:

$$H(\underline{x})=H(\underline{x}_k)+\mathbf{A}(\underline{x}_k)\cdot(\underline{x}-\underline{x}_k)\ \Rightarrow\ \mathbf{A}(\underline{x}_k)\underline{x}=\mathbf{A}(\underline{x}_k)\underline{x}_k-H(\underline{x}_k).$$

קיבלנו תבנית מהצורה $\mathbf{A}\underline{x}=\underline{b}$ המשקפת את האילוצים (לפחות מקומית). רצוננו לבצע עדכון לפתרון אשר מחד יקטין את ערך הפונקציה, ומאידך ישמור על קיום האילוצים. משוואת האיטרציה הנ"ל מבטיחה ירידה, אך ללא הבטחה כי האילוצים יישמרו. נגדיר את אופרטור ההטלה הליניארי על תת-מרחב האילוצים להיות:

$$\mathbf{P}(\underline{x}_k)=I-\mathbf{A}^T(\underline{x}_k)\left[\mathbf{A}(\underline{x}_k)\mathbf{A}^T(\underline{x}_k)\right]^{-1}\mathbf{A}(\underline{x}_k)$$

ונציע את האיטרציה הבאה, בה נעשה צעד SD ומוצאו מוטל לפי הנוסחא שקיבלנו:

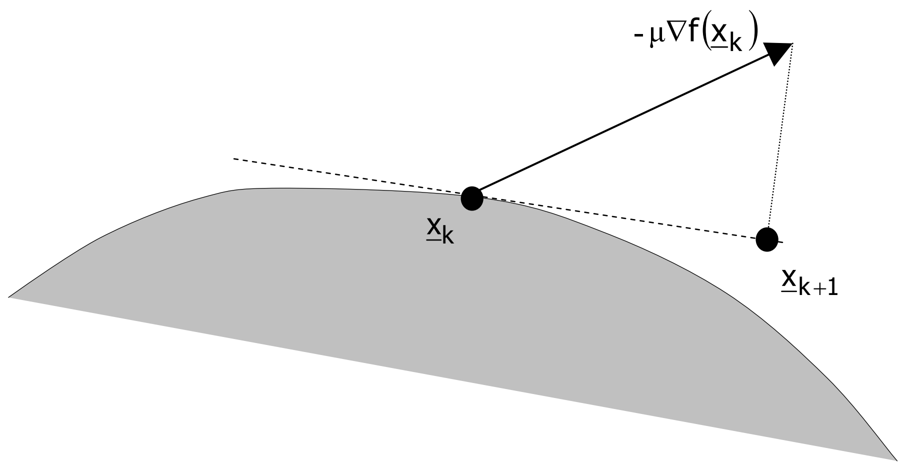
$$\begin{aligned}\underline{x}_{k+1}&=\mathbf{P}(\underline{x}_k)\left[\underline{x}_k-\mu\nabla f(\underline{x}_k)\right]+\mathbf{A}(\underline{x}_k)^T\left(\mathbf{A}(\underline{x}_k)\mathbf{A}(\underline{x}_k)^T\right)^{-1}\left(\mathbf{A}(\underline{x}_k)\underline{x}_k-H(\underline{x}_k)\right)=\\&=\mathbf{P}(\underline{x}_k)\underline{x}_k+\mathbf{A}(\underline{x}_k)^T\left(\mathbf{A}(\underline{x}_k)\mathbf{A}(\underline{x}_k)^T\right)^{-1}\underline{b}-\mu\mathbf{P}(\underline{x}_k)\nabla f(\underline{x}_k)\end{aligned}$$

צמד הגורמים הראשונים בביטוי הנ"ל אינם אלא הטלת הנקודה $\underline{x}_k$ למרחב האילוצים. מכיוון שלפי הנחתנו נקודה זו מקיימת האילוצים, הטלה זו שווה - $\underline{x}_k$, ולכן האיטרציה בפועל צריכה להיות

$$\underline{x}_{k+1}=\underline{x}_k-\mu\mathbf{P}(\underline{x}_k)\nabla f(\underline{x}_k).$$

נקודה חשובה נוספת היא כיוון החיפוש המיושם בפועל - נשים לב כי כיוון העדכון החדש הוא הווקטור $\mathbf{P}(\underline{x}_k)\nabla f(\underline{x}_k)$ - זהו כיוון ירידה כיוון שאופרטור ההטלה מהווה מטריצה חיובית חצי מוגדרת (נובע מהגדרתו).

פעולת ההטלה אשר הצענו תהיה מדויקת רק אם האילוצים ליניאריים. אחרת, הטלה זו כוללת שגיאה אשר עלולה לגרום לכך שהפעלת הטלה זו אינה מביאה לנקודה המקיימת את האילוצים. המחשת בעיה זו מתוארת בציור הבא



רעיון מעניין לפתרון בעיה זו הוא הבא: נניח כי הפעלת האיטרציה וההטלה הביאו לידינו את הנקודה $\underline{y}$. נרצה כעת לבצע תיקון על $\underline{y}$ לשם הבאתה לקיים את האילוצים, ולקבל $\underline{y}^*$ אשר מקיימת $H(\underline{y}^*) = 0$ (הפונקציה הוקטורית H כוללת גם את אילוצי אי-השוויון הפעילים). נציע עדכון מהצורה:

$$\underline{y}^* = \underline{y} + \mathbf{A}^T(\underline{x}_k)\underline{\alpha}$$

אשר מוסיף ל- $\underline{y}$ מרכיב שאינו בתת-המרחב אליו הטלנו קודם. נדרוש:

$$H(\underline{y}^*) = 0 = H\left(\underline{y} + \mathbf{A}^T(\underline{x}_k)\underline{\alpha}\right) \cong H(\underline{y}) + \mathbf{A}(\underline{x}_k)\mathbf{A}^T(\underline{x}_k)\underline{\alpha}$$

$$\Rightarrow \quad \underline{\alpha} = -\left(\mathbf{A}(\underline{x}_k)\mathbf{A}^T(\underline{x}_k)\right)^{-1}H(\underline{y}) \quad \Rightarrow \quad \underline{y}^* = \underline{y} - \mathbf{A}^T(\underline{x}_k)\left(\mathbf{A}(\underline{x}_k)\mathbf{A}^T(\underline{x}_k)\right)^{-1}H(\underline{y})$$

התהליך אשר מוצע כאן קובע דרך איטרטיבית לעדכון הפתרון עד להגעה לקיום האילוצים. הפעלת איטרציה אחת של קשרים אלו לא תספק בשל הקירוב אותו ביצענו - מספר איטרציות יביאו להשגת נקודה המקיימת את האילוצים. גישה זו מתבססת על ההנחה שהנקודה $\underline{y}$ אינה רחוקה מהנקודה $\underline{x}_k$, כיוון שאנו עושים שימוש בגרדיאנטים מנקודה זו לשם חישוב העדכונים השונים.

6. שיטות מבוססות Lagrange לטיפול באילוצים

עבור בעיית אופטימיזציה עם אילוצי שוויון - $\min f(\underline{x})$ S.T. $H(\underline{x}) = 0$, התנאים ההכרחיים אשר צריכים להתקבל בנקודת המינימום הם:

$$\nabla f(\underline{x}) + \nabla H(\underline{x})\underline{\lambda} = 0 \quad ; \quad H(\underline{x}) = 0 .$$

במסגרת השיטות מבוססות הלגרנז' מוצע לפתור מערכת משוואות זו או פונקציות הקשורות אליה ולמצוא בדרך זו את הצמד $\underline{x}^*, \underline{\lambda}^*$. כלומר, מוצעת כאן גישה בה יימצאו m+n נעלמים. מכיוון שנוח לנו להתבונן על בעיות נומריות כבעיות אופטימיזציה, נציע את פתרון מערכת המשוואות כמינימיזציה של הפונקציה המשולבת הבאה:

$$\min_{\underline{x},\underline{\lambda}} M(\underline{x},\underline{\lambda}) = \frac{1}{2}\left\|\nabla f(\underline{x}) + \nabla H(\underline{x})\underline{\lambda}\right\|^2 + \frac{1}{2}\left\|H(\underline{x})\right\|^2 = \frac{1}{2}\left\|\nabla L(\underline{x},\underline{\lambda})\right\|^2 + \frac{1}{2}\left\|H(\underline{x})\right\|^2$$

ידוע לנו כי אנו מחפשים נקודה בה ערך פונקציה זו אפס - דבר זה מהווה מידע חשוב בתהליך ההתכנסות של כל אלגוריתם - אם ניתקע על ערך שונה מאפס, נדע כי אנו בנקודת מינימום לוקלית, ויש לאתחל את האלגוריתם בנקודה שונה.

מגוון של אלגוריתמים ניתנים לבניה ואשר יביאו פונקצית מחיר זו למינימום. אנו לא נתעמק בזאת, ונסתפק באמירה כי גישות אלו זוכות ליתרון נוסף והוא ידיעת הערך באופטימום.

7. קמירות בבעיות עם אילוצים – LP ו-QP

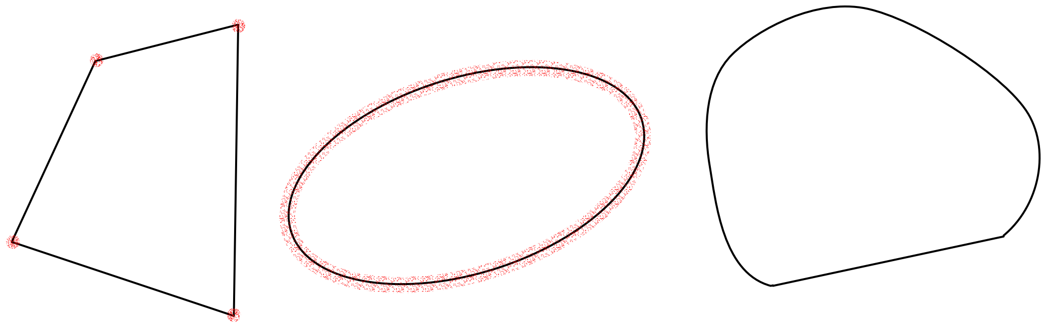
כבר דיברנו בחשיבותן של פונקציות קמורות, וראינו כי פונקציות כאלה מקלות את משימת האופטימיזציה. למעשה, כבר בדיון שניתן התייחסנו לאפשרות שישנם אילוצים, אם כי לא אמרנו זאת מפורשות. התייחסנו לפונקציות קמורות המוגדרות מעל תחום הגדרה קמור. כלומר, הנחנו כי $\underline{x} \in \Omega$, ו-Ω קמורה. אם יש לנו סדרת אילוצים המגדירה קבוצה של פתרונות חוקיים (feasible solutions) אשר מהווה קבוצה קמורה, המשפטים שניתנו בעבר תקפים. לכן נוכל לומר כי אם יש לנו פונקציה $f(\underline{x})$ קמורה מעל תחום הגדרה Ω קמור, אזי:

- הישר המחבר שתי נקודות מתחום ההגדרה על פני הפונקציה מצוי מעל הפונקציה,
- המשיק לפונקציה בכל נקודה נגדיר מישור המצוי מתחת הפונקציה בכל מקום בתחום ההגדרה,
- ההסיין של הפונקציה מעל תחום ההגדרה חיובי חצי מוגדר בכל מקום,
- לפונקציה זו כל נקודת מינימום לוקלי היא גם מינימום גלובלי,
- קבוצת פתרונות לוקליים שקולים יוצרים קבוצה קמורה,

יש עניין אחד בו לא דיברנו ואותו נזכיר כעת – מקסימיזציה של פונקציה קמורה מעל קבוצת אילוצים המגדירה תחום קמור. אנו נראה כי פתרון בעיה זו מוגש על שפת תחום ההגדרה ובנקודת קצה של שפה זו. נקודה $\underline{x}$ בקבוצה קמורה Ω מוגדרת כנקודת

קצה אם לא קיימות שתי נקודות $\underline{y}_1, \underline{y}_2 \in \Omega$ שונות מ- $\underline{x}$ כך שהנקודה $\underline{x}$ מצויה על הישר אשר מתוח ביניהן.

מההגדרה הנ"ל ברור כי כל נקודה המצויה בתוך קבוצה קמורה (כלומר שיש כדור קטן דיו סביבה, אשר מוכל כולו בקבוצה) אינה יכולה להיות נקודת קצה. נקודות קצה יכולות להתקבל רק על שפתה של הקבוצה הקומפקטית. לדוגמה, הנה מספר קבוצות קמורות ונקודות הקצה בהן



נניח, אם כך כי פונקציה קמורה $f(\underline{x})$ המוגדרת מעל קבוצה קמורה Ω מקבלת מקסימום ממש (כלומר רק בנקודה זו ולא גם בסביבתה) בנקודה $\underline{x}^* \in \Omega$. נניח כי $\underline{x}^*$ אינה נקודת קצה של Ω. זאת אומרת שנוכל למתוח קו כלשהו המתחיל בנקודה $\underline{y}_1$, חולף דרך הנקודה $\underline{x}^*$, ומגיע לנקודת יעד $\underline{y}_2$, כאשר $\underline{y}_1, \underline{y}_2 \in \Omega$. הנקודה $\underline{x}^*$ יכולה להיכתב כ- $\underline{x}^* = \alpha \underline{y}_1 + (1-\alpha)\underline{y}_2$, ובשל היות הפונקציה קמורה נקבל:

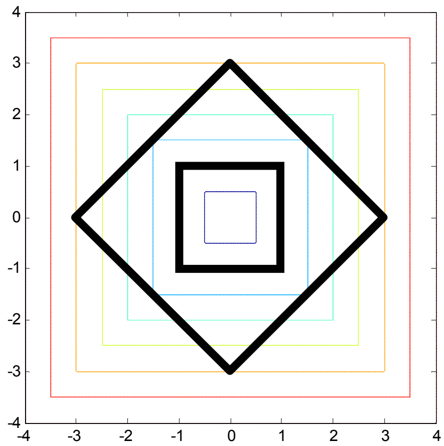
$$f(\underline{x}^*) = f(\alpha \underline{y}_1 + (1-\alpha)\underline{y}_2) \le \alpha f(\underline{y}_1) + (1-\alpha) f(\underline{y}_2) \le \max\{f(\underline{y}_1), f(\underline{y}_2)\}$$

יוצא אם כך שקיימת נקודה אלטרנטיבית $\underline{y}_1$ או $\underline{y}_2$ בה ערך הפונקציה משיג את שיא גובהו המרבי גם כן - זה עומד בסתירה לכך שנקודה $\underline{x}^*$ היא נקודת מקסימום ממש. המסקנה מכאן היא שהמקסימום (ממש) יוגשם רק בנקודת קצה של Ω. זה מביא למשפט הבא:

משפט 4: תהי נתונה פונקציה קמורה $f(\underline{x})$ מעל קבוצה קמורה Ω. אזי, אם לפונקציה זו יש מקסימום ממש בתחום הגדרתה, נקודת המקסימום תתקבל על נקודת קצה של Ω.

מסתבר כי אם נוותר על האמירה "מקסימום ממש", יתקבל כי נקודת המקסימום מתקבלת רק בנקודות שפה של תחום ההגדרה Ω, ולפחות בנקודת קצה אחת. נראה

זאת דרך הדוגמה הבאה - הפונקציה $f(x,y)=|x+y|+|x-y|$ נראית כמתואר בציור הבא.



נניח כי פונקציה זו מוגדרת מעל התחום - $\Omega=\{(x,y)|-1\le x,y\le 1\}$. קל לראות כי שפתה של הקבוצה Ω הוא בדיוק קו הגובה המקסימלי של הפונקציה בתחום ההגדרה, ואמנם, המקסימום במקרה זה מתקבל על כל נקודות שפה ולא רק בקצוות. אם לעומת זאת נחליף את Ω להיות התחום

$$\Omega=\{(x,y)||x|+|y|\le 3\},$$

כפי שמתואר בציור, נקבל כי נקודות המקסימום תהיינה ארבע נקודות הקצה של תחום ההגדרה.

8. בעיות תכנות ליניארי

יש מגוון עצום של בעיות אופטימיזציה משולבות אילוצים שבהן הן המחיר והן תחום האילוצים קמורים. למעשה, הטיפול במשפחת בעיות אלה הוא תחום מחקר פעיל מאוד היום, ואחד המדענים המובילים בו הוא ארקדי נמירובסקי מהפקולטה להנדסת תעשיה וניהול. במסגרת פרק זה נציג משפחת בעיות שזכתה לטיפול נרחב בשל תכיפות הופעתן וחשיבותן – התכנות הליניארי.

בעיית התכנות הליניארי (Linear Programming - LP) הקלאסית הינה בעלת מבנה קנוני מהצורה

$$\min_{\underline{x}} \ \underline{c}^T\underline{x} \quad \text{S.T.} \quad \mathbf{A}\underline{x}=\underline{b} \ \ \& \ \ \underline{x}\ge 0.$$

במבנה זה נניח $b_k\ge 0$ - ניתן להשיג זאת ע"י הכפלת כל משוואת אילוץ שאינה מקיימת זאת ב- (1-). בבעיית תכנות ליניארי אופיינית יתקיים כי m<n כאשר m מספר השורות

ב-**A** ו-n מספר העמודות בה, וזאת על מנת שמרחב הפתרונות החוקיים לא יהיה ריק או כזה המוביל לנקודה אחת (כי אז אין כאן בעיית אופטימיזציה). אנו נניח כי דרגת המטריצה **A** היא m בדיוק, כלומר כל האילוצים הינם בלתי תלויים, ולכן אין אילוץ שמופיע כסמוי באחרים. קל לגלות כאלה ולהשמיטם בטרם תחילת הטיפול.

העניין במשפחת הבעיות של התכנות הליניארי בא מכיוונים שונים. סוגיות רבות בכלכלה, הקצאת משאבים, ניהול, מדעי המחשב, חשמל, ועוד, ניתנים לייצוג כבעיות במבנה הנ"ל.

אחת הסוגיות הראשונות בהן יש לטפל מתייחסת למבנה הבעיה. אלגוריתמים התפורים לבעיה שתוארה מניחים את המבנה שתואר. עם זאת, במקרים רבים מתקבלת בעיה עם הרכב שונה במקצת, ונדרש לבצע עליה שינויים כדי להביאה למבנה הקנוני הנ"ל. נתאר מספר תסריטים אפשריים, ונראה כיצד מתבצעת המרה זו.

1. אילוצי אי-שוויון - נתונה הבעיה הבאה:

$$\min_{\underline{x}} \; \underline{c}^T \underline{x} \quad \text{S.T.} \quad \mathbf{A}\underline{x} \le \underline{b} \;\; \& \;\; \underline{x} \ge 0 .$$

במקרה כזה נציע את הבעיה השקולה הבאה:

$$\min_{\underline{x},\underline{y}} \; \underline{c}^T \underline{x} \quad \text{S.T.} \quad \mathbf{A}\underline{x} + \underline{y} = \underline{b} \;\; \& \;\; \underline{x}, \underline{y} \ge 0 .$$

בבעיה זו אנו עושים שימוש ב-m משתני דמה חדשים ולבעיה חדשה זו יש כעת n+m נעלמים. יתרה מזו – בעיה זו כבר בעלת מבנה קנוני כיוון שזו ניתנת לרישום ע"י

$$\min_{\underline{x},\underline{y}} \; \begin{bmatrix} \underline{c} \\ \underline{0} \end{bmatrix}^T \begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix} \quad \text{S.T.} \quad [\mathbf{A} \; \mathbf{I}] \begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix} = \underline{b} \;\; \& \;\; \begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix} \ge 0 .$$

אם חלק מהאילוצים בשוויון והאחרים באי-שוויון, רק לאלה יש להוסיף משתני דמה. באופן דומה, אם אי-השוויון בכיוון הפוך, אפשרות אחת היא הכפלת האילוץ $\mathbf{A}\underline{x} \ge \underline{b}$ ב-(1-) וטיפול כמקודם, או הכנסת משתני הדמה בסימן שלילי.

2. נתונה הבעיה הקלאסית למעט העובדה שלא נדרש $x_k \ge 0$. במקרה כזה ישנן שתי אפשרויות. הראשונה הינה שימוש בשני משתני דמה u ו- v כך ש- $x_k = u - v$. נחליף בבעיה את כל ההופעות של המשתנה x_k ב- u-v, ונדרוש בנוסף - $u \ge 0, v \ge 0$. כעת הבעיה בעלת מבנה קלאסי, אך ממימד של n+1. שיטה אחרת היא למצוא משוואת אילוץ אחת בה מופיע x_k, ובאמצעותה לייצגו כצירוף ליניארי של אוסף

המשתנים הנותרים. כעת נוכל להציב צירוף אלטרנטיבי זה בכל מקום בו מופיע x_k, וכך תתקבל בעיה חדשה בעלת מבנה קלסי עם n-1 נעלמים.

3. נניח כי נתונה בעיית תכנות ליניארי בה השינוי היחיד ביחס למבנה הקנוני הוא הדרישה $x_k \geq d$. במקרה כזה נגדיר משתנה חדש $y_k = x_k - d$, ונגדיר בעיה חדשה בה יש n נעלמים - x_k יוחלף ע"י y_k בכל מקום בהתאמה, והדרישה הפעם תהיה - $y_k \geq 0$ - כראוי. באופן דומה, כאשר הדרישה בבעיה הנתונה הינה $x_k \leq d$, נגדיר $y_k = d - x_k$.

4. עבור הבעיה הבאה:

$$\min_{\underline{x}} \quad \underline{c}^T |\underline{x}| \quad \text{S.T.} \quad \mathbf{A}\underline{x} \leq \underline{b} \quad \& \quad \underline{x} \geq 0 .$$

ניתן פשוט להשמיט את סימני הערך המוחלט כיוון שממילא נדרש כי כל ערכי x_k יהיו אי-שליליים. אם לעומת זאת הבעיה נתונה ללא אילוץ חיוביות,

$$\min_{\underline{x}} \quad \underline{c}^T |\underline{x}| \quad \text{S.T.} \quad \mathbf{A}\underline{x} \leq \underline{b} ,$$

הטיפול פחות פשוט. במקרה זה נציע את ההצבה הבאה - $\underline{x}=\underline{u}-\underline{v}$. עבור הצבה זו נקבל כי האילוץ יהיה $\mathbf{A}\underline{u}-\mathbf{A}\underline{v}=\underline{b}$ ונדרוש בנוסף $\underline{u}, \underline{v} \geq 0$. פונקצית המחיר תומר להיות $\underline{c}^T \underline{u} + \underline{c}^T \underline{v}$, והבעיה הכוללת תהיה:

$$\min_{\underline{u},\underline{v}} : \quad [\underline{c}^T, \underline{c}^T] \begin{bmatrix} \underline{u} \\ \underline{v} \end{bmatrix} \quad \text{S.T.} \quad [\mathbf{A}, -\mathbf{A}] \begin{bmatrix} \underline{u} \\ \underline{v} \end{bmatrix} = \underline{b} \quad \& \quad \begin{bmatrix} \underline{u} \\ \underline{v} \end{bmatrix} \geq 0 .$$

נושא האילוצים בבעיה הנ"ל ברור, אך נושא פונקצית המחיר תמוה במקצת ומחייב הסבר. ברור כי בכל מקום בו $\underline{u}$ שונה מאפס חייב להיות כי $\underline{v}$ יהיה אפס על מנת לתת ששתי הבעיות שקולות. זאת כיוון שאם x_k חיובי, יש מגוון ערכי u_k ו-v_k שיגשימו אותו. אנו רוצים שהבעיה "תדע" לקחת את החיוביים שבאברי $\underline{x}$ ולשים אותם ב-$\underline{u}$, ושליליים ל-$\underline{v}$. מסתבר (ואנו נראה זאת בהמשך פרק זה) שבעיות תכנות ליניארי מציעות פתרון עם מספר מצומצם ככל האפשר של איברים השונים מאפס (אנו נקרא לתכונה זו - פתרון בסיסי). השלכה ישירה של תכונה זו היא שבפתרון הבעיה הנ"ל, נקבל בדיוק את הפירוק הרצוי של $\underline{x}$, ולכן הבעיות שקולות.

ניתן בצורה דומה להציע עד מגוון וריאציות לבעיות אשר ניתנות להמרה לבעיות תכנות ליניארי קלאסיות. הנה שתי דוגמאות לבעיות בעלות מבנה כזה.

נניח שלרשותנו M תחנות לייצור חשמל במיקומים גיאוגרפיים שונים. כמו כן ישנם N יעדים לצריכת חשמל. עלות להעברת יחידת חשמל מתחנת ייצור m לצרכן n תסומן ע"י $P_{m,n}$. הצרכן ה- n דורש הספק של C_n. התחנה ה- m מייצרת הספק כולל של S_m.

רצוננו לקבוע את הדרך לאספקת החשמל באופן שיהיה הזול ביותר. בתרגום של הבעיה למשוואות נסמן ב - $X_{m,n}$ את כמות החשמל שתסופק מתחנה m לצרכן n. אלו הם הנעלמים של הבעיה אותם יש לקבוע. אזי, צריכים להתקיים האילוצים הבאים -

$$\sum_{n=1}^{N} X_{m,n} \le S_m \quad m = 1,2,\ldots,M \qquad \sum_{m=1}^{M} X_{m,n} = C_n \quad n = 1,2,\ldots,N$$

$$\Rightarrow \sum_{n=1}^{N} C_n = \sum_{m=1}^{M} S_m \qquad X_{m,n} \ge 0 \quad \forall m,n$$

M האילוצים הראשונים מתייחסים לכך שכל תחנה תספק לא יותר ממלוא כוח היצור שלה. N האילוצים הנוספים מבטיחים כי כל צרכן יקבל את הדרוש לו. שוויון בין הצריכה והייצור הכוללים הינו אילוץ על הנתונים ואינו חלק מהבעיה. תחת האילוצים הנ"ל, רצוננו להביא למינימום את הביטוי:

$$Cost = \sum_{m=1}^{M} \sum_{n=1}^{N} X_{m,n} P_{m,n}$$

הבעיה אותה הגדרנו היא בעיית אופטימיזציה בעלת מבנה LP.

כדוגמא אחרת, לרשותנו בארנק $[n_1, n_2, n_3, n_4]$ מטבעות של $[10, 50, 100, 500]$ אגורות בהתאמה. נניח כי משקלו של כל מטבע הוא $[m_1, m_2, m_3, m_4]$ בהתאמה. רצוננו לשלם על לפה עם שווארמה 16.5 שקלים מתוך המטבעות הנ"ל באופן שיוריד מעלינו משקל גדול ככל האפשר. נסמן ב-$[nn_1, nn_2, nn_3, nn_4]$ (<u>שלמים</u> בלבד!) את מספר המטבעות מכל סוג בהתאמה בהן נשתמש לתשלום. נוכל לכתוב את האילוצים הבאים:

$$0 \le nn_k \le n_k \quad 1 \le k \le 4$$

$$10nn_1 + 50nn_2 + 100nn_3 + 500nn_4 = 1650$$

האילוץ הראשון מוודא שאיננו משתמשים במטבעות מעבר לכמות שיש לנו. אילוץ שני מבטיח כי המחיר שישולם הוא זה הנדרש. תחת 5 אילוצים אלו נביא למקסימום את המשקל של המטבעות שישולמו, דהיינו -

$$Weight = \sum_{k=1}^{4} m_k \cdot nn_k$$

גם הפעם התקבלה בעיית תכנות ליניארי, אלא שהפעם נכללים אילוצי אי-שוויון, ובנוסף, אנו מחפשים את הפתרון רק מעל שדה השלמים. לעתים מקובל להתעלם בעניין היות הפתרון מעל השלמים ולהסתפק בקירוב.

כעת נפנה לדיון אשר יתחיל את בניית דרך הפתרון לבעיות אלו. נגדיר תחילה את המושגים "פתרון בסיסי" ו- "פתרון בסיסי חוקי". נסתכל על מערכת המשוואות $\mathbf{A}\underline{x} = \underline{b}$. למערכת זו מספר אינסופי של פתרונות, אשר מגדירים תת-מרחב. "פתרון בסיסי" מבין אלה הוא פתרון של המשוואה בו לכל היותר m איברים שונים מאפס. "פתרון בסיסי חוקי" יהיה פתרון בסיסי אשר כל ערכיו השונים מאפס גדולים מאפס, כך שהוא מקיים גם את אילוץ השוויון וגם את אילוץ החיוביות.

מכיוון ש- A היא מדרגה מלאה, ניתן למצוא בעמודותיה m וקטורים בלתי-תלויים ליניארית. נניח ללא אובדן הכלליות כי עמודות אלה הן הראשונות. אם נכפה $\{\forall\, m+1 \le k \le n \quad x_k \equiv 0\}$ נקבל את מערכת המשוואות הבאה:

$$\mathbf{A}\underline{x} = \begin{bmatrix} | & & | & | & & | \\ \underline{a}_1 & \cdots & \underline{a}_m & \underline{a}_{m+1} & & \underline{a}_n \\ | & & | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ \\ x_m \\ x_{m+1} \\ \\ x_n \end{bmatrix} = \begin{bmatrix} | & & | \\ \underline{a}_1 & \cdots & \underline{a}_m \\ | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \underline{b}$$

וכיוון ש- m העמודות הראשונות בלתי תלויות, ניתן להפוך מטריצה זו לקבל פתרון בסיסי. באופן דומה ניתן לבחור בכל פעם m עמודות אחרות מתוך n הקיימות ולקבל פתרונות בסיסיים (או מצב בו העמודות תלויות ולכן לבחירה ספציפית כזו לא יהיה פתרון בסיסי). סה"כ יהיו לכל היותר:

$$\#_s = \binom{n}{m} = \frac{n!}{(n-m)!m!}$$

פתרונות בסיסיים, וקרוב למדי שחלקם (ואולי אף מרביתם) אינם חוקיים. לדוגמה, אם ישנן 5 משוואות ו-10 נעלמים (n=10, m=5) יש 252 פתרונות בסיסיים אפשריים. בבעיה בה יש 50 משוואות ו- 60 נעלמים יהיו 7.5e10 פתרונות בסיסיים אפשריים. יתכן מצב בו פתרון מערכת של m כאלה משוואות תניב פתרון שחלק מאיבריו אפסים - פירוש הדבר שלפתרון הבסיסי המוצע יש פחות מ- m איברים שונים מאפס - ולכן בהגדרה נאמר "לכל היותר m ...". כעת נציג את המשפט היסודי של התכנות הליניארי, אשר מציג את פשטותה של בעיית תכנות ליניארי כללית.

משפט 5: (המשפט היסודי של התכנות הליניארי). בהינתן בעיית תכנות ליניארי בעלת המבנה הקלאסי:

$$\min_{\underline{x}} \quad \underline{c}^T \underline{x} \qquad \text{S.T.} \quad \mathbf{A}\underline{x} = \underline{b} \ \ \& \ \ \underline{x} \ge 0$$

מתקבל כי:

1. אם קיים פתרון חוקי (כלומר וקטור $\underline{x}$ אשר עונה לאילוצים - עובדה המעידה שקבוצת האילוצים לא ריקה), אזי קיים בהכרח גם פתרון בסיסי חוקי.
2. אם קיים פתרון חוקי אופטימאלי (כלומר פתרון חוקי אשר משיג את המינימום לפונקציה $\underline{c}^T\underline{x}$) אזי בהכרח גם קיים פתרון בסיסי חוקי אופטימאלי.

נוכיח תחילה את הטענה הראשונה של משפט זה. נניח כי נתון בידנו פתרון חוקי $\underline{x}$ בו יש p איברים שונים מאפס. נניח ללא אובדן הכלליות כי אלה הם p איבריו הראשונים של $\underline{x}$. אזי, אם $p \le m$ פירוש הדבר שבידנו פתרון בסיסי והטענה נכונה. לכן, נניח כי p>m. לכן, p העמודות המתייחסות לאיברים אלה (אלה הן p העמודות הראשונות לפי הנחתנו קודם) חייבות להיות תלויות ליניארית. נסמן

$$\hat{\mathbf{A}} = \begin{bmatrix} | & & | \\ \underline{a}_1 & \cdots & \underline{a}_p \\ | & & | \end{bmatrix} \text{ and } \underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}.$$

מצד אחד, קיים וקטור $\underline{y} \neq 0$ כך ש-$\hat{\mathbf{A}}\underline{y} = 0$, ומאידך מתקיים $\hat{\mathbf{A}}\underline{\hat{x}} = \underline{b}$. לכן, ע"י צירוף ליניארי של שתי אלה נקבל

$$\hat{\mathbf{A}}(\underline{\hat{x}} - \varepsilon\underline{y}) = \underline{b}.$$

נזכור כי $\underline{x}$ הוא פתרון חוקי ולכן כל איבריו השונים מאפס חיוביים. הווקטור $\underline{y}$, לעומת זאת, יכול לכלול ערכים חיוביים ושליליים - אנו נניח כי לפחות אחד ערכיו חיובי (אם לא - נכפול את כל ערכיו ב- (1-) וכולם יהיו חיוביים). לכן, עבור $\varepsilon > 0$ וקטן מאוד מתקבל מהמשוואה הנ"ל פתרון חוקי חדש. אם נגדיל את ε עוד ועוד, יגיע הרגע בו אחד המקדמים במשוואה הנ"ל יתאפס. איבר זה יהיה זה בו y_k חיובי, והמנה $\hat{x}_k / y_k$ היא הנמוכה ביותר. עבור $\varepsilon_0 = \hat{x}_k / y_k$ התקבל פתרון חוקי חדש בו יש רק p-1 ערכים שונים מאפס. נמשיך בדרך זו עד שנגיע ל- p=m, והפתרון החוקי אשר יתקבל יהיה בסיסי. באופן זה הוכחנו את הטענה הראשונה במשפט.

נניח כעת כי $\underline{x}$ הינו פתרון חוקי אופטימאלי, ולו p איברים שונים מאפס. אם $p \le m$ פירוש הדבר שבידנו פתרון בסיסי אופטימאלי והטענה נכונה. נניח אם כן כי p>m. כמקודם מתקבל כי עבור ε קטן מאוד (חיובי ושלילי) מתקבל כי $\underline{\hat{x}} - \varepsilon\underline{y}$ הינו פתרון חוקי. ערך הפונקציה לפתרון זה הוא

$$\underline{c}^T(\underline{\hat{x}} - \varepsilon\underline{y}) = \underline{c}^T\underline{\hat{x}} - \varepsilon\underline{c}^T\underline{y}.$$

אם הגורם $\underline{c}^T\underline{y}$ שונה מאפס תתקבל סתירה - נראה זאת - נניח כי $\underline{c}^T\underline{y}$ שונה מאפס. אזי יתקבל כי עבור ערך כלשהו של ε נוכל להנמיך עוד את ערך הפונקציה, ולכן $\underline{x}$ אינו אופטימאלי כפי שנטען. לכן בהכרח נדרוש כי $\underline{c}^T\underline{y}=0$.

כמקודם, נוכל כעת להסיט את ε עד לאיפוס איבר אחד בפתרון. הפתרון החדש יישאר חוקי - כפי שהראינו קודם, ולא ישנה את ערך הפונקציה ולכן יישאר אופטימאלי. נמשיך בדרך זו של איפוס איברים עד לקבלת p=m, והפתרון החוקי אשר יתקבל יהיה בסיסי ואופטימאלי. באופן זה הוכחנו את הטענה השנייה במשפט.

משמעותו של משפט זה הינה שבמקום לחפש במרחב הפתרונות שכולל אינסוף אפשרויות, עלינו לבדוק רק $\#_s$ פתרונות בסיסיים, לנפות את אלו שאינם חוקיים (כיוון שערכיהם גם שליליים), ולבחור את זה אשר מניב מינימום לערך הפונקציה. גישה זו של חיפוש הפתרון באופן ישיר אינה יעילה כמובן עבור בעיות גדולות.

המשפט הנ"ל אינו צריך להפתיענו. את עיקריו כבר ראינו קודם, כשדיברנו בתכונות של פונקציות קמורות מעל תחום הגדרה קמור. כפי שנראה מיד, בעיית התכנות הליניארית כוללת פונקציה שהיא קמורה וקעורה בו-זמנית (בהיותה ליניארית), ותחום האילוצים הוא קבוצה קמורה. לכן, הפתרון צריך להתקבל בקצוות. אנו נראה שקצוות האילוצים אינם אלא אותם פתרונות בסיסיים.

נתחיל עם התחום המוגדר ע"י האילוצים: קבוצת ה- $\underline{x}$ המקיימת את האילוצים $\mathbf{A}\underline{x}=\underline{b}\ \ \&\ \ \underline{x}\ge 0$ היא קבוצה קמורה. אם שתי נקודות $\underline{x}$ ו- $\underline{y}$ שייכות לקבוצה זו, קל לראות כי כל צירוף קמור שלהן שייך לקבוצה. מיהן נקודות הקיצון של קבוצה זו? נראה כי אלו הפתרונות הבסיסיים.

נניח כי $\underline{x}$ הוא פתרון בסיסי חוקי. נראה כי הוא נקודת קיצון ע"י כך שנראה שאין שתי נקודות שהינן פתרונות חוקיים אשר צירופן הקמור מניב את $\underline{x}$. נניח כי קיימות שתי נקודות כאלה - $\underline{y}_1,\underline{y}_2$. אזי - $\underline{x}=\alpha\underline{y}_1+(1-\alpha)\underline{y}_2\ \ \exists 0<\alpha<1$. מכיוון שאיבריו של $\underline{x}$ הם אפס בתחום $m+1\div n$, ומכיוון שכל איברי $\underline{y}_1,\underline{y}_2$ צריכים להיות חיוביים, בהכרח נקבל כי גם איברי $\underline{y}_1,\underline{y}_2$ בתחום $m+1\div n$ הם אפסים.

כיוון ש- m הערכים של כל פתרון אפשרי הם למעשה מקדמי הצירוף אשר ייתן $\hat{\mathbf{A}}\underline{\hat{x}} = \underline{b}$, ועמודות אלה בלתי תלויות ליניארית, נקבל כי בהכרח $\underline{x} = \underline{y}_1 = \underline{y}_2$, ולכן לא קיימות צמד נקודות פתרון חוקיות אשר יתנו את $\underline{x}$. בכך הראינו כי אף פתרון חוקי הוא בסיסי, אזי הוא נקודת קצה של תחום ההגדרה.

נניח כי $\underline{x}$ היא נקודת קיצון של K, ונניח כי נקודה זו כוללת $p > m$ ערכים שונים מאפס. נראה כי פירוש הדבר הוא שהיא אינה נקודת קיצון. כבר ראינו קודם כי ניתן במקרה זה להציע פתרונות חדשים מהצורה:

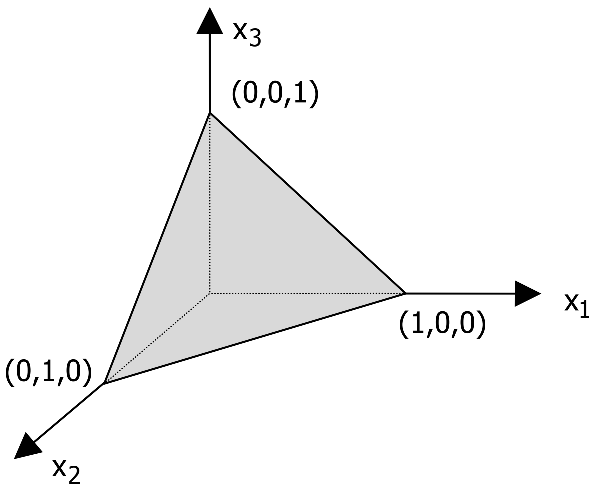
$$\hat{\mathbf{A}}\left(\underline{\hat{x}} - \varepsilon\underline{y}\right) = \underline{b} \quad \text{and} \quad \hat{\mathbf{A}}\left(\underline{\hat{x}} + \varepsilon\underline{y}\right) = \underline{b}$$

כאשר $\underline{y}$ הם מקדמי הצירוף הליניארי שנובע מהיותן של עמודות אלו תלויות. עבור ε קטן דיו (קטן מהערך הקטן ביותר של $\underline{x}$ בערך מוחלט) שני פתרונות אלו חוקיים, וצירופם נותן את $\underline{x}$. לכן לא יתכן כי $\underline{x}$ היא נקודת קיצון. לכן לא יתכן כי נקודת קיצון תכלול יותר מ- m איברים שונים מאפס. למעשה, קיבלנו את המשפט הבא:

משפט 6: יהי K המצולע הרב מימדי הקמור המוגדר ע"י - $\underline{x} \geq 0 \;\;\&\;\; \mathbf{A}\underline{x} = \underline{b}$. הווקטור $\underline{x}$ הוא נקודת קיצון של מצולע רב-מימדי זה אם ורק אם הוא פתרון בסיסי חוקי של המערכת $\underline{x} \geq 0 \;\;\&\;\; \mathbf{A}\underline{x} = \underline{b}$.

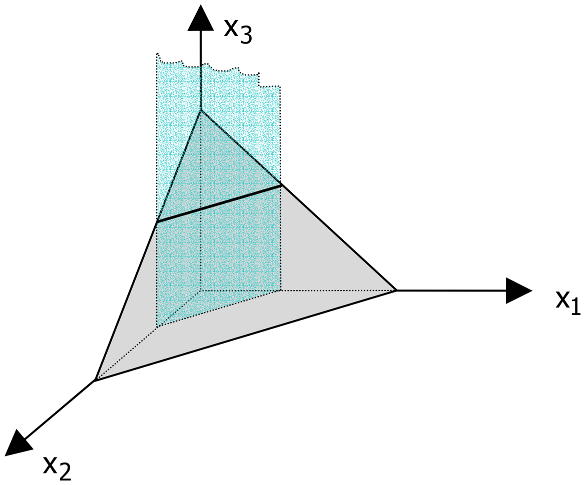
ממשפט זה נובע כי לקבוצה K יש לכל היותר $\#_s$ נקודת קיצון (קבוצת החוקיים שבהם קטנה יותר).

לדוגמה, עבור האילוצים - $\begin{matrix} x_1 + x_2 + x_3 = 1 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0 \end{matrix}$ קבוצת הפתרונות נראית כך:



וכצפוי, יש שלוש נקודת קיצון ל- K, המתייחסות לשלוש נקודת פתרון בסיסיות חוקיות, בהן רק איבר אחד שונה מאפס.

כדוגמה אחרת, נניח כי מערכת האילוצים כוללת בנוסף את המשוואה $2x_1 + 2x_2 = 1$.
הקבוצה K מתוארת בציור הבא:



הפתרונות הבסיסיים הפעם הם:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \begin{matrix} \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \\ \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}\begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \\ \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = ? \end{matrix}.$$

בפתרון הבסיסי השלישי אין כלל פתרון, כך שקיבלנו שני פתרונות בסיסיים חוקיים, וכן שתי נקודות קיצון ל- K.

ככלל, כאשר פתרון בסיסי נעלם זה יכול לקרות בשל כמה סיבות (I) הוא לא קיים, (II) הוא זהה לאחד האחרים, או (III) הוא לא חוקי.

לסיכום עד כה, ראינו כי פתרון בעיית תכנות ליניארית כרוך באיתור פתרון בסיסי וחוקי אשר יגשים את המינימום של הפונקציה. במונחים גיאומטריים, נקודת פתרון זו הינה נקודת קצה של קבוצת הפתרונות החוקיים. נקודה אחרונה היא שישנו מספר סופי (אך יכול להיות גדול) של נקודות כאלה.

בשל האמור לעיל, האמונה הייתה במשך שנים שאין צורך, כבאלגוריתמים איטרטיביים קודמים שהוצגו, לחפש מעל הרצף, ודי יהיה לסרוק בנקודות מסוימות את מרחב

הפתרונות החוקיים - עובדה הצפויה להקל את פתרון הבעיה. ובאמת, אלגוריתם ה-simplex שהוצע ע"י Danzig (Stanford, 1947) הציע "טיול" על שפתה של קבוצת האילוצים, הקופץ מפתרון בסיסי חוקי אחר לאחר, תוך כדי הורדת ערך הפונקציה. משך שנים ארוכות נחשב אלגוריתם זה לגישת פתרון קלאסית ונכונה ביותר. עם זאת, היה ברור שאלגוריתם זה סובל מבעיות נומריות רבות ההולכות ומחריפות עם הגדלת מימד הבעיה (מספר הנעלמים). בבעיות LP בנות אלפי נעלמים ויותר, שיטת ה-Simplex קורסת, ונדרשת אלטרנטיבה.

בשנת 1984 הציג Karmarkar (חוקר הודי ידוע בתחום האופטימיזציה) גישה חדשה לפתרון בעיות LP, המבוססת על שילוב של שיטות המחסום וההטלה. הצעתו עוררה גל של עבודות מחקר בתחום וכך נוצרה משפחה חדשה של שיטות הקרויות אלגוריתמי Interior Point או Central Path. יש אלגוריתמים רבים במשפחה זו - אנו נראה את הבסיסי שבהם ורק ברמתו העקרונית. גישה זו עולה על ה- Simplex כמעט בכל היבט - פשטות, קצב התכנסות, חסינות לבעיות נומריות ועוד.

הבעיה שעל הפרק היא:

$$\min_{\underline{x}} \ \underline{c}^T\underline{x} \quad \text{S.T.} \quad \mathbf{A}\underline{x}=\underline{b} \ \ \& \ \ \underline{x}\ge 0 .$$

יש לנו כאן שני סוגי אילוצים - אילוצים ליניאריים הקלים לטיפול ע"י הטלה (נזכור כי בשיטת ההטלה, התוצאה המתקבלת מדייקת ואינה כרוכה בקירוב), ואילוצי אי-שוויון פשוטים הניתנים לטיפול ע"י שיטת המחסום. נציע את פונקצית המחסום הבאה:

$$G(\underline{x})=-\sum_{k=1}^{n}\log\{x_k\} ,$$

וכעת נעבור לפתרון הבעיה השקולה הבאה:

$$\min_{\underline{x}} \ \underline{c}^T\underline{x}-\frac{1}{\text{Const}}\cdot\sum_{k=1}^{n}\log\{x_k\} \ \ \text{S.T.} \ \ \mathbf{A}\underline{x}=\underline{b}$$

ואת בעיה זו נפתור ע"י איטרציה של אלגוריתם כמו SD, NSD, DNSD, או אפילו ניוטון, כשלאחריה אנו מטילים על אילוצי השוויון את התוצאה לקבלת נקודה המקיימת את האילוצים.

נקודה שנייה חשובה כאן היא מבנה הפונקציה המובאת למינימום. זו פונקציה שחישוב נגזרותיה ואחסונם קל מאוד. זאת כיוון שאם נגדיר

$$f(\underline{x})=\underline{c}^T\underline{x}-\frac{1}{\text{Const}}\cdot\sum_{k=1}^{n}\log\{x_k\} ,$$

אזי

$$\nabla f(\underline{x})=\begin{bmatrix} c_1-\dfrac{1}{\text{Const}\cdot x_1} \\ c_2-\dfrac{1}{\text{Const}\cdot x_2} \\ \vdots \\ c_n-\dfrac{1}{\text{Const}\cdot x_n} \end{bmatrix} \quad \text{and} \quad \nabla^2 f(\underline{x})=\frac{1}{\text{Const}}\begin{bmatrix} \dfrac{1}{x_1^2} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{x_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{1}{x_n^2} \end{bmatrix}$$

כפי שרואים, ההסיין הינו מטריצה אלכסונית - אחסונו קל והיפוכו טריוויאליים. לכן ניתן להשתמש באלגוריתם מסדר שני כדוגמת ניוטון, ולהאיץ את ההתכנסות. כיוון שהפונקציה המשולבת (יחד עם המחסום) היא קמורה ממש אין גם כל חשש להינעלות על נקודת מינימום לוקלית.

האתחול לאלגוריתם חייב להוביל אותנו לנקודה בתוך תחום האילוצים. נקודה זו תיקבע ע"י קביעת הקבוע Const להיות כמעט אפס, ופתרון הבעיה

$$\max_{\underline{x}} \frac{1}{\text{Const}}\cdot\sum_{k=1}^{n}\log\{x_k\} \quad \text{S.T.} \quad \mathbf{A}\underline{x}=\underline{b}$$

נקודה זו מוגדרת כ-"מרכז האנליטי" של אילוצי בעיית ה- LP, וקל לקבלה ע"י אלגוריתם ניוטון (גם כאן הפונקציה קמורה וקלה לגזירה). מנקודת מרכז זו ע"י הגדלת הקבוע וביצוע 1-2 איטרציות של ניוטון (רצוי עם חיפוש על ישר כדי לא להישפך מחוץ לתחום המותר) אנו מתכנסים לנקודה $\underline{x}^*$ של בעיית ה- LP המקורית. בעוד שה- Simplex "מטייל" על גבולות תחום האילוצים ויכול לבצע חיפוש מייגע, לוקחת גישה זו "קיצור-דרך" בתוך תחום האילוצים, וישר לנקודה הרלוונטית.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 9 – התמרת פורייה ברצף

נושאים שייסקרו:

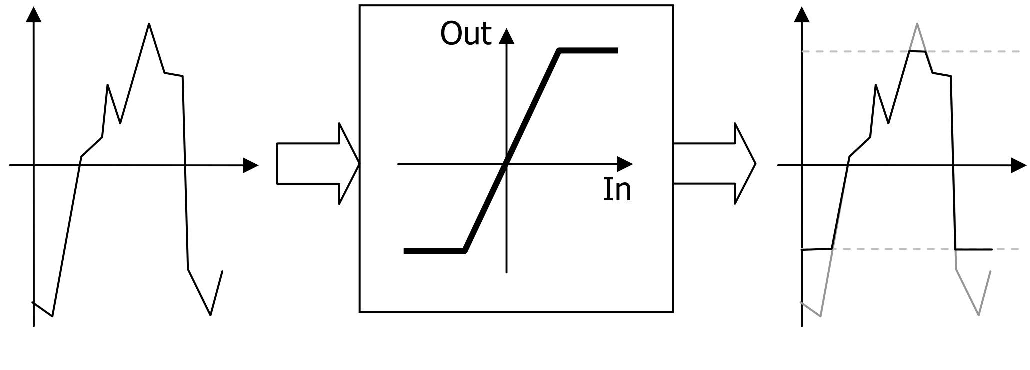
1. אותות ומערכות
2. מערכות קבועות בזמן
3. מערכות ליניאריות
4. פונקצית ההלם והתגובה לה
5. פעולת הקונבולוציה
6. התמרת הפורייה האינטגראלית
7. תכונות התמרת הפורייה האינטגראלית

---

1. אותות ומערכות

החיים מזמנים לנו שפע רחב של אותות מסוגים שונים. אות הינו פונקציה המתארת גודל (או גדלים) מסוימים כפונקציה של הזמן, המקום, או תחום הגדרה אחר. לדוגמה, גובה של אדם מיום לידתו ועד מותו היא אות המתאר אורך כפונקציה של הזמן. דוגמה אחרת לאות היא הלחץ הברומטרי במקום מסוים כפונקציה של הזמן. אנו נחשוב על אות כפונקציה $s(t)$, המוגדרת מעל הציר הרציף.

מה נוכל לעשות על אותות? מגוון רחב של פעולות, ואותן נכנה "מערכות"! בהינתן האות $s(t)$ נוכל להכפילו בקבוע a ולקבל אות חדש $a \cdot s(t)$. נוכל להעלות האות בריבוע, כלומר ליצור אות חדש $g(t) = s^2(t)$, ובעצם נוכל להציע כל פונקציה מוכרת לנו ($\sin$, $\cos$, $\log$, $\exp$, ועוד) אשר תפעל על האות ותיתן $g(t) = f\{s(t)\}$. כל אפשרויות אלו קרויות מערכות חסרות זיכרון כיוון שלקביעת g בזמן t כל שנדרש הוא לדעת את s באותו זמן ואת עקומת ההתאמה של הערכים f, אותה ניתן לתאר כגרף מהצורה הבאה:



נוכל גם להציע מערכות עם זיכרון. לדוגמה, מערכת מהצורה g(t)=s(t)+s(t-1) היא מערכת המפעילה זיכרון. כך גם המערכת

$$g(t) = \int_{-1}^{1} s(t-x)dx$$

אשר מחשבת את השטח מתחת לאות באינטרוול [-1,+1] סביב נקודת הזמן x.

מערכות עם זיכרון נשענות על מרכיב בעל חשיבות מרכזית – הזזת זמן. המערכת הבסיסית g(t)=s(t-d) משהה את אות הכניסה בשיעור d ואז מספקת אותו במוצאה. באופן דומה, המערכת g(t)=s(t+d) מקדימה את אות הכניסה בשיעור דומה.

נסתכל על שתי מערכות T_1 ו-T_2 המשתמשות בזיכרון ונבחין בהבדל מהותי ביניהן:

$$g_1(t) = T_1\{s(t)\} = \int_{0}^{1} s(t-x)dx \qquad g_2(t) = T_2\{s(t)\} = \int_{-1}^{1} s(t-x)dx$$

המערכת T_1 משתמשת בערכי אות הכניסה s(t) בזמנים [t-1,t] על מנת לקבוע את מוצא המערכת בזמן t. על מערכת זו נאמר כי היא סיבתית. המערכת השנייה אינה סיבתית כיוון שלשם חישוב מוצאה בזמן t היא צריכה לדעת את ערכי הכניסה בעבר ובעתיד, בתחום הזמנים [t-1,t+1]. סיבתיות הינה תכונה חשובה למערכות הפועלות ב-"זמן אמת" (כלומר, בכל רגע נתון מוציאות מוצא המתייחס לכל המידע שהצטבר עד לרגע זה).

מערכת יכולה לכלול מרכיבי משוב בתוכה. לדוגמה, המערכת $g(t) = g(t-1)s(t)$ משתמשת בערכי העבר כדי לקבוע את מוצאה בזמן t. מערכות כאלה, הנקראות מערכות רקורסיביות, מאופיינות בכך שהזיכרון האפקטיבי שלהן אינסופי (הן כוללות השפעת הכניסה לאורכי זמן ארוכים מאוד).

2. מערכות קבועות בזמן

נאמר על מערכת כי היא קבועה בזמן אם הפעלתה על אות המוזז בזמן מניבה את אותו המוצא כמו הפעלתה על האות המקורי והזזת מוצאה בזמן. כלומר, עבור המערכת T, אם לאות שרירותי s(t) המוצא הוא $g(t) = T\{s(t)\}$, אזי אם מתקיים כי $\forall d\ \ g(t+d) = T\{s(t+d)\}$, הרי שהמערכת קבועה בזמן. אינטואיטיבית, פירוש הדבר שהמערכת פועלת בצורה זהה בזמנים שונים. לדוגמה, המערכת $g(t) = t \cdot s(t)$ אינה קבועה בזמן כי הזזת מוצאה נותנת

$$g(t+d) = (t+d) \cdot s(t+d),$$

בעוד שהזזת הכניסה והזנתה תיתן

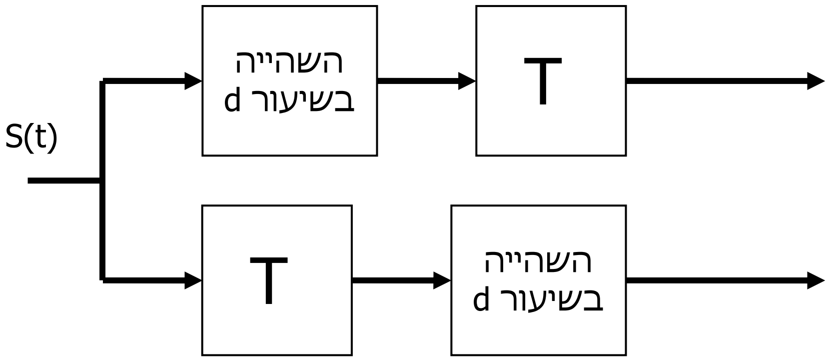
$$.\, t \cdot s(t+d)$$

לעומת זאת, המערכת $g(t) = s(t-1) \cdot s(t)$ קבועה בזמן כיוון שהזזת המוצא נותנת

$$g(t+d) = s(t-1+d) \cdot s(t+d)$$

והזזת הכניסה מביאה לאותו ביטוי.

בעצם, בהגדרת התכונה "קביעות בזמן" אנו שואלים אם החלפת הסדר בין שתי מערכות – האחת השהייה והאחרת המערכת שבמבחן – תיתן התנהגות זהה.



נושא זה של שרשור מערכות וחיבורם בדרכים שונות חשוב. אנו נאמר על שתי מערכות T_1 ו- T_2 כי הן משתחלפות עם מתקיים כי

$$.\, T_1\{T_2\{s(t)\}\} = T_2\{T_1\{s(t)\}\}$$

לדוגמה, עבור המערכות שהגדרנו קודם

$$g_1(t) = T_1\{s(t)\} = \int_0^1 s(t-x)dx \qquad g_2(t) = T_2\{s(t)\} = \int_{-1}^1 s(t-x)dx$$

יתקבל כי

$$T_1\{T_2\{s(t)\}\} = \int_0^1 \left(\int_{-1}^1 s(t-u+x)dx \right) du$$

$$T_2\{T_1\{s(t)\}\} = \int_{-1}^1 \left(\int_0^1 s(t+u-x)dx \right) du = \int_0^1 \int_{-1}^1 s(t-u+x)dxdu$$

ושתי מערכות אלו זהות. באופן דומה, שתי המערכות חסרות הזיכרון $g_1(t) = s^2(t)$ ו- $g_2(t) = \sqrt{|s(t)|}$ גם כן משתחלפות כיוון ש -

$$T_1\{T_2\{s(t)\}\} = \left(\sqrt{|s(t)|}\right)^2 = |s(t)|$$
$$T_2\{T_1\{s(t)\}\} = \sqrt{|s^2(t)|} = |s(t)|$$
.

לעומת זאת, שתי המערכות $g_1(t) = s^2(t)$ ו- $g_2(t) = s(t) + s(t-1)$ אינן משתחלפות כי

$$T_1\{T_2\{s(t)\}\} = (s(t) + s(t-1))^2$$
$$T_2\{T_1\{s(t)\}\} = s^2(t) + s^2(t-1)$$
.

כשם שניתן לחבר מערכות בטור בזו אחר זו, ניתן גם לחברן במקביל ולסכם את מוצאיהן. בצורה זו, אם נתונות שתי מערכות T_1 ו- T_2, חיבורן ייתן את המערכת השקולה

$$g(t) = T_1\{s(t)\} + T_2\{s(t)\}.$$

3. מערכות ליניאריות

מתוך מגוון המערכות האפשריות, ישנה תת-משפחה רחבה בה נרבה לעסוק, והיא המשפחה הליניארית. מערכת מוגדרת כליניארית אם היא מקיימת את שתי התכונות הבאות:

- תכונת ההומוגניות: עבור כל אות $s(t)$ וסקלר a מתקיים כי $T\{a \cdot s(t)\} = a \cdot T\{s(t)\}$.
- תכונת האדיטיביות: עבור כל שני אותות $s_1(t)$ ו- $s_2(t)$ מתקיים כי $T\{s_1(t) + s_2(t)\} = T\{s_1(t)\} + T\{s_2(t)\}$.

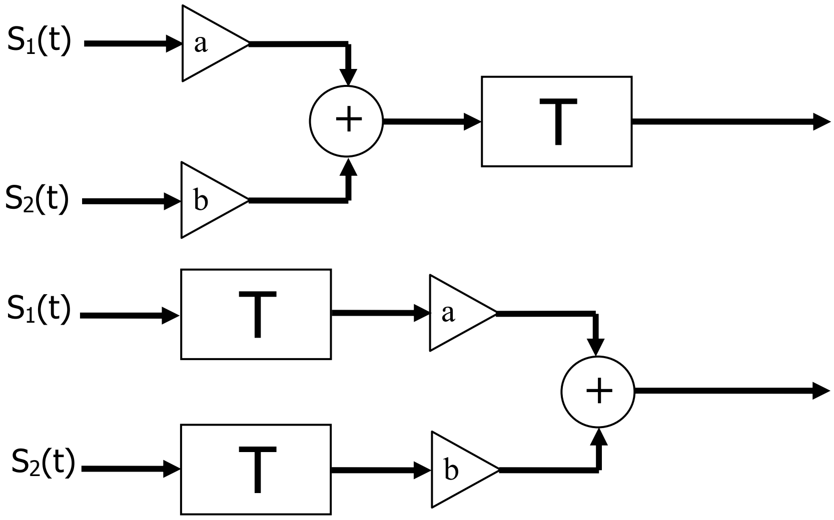
ניתן להחליף שתי תכונות אלו בתכונה אחת שקולה לחלוטין האומרת שלכל שני אותות $s_1(t)$ ו- $s_2(t)$ ולכל שני סקלרים a ו-b, יתקיים כי

$$T\{a \cdot s_1(t) + b \cdot s_2(t)\} = a \cdot T\{s_1(t)\} + b \cdot T\{s_2(t)\}.$$

ברמה אינטואיטיבית נאמר כי מערכת היא ליניארית אם מוצאה הוא צירוף ליניארי של ערכי הכניסה. כשאנו אומרים צירוף ליניארי, מותר שיהיו משקלים שונים בצירוף זה, ואף משקלים תלויי זמן.

ניתן מספר דוגמאות להמחיש זאת:

- המערכת $g(t) = s(t) + 3s(t-1)$ היא מערכת ליניארית. למעשה זו מערכת סיבתית, עם זיכרון, והיא גם קבועה בזמן.



- המערכת $g(t)=s(t)+\sqrt{t}\cdot s(t-1)$ תלויה בזמן, אך היא עדיין ליניארית, כי היא בונה את מוצאה כצירוף ליניארי של אות הכניסה.
- המערכת $g(t)=2s(t)+3$ אינה ליניארית. קל לראות זאת לפי ההגדרה, כיוון שתכונת ההומוגניות $T\{a\cdot s(t)\}=2a\cdot s(t)+3\neq a\cdot T\{s(t)\}=2a\cdot s(t)+3a$ לא מתקיימת.
- באופן דומה, המערכת $g(t)=\int_{-\infty}^{t}f(x)s(x)dx$ עבור פונקציה כלשהי ידועה מראש f תהיה ליניארית
- המערכת $g(t)=s(t)s(t-1)$ אינה ליניארית כי היא מצרפת את ערכי אות הכניסה באופן שאינו צירוף ליניארי.
- המערכת $g(t)=s'(t)=ds(t)/dt$ (גזירה) היא מערכת ליניארית! נראה זאת לפי ההגדרה:

$$\begin{aligned}T\{a\cdot s_1(t)+b\cdot s_2(t)\}&=\frac{d}{dt}\left[a\cdot s_1(t)+b\cdot s_2(t)\right]=\\&=a\frac{ds_1(t)}{dt}+b\frac{ds_2(t)}{dt}=\\&=a\cdot T\{s_1(t)\}+b\cdot T\{s_2(t)\}\end{aligned}$$

למעשה, מערכת זו גם קבועה הזמן כיוון שאם $g(t)=ds(t)/dt$, אזי הזזת הכניסה תביא להזזה דומה במוצא, דהיינו $g(t-d)=ds(t-d)/dt$.

תכונה מעניינת של מערכת ליניארית היא היכולת לשחלפה עם פעולת סיכום. זוהי הבחנה טריוויאלית הנובעת מההגדרה

$$T\{a\cdot s_1(t)+b\cdot s_2(t)\}=a\cdot T\{s_1(t)\}+b\cdot T\{s_2(t)\}.$$

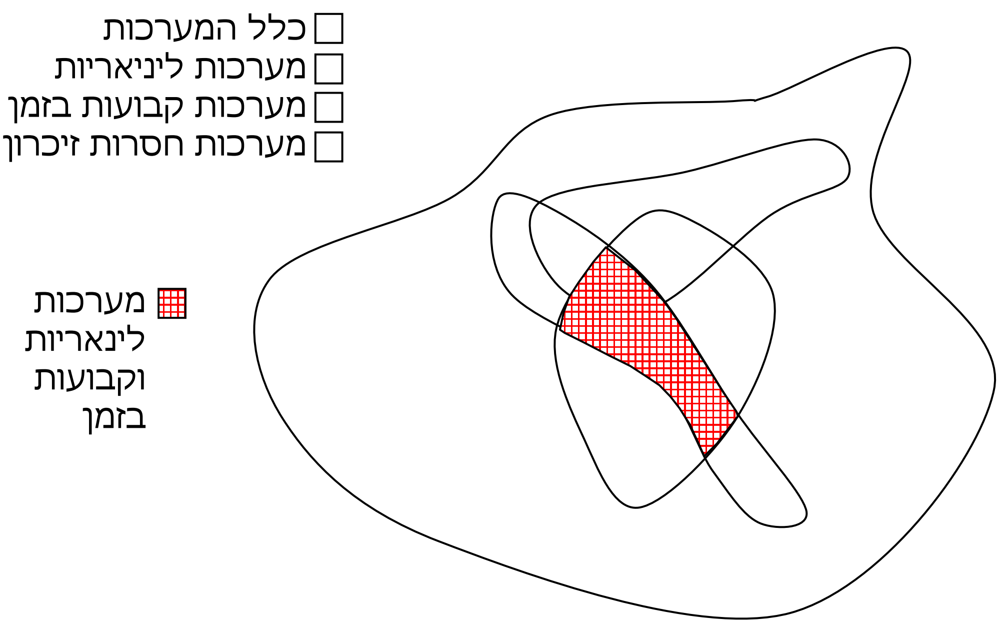
גם אינטגרל היא פעולת סיכום, וככזו, גם היא יכולה להשתחלף עם המערכת הליניארית. לכן, עבור אות $s_2(t)$ שהתקבל ממוצאה של המערכת (זוהי מערכת ליניארית, קבועה בזמן)

$$s_2(t) = \int_{-1}^{1} s_1(t-x)dx ,$$

אם בידינו מערכת ליניארית שנייה T אשר תפעל על $s_2(t)$, נקבל כי

$$T\{s_2(t)\} = T\left\{\int_{-1}^{1} s_1(t-x)dx\right\} = \int_{-1}^{1} T\{s_1(t-x)\}dx$$

ופירוש התוצאה הוא שניתן להפעיל את T קודם ואחר-כך את האינטגרל. אנו נפגוש תכונה זו שוב בצורתה הרחבה יותר בהמשך.



מה הקשר בין כל הדיון הנ"ל והפרקים הקודמים בהם דנו? מסתבר שהקשר הדוק. אם נחשוב על האות שלנו כווקטור $\underline{s}$, הכפלתו במטריצה $\mathbf{M}$, $\underline{y} = \mathbf{M}\underline{s}$, יכולה להיחשב כפעולתה של מערכת. מערכת כזו היא בהכרח ליניארית (נובע מידית מההגדרה של ליניאריות, ותכונות של הכפלה במטריצה). בדיון במטריצות ווקטורים, מימד האות סופי, וכך גם מימד המערכת הפועלת עליו. בדיון כעת אנו לא מניחים אות במימד סופי אלא חושבים עליו כפונקציה ברצף. לכן המערכת אינה מטריצה אלא גורם "אבסטרקטי" יותר, אך האנלוגיה למטריצה מרכזי וחשוב. אגב, היות המערכת קבועה בזמן מיתרגם בייצוג וקטורי מטריצי למבנים מאוד מיוחדים של המטריצה – אנו נדון בכך כשנמיר את כל הדיון הזה לדיון באותות בזמן בדיד ותמך מוגבל.

4. פונקצית ההלם והתגובה לה

לפני שנתחיל להרחיב על מערכות ליניאריות וקבועות בזמן, נדבר על אות מיוחד שיעניין אותנו לצורך ניתוח מערכות אלו – פונקצית ההלם. פונקציה זו מוזרה מעט וחולנית מבחינה מתמטית. לכן דרך הגדרתה עקיפה, ואנו נראה שתי דרכים לאפיונה.
נגדיר את האות המלבני הבא

$$s_T(t) = \begin{cases} \frac{1}{2T} & -T < t < T \\ 0 & \text{otherwise} \end{cases}.$$

נסתכל על אות זה לערכי T משתנים ההולכים וקטנים ושואפים לאפס. נקבל כי המלבן גדל לגובה אך מתכווץ ברוחבו, ובכל מקרה, לכל בחירה של T שטחו יהיה 1. כלומר,

$$\forall T, \quad \int_{-\infty}^{\infty} s_T(t)dt = 1.$$

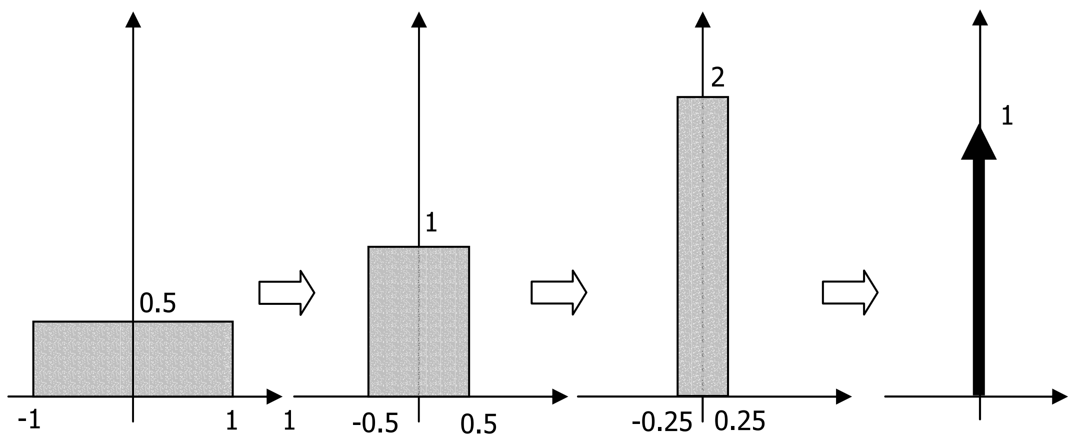
בגבול נקבל פונקציה שרוחבה אפס וגובהה אינסוף וזוהי פונקצית ההלם, המסומנת ב- $\delta(t)$:

$$\delta(t) = \lim_{T\to 0} s_T(t) = \lim_{T\to 0} \begin{cases} \frac{1}{2T} & -T < t < T \\ 0 & \text{אחרת} \end{cases}.$$

נוכל לפיכך לומר כי השטח מתחת לפונקצית ההלם גם הוא 1, כלומר,

$$\int_{-\infty}^{\infty} \delta(t)dt = 1.$$

בציור נתאר את אות ההלם כחץ אנכי ונציין את שטחו במקום גובהו, כמו בציור הבא.



מהידיעה ש- $\int_{-\infty}^{\infty} \delta(t)dt$ ברור כי

$$a \cdot \int_{-\infty}^{\infty} \delta(t)dt = \int_{-\infty}^{\infty} [a \cdot \delta(t)]dt = a.$$

כלומר, ניתן לכפול פונקציה זו בקבוע a ולקבל פונקציה חדשה בעלת ערך אינסוף בראשית, אך השטח מתחתיה גדול פי a.

נניח כי בידינו שני אותות, s(t) כלשהו ופונקצית ההלם, ורצוננו לחשב את השטח מתחת מכפלתם. שטח זה יהיה

$$\int_{-\infty}^{\infty}\delta(t)s(t)dt = s(0) \ ,$$

כיוון שערכי האות בכל מקום מלבד בראשית אינם רלוונטיים כי הם מוכפלים באפס. במקום האפס s(0) מכפיל את שיא ההלם, וכך משפיע את השטח הכולל, ממש כמו שהקבוע a עשה זאת קודם. באופן דומה, התוצאה הבאה

$$\int_{-\infty}^{\infty}\delta(t)s(t+a)dt = \int_{-\infty}^{\infty}\delta(z-a)s(z)dz = s(a) \ ,$$

קובעת כי ערך האות s(t) רלוונטי רק במקום בו שיא ההלם פוגש את ערך הפונקציה, ובמקרה זה ב-a. האגף המרכזי במשוואה מתקבל משינוי משתנה האינטגרציה. אם נכתוב את המשוואה שקיבלנו בצורה שונה מעט, נאמר

$$\int_{-\infty}^{\infty}\delta(x)s(t+x)dx = \int_{-\infty}^{\infty}\delta(t-x)s(x)dx = s(t) \ .$$

משמעות משוואה זו שהאות s(t) יכול להיות מתואר כצירוף ליניארי של אותות הלם מוזזות $\delta(t-x)$ כשהן משוקללות ע"י ערכי האות עצמו s(x). לייצוג זה תהיה חשיבות רבה בבואנו לנתח מערכות הפועלות על אותות.

למרות שפונקצית ההלם חולה, הפונקציה הקדומה לה (האינטגרל שלה) מוגדרת היטב והינה פונקצית מדרגה פשוטה

$$\int_{-\infty}^{t}\delta(x)dx = \mu(t) = \begin{cases} 1 & t \ge 0 \\ 0 & \text{אחרת} \end{cases} \ .$$

לא נוכל ברוח דומה לומר מה הנגזרת של פונקצית ההלם, אך נוכל להראות את הקשר הבא

$$\int_{-\infty}^{\infty}\frac{d\delta(x)}{dx}s(x)dx = \left[\delta(x)s(x)\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\frac{ds(x)}{dx}\delta(x)dx = -s'(0)$$

כשהשתמשנו בתכונת האינטגרציה בחלקים.

נחזור כעת למערכות ונניח כי בידינו מערכת ליניארית וקבועה בזמן. אם נזין למערכת זו הלם $\delta(t)$, נקבל מוצא אשר ייקרא התגובה להלם ויסומן $h(t)$, כלומר $T\{\delta(t)\} = h(t)$. לדוגמה, עבור המערכת $g(t) = s(t) + s(t-1)$ תגובתה להלם פשוטה למדי ומתקבלת פשוט ע"י הצבה $s(t) = \delta(t)$ הנותנת כי $h(t) = \delta(t) + \delta(t-1)$. כדוגמה שנייה, מתקבל כי עבור המערכת

$$g(t) = \int_{-1}^{2} s(t-x)dx ,$$

התגובה להלם היא

$$h(t) = \int_{-1}^{2} \delta(t-x)dx \underset{t-x=z}{\underset{\uparrow}{=}} -\int_{t+1}^{t-2} \delta(z)dz = \int_{t-2}^{t+1} \delta(z)dz = \begin{cases} 0 & t > 2 \\ 0 & t < -1 \\ 1 & -1 \le t \le 2 \end{cases}$$

המקרה הראשון מתייחס לגבול תחתון חיובי ולכן אנרגיית ההלם לא מוכלת בתחום האינטרגרציה. באופן דומה, המקרה השני מתייחס לגבול עליון שלילי.

נניח כעת כי הזנו אות כלשהו אחר $s(t)$ למערכת T. מהו הקשר בין מוצא המערכת כעת והתגובה להלם $h(t)$? לצורך בניית התשובה נשתמש בתכונה שהראינו קודם לכן על היכולת לפרק אות לצירוף ליניארי של פונקציות הלם

$$\int_{-\infty}^{\infty} \delta(t-x)s(x)dx = s(t) .$$

לכן,

$$g(t) = T\{s(t)\} = T\left\{\int_{-\infty}^{\infty} s(x)\delta(t-x)dx\right\} = \int_{-\infty}^{\infty} T\{\delta(t-x)\}s(x)dx = \int_{-\infty}^{\infty} h(t-x)s(x)dx$$

כשהכנסנו את פעולת המערכת לתוך האינטגרל עשינו שימוש בתכונת הליניאריות של המערכת שמשמעה כי מערכת ליניארית על סכום של איברים שקולה להפעלת המערכת על כל אחד מאיברים אלה וסיכום המוצאים. ההחלפה $T\{\delta(t-x)\} = h(t-x)$ מותרת ונכונה בשל הנחתנו כי המערכת T קבועה בזמן. המסקנה היא שלמערכות ליניאריות וקבועות בזמן, ידיעת התגובה להלם פירושה ידיעת המוצא של המערכת לכל כניסה אפשרית לפי הקשר

$$g(t) = \int_{-\infty}^{\infty} h(t-x)s(x)dx = \int_{-\infty}^{\infty} h(x)s(t-x)dx .$$

נאמר יותר מזאת – כל מערכת ליניארית ניתנת לאפיון ע"י תגובתה להלם. מסתבר כי תיתכנה שתי מערכות ליניאריות שונות אשר להן אותה תגובה להלם, אך אנו נתעלם מבעיה זו ונניח התאמה חד-חד-ערכית בין שני המושגים.

5. פעולת הקונבולוציה

המשוואה שנכתבה כקשר בין תגובת ההלם לאות הכניסה ידועה בשם "קונבולוציה". אנו מסמנים אותה ע"י הפעולה $\otimes$,

$$g(t)=\int_{-\infty}^{\infty}h(t-x)s(x)dx=\int_{-\infty}^{\infty}h(t)s(t-x)dx=h(t)\otimes s(t).$$

נמחיש כיצד מחושבת הקונבולוציה בפועל באופן גרפי. אם נתייחס לביטוי

$$g(t)=\int_{-\infty}^{\infty}h(t-x)s(x)dx=h(t)\otimes s(t)$$

הרי שעלינו לבצע השלבים הבאים:

- אנו נתייחס לאינטגרל ע"י ציור האותות מעל הציר x.
- נצייר את s(x) על מערכת צירים זו.
- יש לקחת את התגובה להלם h(x) ולהפוך אותה בתמונת מראה לקבלת h(-x). לאחר מכן יש להזיזה בשיעור t קדימה לקבלת h(t-x).
- יש לכפול את h(t-x) באות s(x).
- יש לחשב האינטגרל מתחת שטח המכפלה.
- ע"י הסעת h מעל s (לסריקת ערכי הזמן t) נקבל את המוצא בכל נקודת זמן.

נראה דוגמה לשם המחשת התהליך ונבנה אותה עבור המערכת שראינו כבר:

$$h(t)=\int_{-1}^{2}\delta(t-x)dx=\begin{cases}1 & -1<t<2\\ 0 & \text{אחרת}\end{cases}.$$

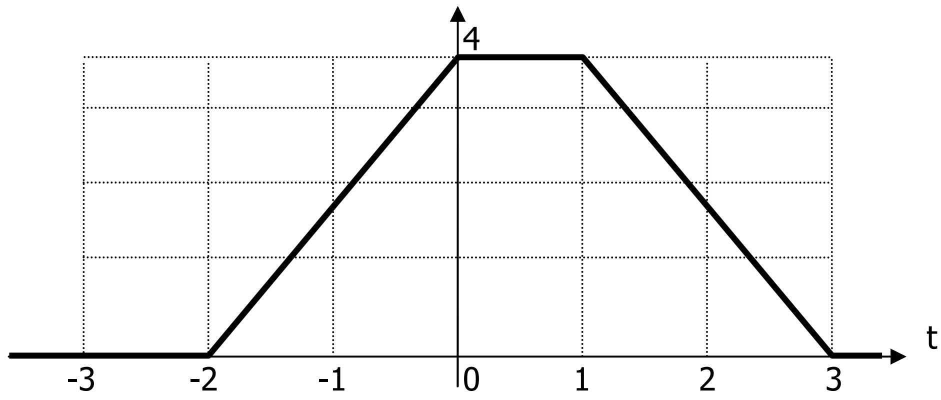
נניח כי אות הכניסה שלנו הוא

$$s(t)=\begin{cases}2 & -1<t<1\\ 0 & \text{אחרת}\end{cases}.$$

נציע פתרון אנליטי ופתרון גרפי של הבעיה. נתחיל ברישום אנליטי של הקונבולוציה:

$$g(t)=\int_{-\infty}^{\infty}h(t-x)s(x)dx\underset{\substack{\uparrow\\ \text{limits}\\ \text{of } s(x)}}{=}2\int_{-1}^{1}h(t-x)dx\underset{\substack{\uparrow\\ t-x=z}}{=}2\int_{t-1}^{t+1}h(z)dz$$

מביטוי זה פשוט יחסית לחשב את g(t) – זהו השטח מתחת תגובת ההלם באינטרוול סימטרי סביב t ברוחב 2, ולכן המוצא ייראה כך:



בהתייחס לציור הבא, ברור כי עבור $t<-2$ המוצא הוא אפס. כך גם הדבר עבור $t>3$. ערכי השיא במוצא מתקבלים עבור מצב של חפיפה מלאה בין שתי הפונקציות וזה קורה עבור $0<t<1$, ואז הערך הוא 4. בתחום $-2<t<0$ מטפס ערך הקונבולוציה ליניארית עם הגדלה של החפיפה, מאפס עד 4. באופן דומה, בתחום $1<t<3$ דועך המוצא מ-4 לאפס.

נתאר בקצרה מספר תכונות כלליות של קונבולוציה ללא הוכחתן (זו מיידית מתוך ההגדרה):

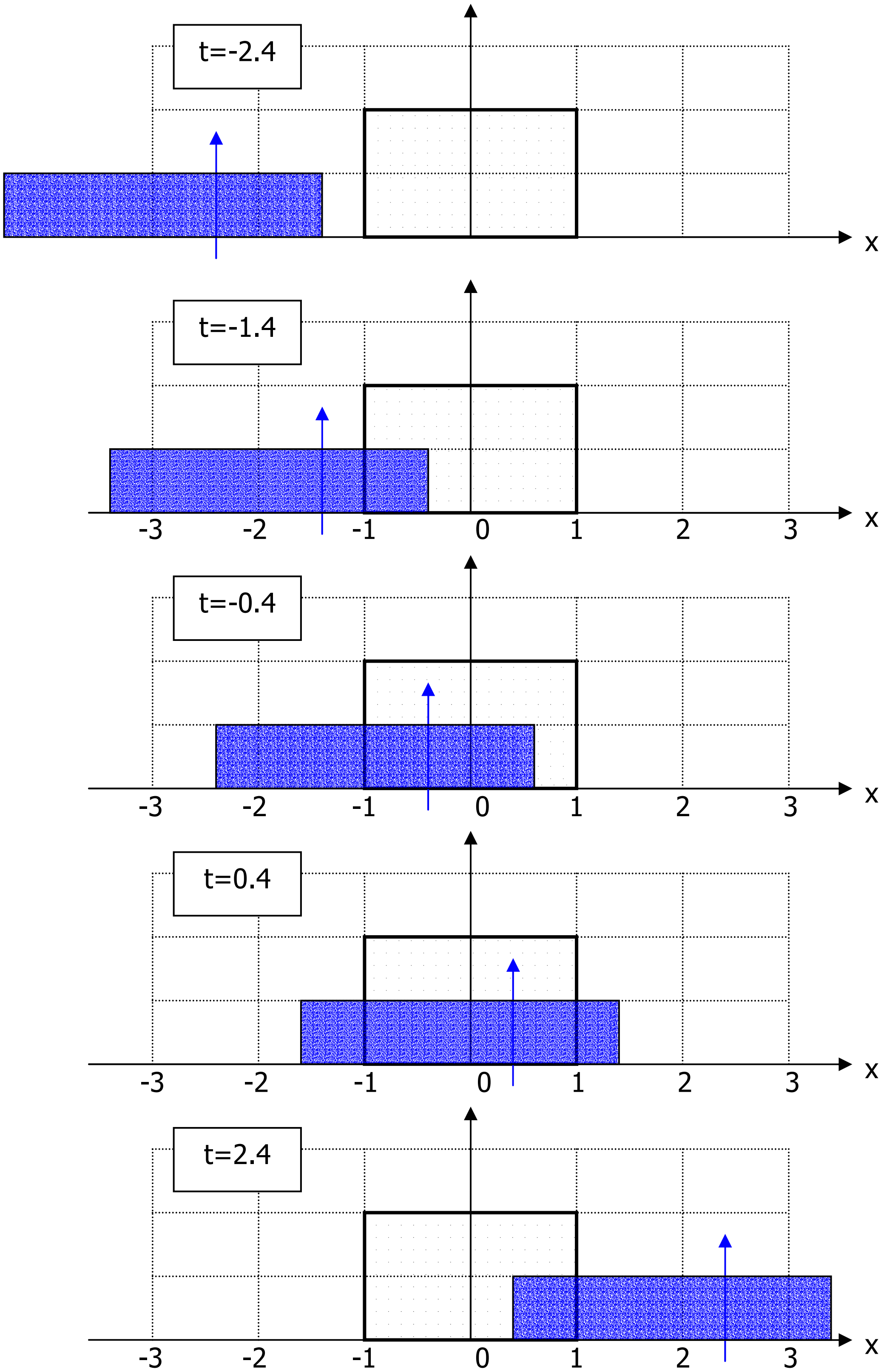
משפט 1: פעולת הקונבולוציה מקיימת את התכונות הבאות -

- קומוטטיביות - $s(t)\otimes g(t)=g(t)\otimes s(t)$.
- אסוציטיביות - $s(t)\otimes\left(g(t)\otimes f(t)\right)=\left(s(t)\otimes g(t)\right)\otimes f(t)$.
- דיסטריביוטיות - $s(t)\otimes\left[g(t)+f(t)\right]=s(t)\otimes g(t)+s(t)\otimes f(t)$.
- מכפלה בסקלר - $a\cdot\left[s(t)\otimes g(t)\right]=\left[as(t)\right]\otimes g(t)=\left[ag(t)\right]\otimes s(t)$
- הזזה - $s(t)\otimes g(t-a)=s(t-a)\otimes g(t)$.
- גזירה - $\left[s(t)\otimes g(t)\right]'=\left[s(t)\right]'\otimes g(t)=\left[g(t)\right]'\otimes s(t)$.

תכונה בסיסית של מערכות ליניאריות קבועות בזמן היא היותן משתחלפות. אנו נפגוש תכונה זו במספר צורות, ולמעשה, עניינו במערכות מסוג זה נובע בין היתר מתכונה חשובה זו. למעשה, מתכונות הקונבולוציה קל לראות כי התכונה הנ"ל מתקיימת כיוון ש:

$$h_2(t)\otimes\left(h_1(t)\otimes s(t)\right)=\left(h_1(t)\otimes h_2(t)\right)\otimes s(t)=\left(h_2(t)\otimes h_1(t)\right)\otimes s(t).$$

כשעשינו שימוש בתכונות האסוציאטיביות והקומוטטיביות.


t=-2.4
x
t=-1.4
-3
-2
-1
0
1
2
3
x
t=-0.4
-3
-2
-1
0
1
2
3
x
t=0.4
-3
-2
-1
0
1
2
3
x
t=2.4
-3
-2
-1
0
1
2
3
x

תכונה שנייה בעלת חשיבות היא היכולת להבחין אם מערכת סיבתית עפ"י תגובתה להלם. אם התגובה להלם היא אפס בזמנים שליליים, בהכרח מתקיים כי המערכת סיבתית. זאת כיוון שעבור תגובה להלם h(t) המתאפסת לזמנים שליליים מתקבל כי

$$g(t) = h(t) \otimes s(t) = \int_{-\infty}^{\infty} h(x)s(t-x)dx \underset{t<0:\ h(x)=0}{\underset{\uparrow}{=}} \int_{0}^{\infty} h(x)s(t-x)dx ,$$

כלומר גבולות האינטגרציה משתנים. לכן מוצא המערכת בזמן t תלוי אך ורק בערכי הכניסה בזמנים t-x עבור x חיובי, דהיינו – סיבתיות.

6. התמרת הפורייה האינטגראלית

אנו עוסקים כעת במערכות ליניאריות וקבועות בזמן. קיימת משפחה של אותות s(t) מעניינים אשר מצליחים לעבור מערכות מסוג זה "ללא פגע", כלומר מוצא המערכת נראה כמו s(t) למעט שינוי בגודל

$$g(t) = T\{s(t)\} = a \cdot s(t) .$$

משוואה זו מזכירה מאוד את המשוואה המגדירה ערכים ווקטורים עצמיים וזה לא מקרה! כאן אנו מתעניינים באותות עצמיים ובערכים התואמים להם המתייחסים למערכות מסוג זה. או נראה כי אותות אלו מאפשרים ניתוח מאוד יעיל של מערכות ליניאריות וקבועות בזמן.

אחת התגליות של ג'וסף פורייה (1800~) היא שמשפחת אותות כאלה קיימת. משפחה זו היא משפחת האותות

$$s(t) = \exp\{j2\pi ft\} = \cos(2\pi ft) + j\sin(2\pi ft) .$$

זהו אות מרוכב הרמוני (מחזורי) בעל תדר f – כלומר ביחידת זמן אחת הוא משלים f מחזורים שלמים. נזין אות זה למערכת ליניארית וקבועה בזמן המאופיינת ע"י תגובתה להלם h(t) ונקבל

$$g(t) = h(t) \otimes s(t) = \int_{-\infty}^{\infty} h(x)s(t-x)dx =$$

$$= \int_{-\infty}^{\infty} h(x)\exp\{j2\pi f(t-x)\}dx =$$

$$= \exp\{j2\pi ft\}\int_{-\infty}^{\infty} h(x)\exp\{-j2\pi fx\}dx = H(f) \cdot \exp\{j2\pi ft\}$$

קיבלנו כי מוצא המערכת גם הוא אות מרוכב בתדר f וכל שהשתנה הוא שהוא כעת מוכפל במספר קבוע התלוי בשני גורמים - תגובת ההלם והתדר f. התקבל

ביטוי מעניין המכפיל את אות הכניסה – ביטוי אינטגרלי זה לוקח את התגובה להלם שהינה פונקציה של הזמן, וממיר אותה לפונקציה חדשה שהיא פונקציה של התדר.

ביטוי זה מוכר בשם "התמרת פורייה", והדרך בה הצגנו אותה כאן הייתה גורמת לפורייה (ראה תמונה) להתהפך בקברו! יוסף פורייה הוא מתמטיקאי צרפתי בן המאה השמונה-עשרה אשר פיתח התמרה זו לצד תוצאות מתמטיות רבות אחרות. להתמרה זו חשיבות רבה בעיבוד אותות, בשל היותה מקילה לניתוח ואפיון של מערכות.

נסכם אם כך: בהינתן מערכת ליניארית וקבועה בזמן בעלת התגובה להלם $h(t)$, התמרת הפורייה של התגובה להלם המסומנת $H(f)$ היא פונקציה של התדר המוגדרת ע"י

$$H(f) = F\{h(t)\} = \int_{-\infty}^{\infty} h(t)\exp\{-j2\pi ft\}dt$$

לפעמים בספרות הגדרה זו ניתנת עם מקדם אך אנו נתעלם מכך כאן.

שאלה ראשונה שעלינו לשאול היא – האם תמיד נוכל לחשב את התמרת הפורייה? האם זו מוגדרת לכל פונקצית תגובה להלם $h(t)$? התשובה לכך היא שלפונקציות המקיימות

$$\int_{-\infty}^{\infty} |h(t)|^2 dt < \infty \text{ ,}$$

התמרת פורייה מוגדרת היטב. דרישה זו פירושה שאנו עוסקים בפונקציות השייכות למרחב הילברט המסומן L^2 (שאומר כי החברים בו בעלי אינטגרל סופי כשזה מחושב על ערך מוחלט של ריבועם). אנו לא נוכיח קביעה זו כאן.

שאלה שנייה שיש לשאול היא – האם הפונקציה $H(f)$ מהווה ייצוג מלא של המערכת, ממש כשם ש- $h(t)$ הינה? במילים אחרות, האם התמרה זו מאבדת מידע? התשובה היא

שאין כל אבדן מידע והדרך להראות זאת היא ע"י תהליך מתמטי מוגדר היטב של התמרה הפוכה אשר יחזיר מ- H(f) ל-h(t). נוסחת ההתמרה ההפוכה הינה

$$F^{-1}\{H(f)\} = \int_{-\infty}^{\infty} H(f)\exp\{j2\pi ft\}df,$$

נראה כעת כי נוסחה זו מובילה להיפוך הנדרש. נציב לתוכה את ההתמרה קדימה, ונראה כי חוזרים לפונקצית המקור,

$$\begin{aligned} F^{-1}\{H(f)\} &= \int_{-\infty}^{\infty} H(f)\exp\{j2\pi ft\}df = \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} h(x)\exp\{-j2\pi fx\}dx\right)\exp\{j2\pi ft\}df = \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x)\exp\{j2\pi f(t-x)\}dxdf = \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t-x)\exp\{j2\pi fx\}dxdf = \\ &= \int_{-\infty}^{\infty} h(t-x)\left(\int_{-\infty}^{\infty}\exp\{j2\pi fx\}df\right)dx = \\ &= \int_{-\infty}^{\infty} h(t-x)\delta(x)dx = h(t). \end{aligned}$$

התהליך הנ"ל למעשה הראה את התמרת הפורייה ההפוכה, המעבירה אותנו מ-H(f) ל-h(t). השתמשנו בתכונה

$$\int_{-\infty}^{\infty}\exp\{j2\pi fx\}df = \delta(x),$$

כלומר, עבור כל x שאינו אפס יתקבל כי אנו סוכמים מעל אות מחזורי ולכן האינטגרל מתאפס. כאשר x=0 אנו סוכמים מעל אות קבוע ואז מקבלים אינסוף.

קיבלנו אם כך דרך אלטרנטיבית לתיאור מערכת ליניארית וקבועה בזמן, ע"י ההתייחסות ל- H(f). מה משמעותה? על מנת להבהיר זאת, נתחיל במשמעות התמרת הפורייה על אות כלשהו s(t):

$$S(f) = F\{s(t)\} = \int_{-\infty}^{\infty} s(t)\exp\{-j2\pi ft\}dt.$$

מהי המשמעות של S(f)? למעשה אנו מתארים את אות הכניסה בדרך שונה אך ממצה. בעוד s(t) אומר לנו כיצד בנוי האות ע"י תיאור ערכיו בכל נקודת זמן t, הפונקציה S(f)

אומרת לנו כיצד בנוי האות כאוסף של אותות הרמוניים $\exp\{j2\pi ft\}$. נראה זאת דרך דוגמא פשוטה – נניח עי האות שלנו כולל הרמוניה אחת בתדר f_0 והוא

$$s(t) = \exp\{j2\pi f_0 t\},$$

אזי התמרת הפורייה שלו היא

$$S(f) = \int_{-\infty}^{\infty} s(t)\exp\{-j2\pi ft\}dt = \int_{-\infty}^{\infty} \exp\{-j2\pi(f-f_0)t\}dt = \delta(f-f_0).$$

כלומר, מהתמרת הפורייה אנו מזהים כי מדובר באות בעל הרמוניה יחידה, ואשר תדרה הוא f_0. באופן דומה, אם האות שלנו הוא אות סינוסואידלי רציף:

$$s(t) = \sin(2\pi f_0 t) = \frac{\exp\{j2\pi f_0 t\} - \exp\{-j2\pi f_0 t\}}{2j},$$

אזי התמרת הפורייה שלו היא

$$\begin{aligned} S(f) &= \int_{-\infty}^{\infty} s(t)\exp\{-j2\pi ft\}dt = \\ &= \int_{-\infty}^{\infty} \frac{\exp\{j2\pi(f_0-f)t\} - \exp\{-j2\pi(f+f_0)t\}}{2j}dt = \frac{1}{2j}\left[\delta(f-f_0) - \delta(f+f_0)\right] \end{aligned}$$

כלומר, לאות זה יש שתי הרמוניות בתדרים f_0 ו- $-f_0$.

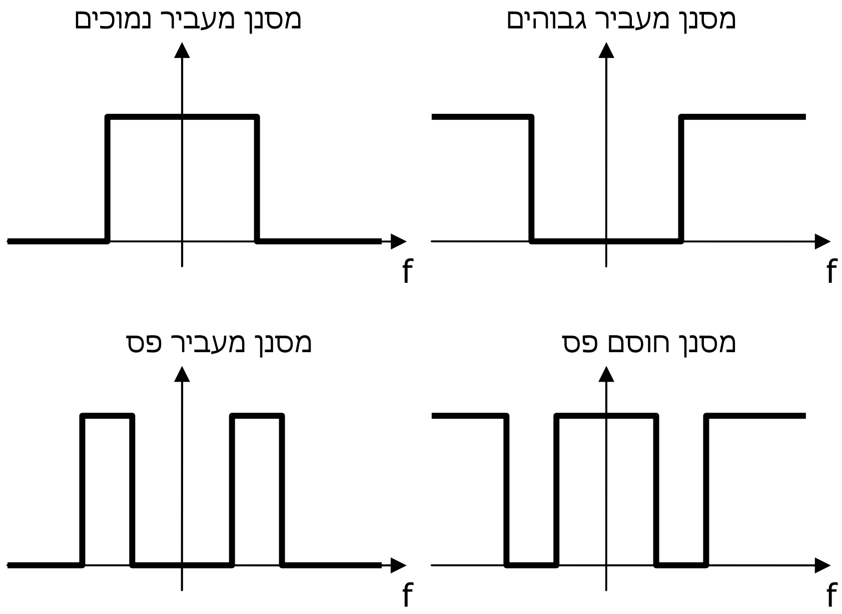
נחזור לשאלה המקורית: בהינתן מערכת עם $H(f)$ כהתמרת הפורייה של תגובתה להלם, מה משמעות של פונקציה זו? אם נזין למערכת זו אות הרמוני יחיד בתדר f_0 ראינו כבר כי מוצאה יהיה אותו אות (באותו תדר) כשהוא מוכפל ב-$H(f)$, כלומר, $H(f)$ היא תגובת התדר של המערכת הנתונה. לדוגמה, אם $H(f)$ מתאפסת לטווח מסוים של תדרים, פירוש הדבר הוא שהמערכת הנתונה לנו חוסמת תדרים אלו. מקובל לצייר את הערך המוחלט של פונקצית תגובת התדר כדי לתאר את פעולת המערכת - ראה את הציור הבא לתיאור מסננים קלאסיים.

גם לפאזה (הרי $H(f)$ מרוכב!) יש תפקיד חשוב, והיא קובעת כיצד מוזז האות במוצא. לדוגמה, אם $H(f_0) = 0.01 \cdot \exp\{j\pi/4\}$, פירוש הדבר שמוצא המערכת בתדר זה יונחת פי 100, ויוסט בכדי רבע מחזור קדימה, כיוון ש:

$$\begin{aligned} H(f_0)\exp\{j2\pi f_0 t\} &= 0.01 \cdot \exp\{j\pi/4\}\exp\{j2\pi f_0 t\} = \\ &= 0.01\left[\cos(2\pi f_0 t + \pi/4) + j\sin(2\pi f_0 t + \pi/4)\right] \end{aligned}$$

כאשר נזין למערכת אות כלשהו $s(t)$, הרי שראינו כי ניתן לפרקו לצירוף ליניארי של אותות הרמוניים. בשל ליניאריות המערכת, כל אחד מאלה יוכפל בתגובת התדר בתדר המתאים לו, וכך ייקבע המוצא. אנו נראה מיד דרך פשוטה לניסוח תכונה זו באופן מתמטי נקי, והיא אחת הסיבות לפופולאריות של התמרת הפורייה.

באנלוגיה לדיון הקודם בערכים עצמיים, קיבלנו כאן כי למערכות ליניאריות וקבועות בזמן יש משפחת אותות עצמיים שעוברים דרך המערכת ללא פגע – כלומר מוצא המערכת עבורם זהה לכניסתנו עד כדי הכפלה בקבוע. אוסף קבועים אלו מתקבל ע"י התמרת הפורייה, כך שאלו הם הערכים העצמיים של המערכת.



7. תכונות התמרת פורייה האינטגראלית

כעת נציג עם הוכחות בסיסיות סידרה של תכונות של התמרת הפורייה. מסתבר כי בעזרת תכונות אלה ניתן לחשב התמרה של אות בעל הרכב מורכב, ע"י פירוקו.

משפט 2: התמרת פורייה האינטגראלית ותכונותיה -

- הגדרה - עבור האות $g(t)$ התמרת הפורייה נתונה ע"י

$$G(f) = F\{g(t)\} = \int_{-\infty}^{\infty} g(t)\exp\{-j2\pi ft\}dt ,$$

וההתמרה חזרה ע"י $g(t) = F^{-1}\{G(f)\} = \int_{-\infty}^{\infty} G(f)\exp\{j2\pi ft\}df$.

- ליניאריות – התמרת הפורייה ליניארית, כלומר, עבור כל שני אותות $s(t)$ ו-$g(t)$, ושני סקלרים a ו-b, מתקיים כי

$$F\{a\cdot s(t)+b\cdot g(t)\} = \int_{-\infty}^{\infty}\left(a\cdot s(t)+b\cdot g(t)\right)\exp\{-j2\pi ft\}dt =$$

$$= a\cdot\int_{-\infty}^{\infty}s(t)\exp\{-j2\pi ft\}dt + b\cdot\int_{-\infty}^{\infty}g(t)\exp\{-j2\pi ft\}dt =$$

$$= a\cdot F\{s(t)\}+b\cdot F\{g(t)\}$$

תכונה זו נובעת מהיות ההתמרה מבוססת על סכימה אינטגראלית, בה ניתן להוציא מקדמים החוצה ולפרק הסכימה לכל נתח בנפרד.

- התמרה פעמיים – עבור האות $g(t)$ שהתמרתו היא $G(f)$, יתקבל

$$F\{G(f)\} = \int_{-\infty}^{\infty}G(f)\exp\{-j2\pi ft\}df = g(-t).$$

- ממוצע – ממוצע האות $g(t)$ אינו אלא $G(0)$:

$$\int_{-\infty}^{\infty}g(t)dt = \int_{-\infty}^{\infty}g(t)\exp\{-j2\pi t\cdot 0\}dt = F\{g(t)\}\big|_{f=0} = G(0).$$

- ממשיות וסימטריות – בהינתן לנו אות $g(t)$ ממשי טהור (כלומר ללא מרכיב מדומה), להתמרת הפורייה שלו מבנה סימטרי צמוד,

$$G(f) = \int_{-\infty}^{\infty}g(t)\exp\{-j2\pi ft\}dt$$

$$\overline{G}(f) = \overline{\int_{-\infty}^{\infty}g(t)\exp\{-j2\pi ft\}dt} = \int_{-\infty}^{\infty}g(t)\exp\{j2\pi ft\}dt = G(-f)$$

השתמשנו בתכונה של פעולת הצמוד שהיא מתחלפת עם סכימה. משמעות תוצאה זו היא שעבור אותות ממשיים לא צריך לדעת את ערכי $G(f)$ לערכי f שליליים כיוון שאלו משתמעים מערכי G בתדרים החיוביים. מסיבה זו, אגב, המסננים הקנוניים שהוצגו קודם סימטריים סביב האפס בערכם המוחלט – אנו מניחים כי תגובתם להלם ממשית.

לתוצאה זו גם אפקט הפוך – אם $g(t)$ הוא אות סימטרי צמוד, כלומר $\overline{g(t)} = g(-t)$, יתקבל כי התמרת הפורייה ממשית, כיוון ש:

$$G(f) = \int_{-\infty}^{\infty}g(t)\exp\{-j2\pi ft\}dt = \int_{0}^{\infty}g(t)\exp\{-j2\pi ft\}dt + \int_{-\infty}^{0}g(t)\exp\{-j2\pi ft\}dt =$$

ע"י

$$\underset{u=-t}{\overset{}{\underset{\uparrow}{=}}} \int_{0}^{\infty}g(t)\exp\{-j2\pi ft\}dt + \int_{0}^{\infty}g(-u)\exp\{j2\pi fu\}du$$

שינוי משתנה בחלק הסכימה השני נקבל

$$G(f) \underset{u=-t}{\overset{\uparrow}{=}} \int_{0}^{\infty} g(t)\exp\{-j2\pi ft\}dt + \int_{0}^{\infty} \overline{g(t)}\exp\{j2\pi ft\}du =$$

$$\underset{u=-t}{\overset{\uparrow}{=}} \int_{0}^{\infty} g(t)\exp\{-j2\pi ft\}dt + \overline{\int_{0}^{\infty} g(t)\exp\{-j2\pi ft\}du}$$

בתוצאה זו השתמשנו בעובדה שסכום של שני מספרים צמודים זה לזה הוא מספר ממשי טהור.

כמקרה אחרון מעניין, אם האות הנתון לנו $g(t)$ גם ממשי וגם סימטרי יתקבל כי $G(f)$ ממשי וסימטרי אף הוא. הוכחה לכך לא נחוצה כיוון שהיא בנויה על שתי הטענות הקודמות.

- הזזה ואפנון – אם ידוע לנו כי התמרתו של $g(t)$ היא $G(f)$, ההתמרה של אות מוזז תתקבל כשהיא מאופננת:

$$\int_{-\infty}^{\infty} g(t-d)\exp\{-j2\pi ft\}dt \underset{t-d=u}{\overset{\uparrow}{=}} \int_{-\infty}^{\infty} g(u)\exp\{-j2\pi f(u+d)\}du =$$

$$= \exp\{-j2\pi fd\}\int_{-\infty}^{\infty} g(t)\exp\{-j2\pi ft\}dt = \exp\{-j2\pi fd\}G(f)$$

משמעות האפנון הפעלת התמרת הפורייה בפונקציה הרמונית בתדר עם זמן קבוע ושווה למידת ההשהיה d. באופן דומה, אם לקחנו את האות $g(t)$ ואפננו אותו ע"י הכפלתו ב- $\exp\{-j2\pi f_0 t\}$, ההתמרה תהיה מוזזת -

$$\int_{-\infty}^{\infty} g(t)\exp\{-j2\pi f_0 t\}\exp\{-j2\pi ft\}dt =$$

$$= \int_{-\infty}^{\infty} g(t)\exp\{-j2\pi(f+f_0)t\}dt = G(f+f_0)$$

- הגדלה – אם אות הכניסה שלנו עובר שינויי סקלת ציר זמן ההתמרה תשנה סקלה בצורה דומה והפוכה לה. בהנחה ששינוי הסקלה a הוא חיובי נקבל

$$F\{g(at)\} = \int_{-\infty}^{\infty} g(at)\exp\{-j2\pi ft\}dt$$

$$\underset{at=u}{\overset{\uparrow}{=}} \frac{1}{a}\int_{-\infty}^{\infty} g(u)\exp\{-j2\pi fu/a\}du = \frac{1}{a}G(f/a)$$

- גזירה – אם ידוע כי $g(t)$ ו- $G(f)$ הם צמד פורייירי ונניח כי ידוע כי $g(t)$ מתאפס לקראת האינסוף משני הכיוונים. אזי התמרת הפורייה של נגזרתו של $g(t)$ תהיה

$$F\left\{\frac{d}{dt}g(t)\right\} = \int_{-\infty}^{\infty} \frac{d}{dt}g(t)\exp\{-j2\pi ft\}dt =$$

$$= \left[g(t)\exp\{-j2\pi ft\}\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{d}{dt}\exp\{-j2\pi ft\}g(t)dt =$$

$$= j2\pi f \cdot \int_{-\infty}^{\infty} \exp\{-j2\pi ft\}g(t)dt = j2\pi f \cdot G(f)$$

נשים לב לכך שהנחנו כי הגורם הראשון באינטגרציה בחלקים מתאפס, והנחה זו נובעת מכך שהפונקציה $g(t)$ דועכת לקראת האינסוף.

ישנן עוד תכונות חשובות ואותן נעלה בהמשך. מהתכונות הנ"ל חשוב להבין כי לא פעם בהינתן פונקציה, חישוב התמרתה לא צריך להיעשות ע"י ההגדרה ישירות. כתחליף, יש להשתמש בתכונות שתוארו קודם, ובצמדים הבסיסיים המוכרים. בטבלה הבאה אנו מציגים מספר דוגמאות לצמדים אופייניים לאות $g(t)$ והתמרתו $G(f)$.

האות $g(t)$	ההתמרה $G(f)$
$\delta(t)$	1
$\delta(t-t_0)$	$\exp\{-j2\pi ft_0\}$ [תכונת ההזזה]
1	$\delta(f)$
$\exp\{-j2\pi f_0 t\}$	$\delta(f-f_0)$ [תכונת האפנון]
$\cos(2\pi f_0 t)$	$\left[\delta(f-f_0)+\delta(f+f_0)\right]/2$
$\sin(2\pi f_0 t)$	$\left[\delta(f-f_0)-\delta(f+f_0)\right]/2j$
$\exp(-2\pi a\|t\|)$	$a/(a^2+f^2)/\pi$
$\exp(-at^2)$ [גאוסיאן]	$\sqrt{\pi/a}\cdot\exp(-\pi^2 f^2/a)$
$\mu(t+d)-\mu(t-d)$ [פונקצית מלבן]	$2d\cdot\sin c\{2\pi fd\}$
tri(t,2d) [פונקצית משולש]	$(2d\cdot\sin c\{2\pi fd\})^2$

הצמד הקשור לפונקצית המלבן מתקבל באופן הבא:

$$\int_{-\infty}^{\infty}\left(\mu(t+d)-\mu(t-d)\right)\exp\{-j2\pi ft\}dt=\int_{-d}^{d}\exp\{-j2\pi ft\}dt=\left[\frac{\exp\{-j2\pi ft\}}{-j2\pi f}\right]_{-d}^{d}=$$
$$=\frac{\exp\{-j2\pi fd\}}{-j2\pi f}-\frac{\exp\{j2\pi fd\}}{-j2\pi f}=$$
$$=\frac{\exp\{j2\pi fd\}-\exp\{-j2\pi fd\}}{j2\pi f}=$$
$$=\frac{\sin\{2\pi fd\}}{\pi f}=2d\cdot \operatorname{sinc}\{2\pi fd\}$$

כשאנו מגדירים $\operatorname{sinc}(x)=\sin(x)/x$. פונקצית המשולש מוגדרת ע"י

$$s(t)=\begin{cases}0 & |t|>2d\\ 2d-t & 0\le t\le 2d\\ t+2d & -2d\le t\le 0\end{cases}$$

ואת התמרתה נצדיק בהמשך, כשנדון בתכונת הקונבולוציה.

תכונה מאוד חשובה של התמרת הפורייה היא שימור אנרגיה. תכונה זו, המוכרת בשם "משפט פרסבל", טוענת כי

$$.\ \int_{-\infty}^{\infty}|g(t)|^2dt=\int_{-\infty}^{\infty}|G(f)|^2df$$

תכונה זו היא למעשה מקרה פרטי של תכונה אחרת חשובה – שימור מכפלה פנימית, אשר אותה נוכיח.

משפט 3: עבור שני אותות $g(t)$ ו-$s(t)$ והתמרתם – $G(f)$ ו-$S(f)$, מכפלה פנימית בין שתי פונקציות בזמן מוגדרת ע"י

$$.\langle g(t),s(t)\rangle_t=\int_{-\infty}^{\infty}g(t)\overline{s(t)}dt$$

מתקיים כי מכפלה פנימית בתדר של פונקציות ההתמרה תוביל לאותה תוצאה,

$$\langle G(f),S(f)\rangle_f=\int_{-\infty}^{\infty}G(f)\overline{S(f)}df=\langle g(t),s(t)\rangle_t=\int_{-\infty}^{\infty}g(t)\overline{s(t)}dt$$

הוכחת תכונה זו נעשית ע"י שינוי סדר הסכימה

$$\left\langle G(f),S(f)\right\rangle_f = \int_{-\infty}^{\infty} G(f)\overline{S(f)}df = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t)\overline{s(u)}\left(\int_{-\infty}^{\infty}\exp\{j2\pi f(u-t)\}df\right)dudt =$$

$$= \int_{-\infty}^{\infty} g(t)\left(\int_{-\infty}^{\infty}\overline{s(u)}\delta(u-t)du\right)dt = \int_{-\infty}^{\infty} g(t)\overline{s(t)}dt = \left\langle g(t),s(t)\right\rangle_t$$

לכן, כמקרה פרטי, הצבת $s(t)=g(t)$ תיתן את תוצאת פרסבל:

$$\int_{-\infty}^{\infty}|g(t)|^2dt = \int_{-\infty}^{\infty}|G(f)|^2df\ .$$

תכונה חשובה אחרת היא תכונת הקונבולוציה, וגרסה דומה לה הקרויה תכונת הקורלציה, אשר דומה לתכונת המכפלה הפנימית שהוצגה קודם.

משפט 4: עבור שני אותות $g(t)$ ו-$s(t)$, והתמרתם – $G(f)$ ו-$S(f)$, הקונבולוציה בין שתי הפונקציות בזמן נתונה ע"י

$$g(t)\otimes s(t) = \int_{-\infty}^{\infty} g(u)s(t-u)du\ .$$

מתקיים כי התמרת הפורייה של הקונבולוציה מומרת למכפלה רגילה.

הוכחת תכונה זו קלה, וגם היא נובעת ישירות מההגדרה,

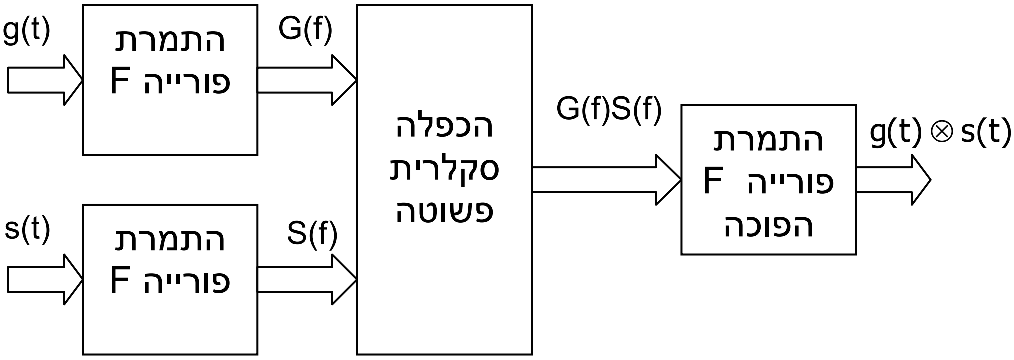
$$F\{g(t)\otimes s(t)\} = F\left\{\int_{-\infty}^{\infty} g(u)s(t-u)du\right\} = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} g(u)s(t-u)du\right)\cdot\exp\{-j2\pi ft\}dt =$$

$$= \int_{-\infty}^{\infty} g(u)\left(\int_{-\infty}^{\infty} s(t-u)\cdot\exp\{-j2\pi ft\}dt\right)du =$$

$$\underset{t-u=z}{\underset{\uparrow}{=}} \int_{-\infty}^{\infty} g(u)\left(\int_{-\infty}^{\infty} s(z)\cdot\exp\{-j2\pi f(z+u)\}dz\right)du =$$

$$= \left(\int_{-\infty}^{\infty} g(u)\exp\{-j2\pi fu\}du\right)\left(\int_{-\infty}^{\infty} s(z)\cdot\exp\{-j2\pi fz\}dz\right) = G(f)S(f)$$

משמעות תוצאה זו היא שבמקום לחשב קונבולוציה (עניין מורכב בד"כ), נוכל לעבור להתמרות הפורייה של שני האותות המעורבים, להכפילם זה בזה סקלרית – עניין קל – ואז לבצע התמרה הפוכה חזרה לזמן.



כדוגמה, עבור אות כניסה מלבני $s(t)=\mu(t+d)-\mu(t-d)$ המוזן למערכת בעת תגובה להלם $h(t)=\mu(t+d)-\mu(t-d)$, מוצא המערכת יהיה

$$F^{-1}\{F\{\mu(t+d)-\mu(t-d)\}\cdot F\{\mu(t+d)-\mu(t-d)\}\}=F^{-1}\left\{\left(\frac{\sin\{2\pi fd\}}{\pi f}\right)^2\right\}$$

אם נבצע את הקונבולוציה בצורה ישירה נקבל את פונקצית המשולש, כיוון ש:

$$\int_{-\infty}^{\infty}(\mu(u+d)-\mu(u-d))(\mu(t-u+d)-\mu(t-u-d))du=$$

$$=\int_{-d}^{d}(\mu(t-u+d)-\mu(t-u-d))du=\int_{-d}^{d}\mu(t-u+d)du-\int_{-d}^{d}\mu(t-u-d)du=$$

$$\underset{z=t-u+d}{\overset{\uparrow}{=}}\int_{t}^{t+2d}\mu(z)dz\ \underset{z=t-u-d}{\overset{\uparrow}{-}}\int_{t-2d}^{t}\mu(z)dz$$

נשתמש בתכונה שאינטגרל על-פני פונקצית מדרגה מקיים

$$\int_a^b\mu(z)dz=\begin{cases}b-a & b>a>0\\ b & b>0>a\\ 0 & 0>b>a\end{cases}$$

ונקבל את פונקצית המשולש, כי

$$\int_{t}^{t+2d}\mu(z)dz\ \underset{z=t-u-d}{\overset{\uparrow}{-}}\int_{t-2d}^{t}\mu(z)dz=\begin{cases}2d & t>0\\ t+2d & 0\ge t>-2d\\ 0 & -2d\ge t\end{cases}-\begin{cases}2d & t>2d\\ t & 2d\ge t>0\\ 0 & 0\ge t\end{cases}=$$

$$=\begin{cases}2d-2d=0 & t>2d\\ 2d-t & 2d\ge t>0\\ t+2d & 0\ge t>-2d\\ 0 & -2d>t\end{cases}$$

תכונה דומה היא תכונת הקורלציה. פעולת הקורלציה בין אותות מוגדרת ע"י האינטגרל:

$$.\,\mathrm{Corr}\{g(t), s(t)\} = \int_{-\infty}^{\infty} g(u)\overline{s(u-t)}du$$

חישוב קורלציה חשוב בעיבוד אותות ותמונות – הקורלציה מראה כמה דומים שני אותות. התמרת תדר מובילה ל:

$$F\{\mathrm{Corr}\{g(t), s(t)\}\} = F\left\{\int_{-\infty}^{\infty} g(u)\overline{s(u-t)}du\right\} = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} g(u)\overline{s(u-t)}du\right)\cdot \exp\{-j2\pi ft\}dt =$$

$$= \int_{-\infty}^{\infty} g(u)\left(\int_{-\infty}^{\infty}\overline{s(u-t)}\cdot \exp\{-j2\pi ft\}dt\right)du =$$

$$= \int_{-\infty}^{\infty} g(u)\exp\{-j2\pi fu\}\left(\int_{-\infty}^{\infty}\overline{s(u-t)}\cdot \exp\{-j2\pi f(t-u)\}dt\right)du =$$

$$= \left(\int_{-\infty}^{\infty} g(u)\exp\{-j2\pi fu\}du\right)\left(\int_{-\infty}^{\infty}\overline{s(t)}\cdot \exp\{j2\pi ft\}dt\right) =$$

$$= \left(\int_{-\infty}^{\infty} g(u)\exp\{-j2\pi fu\}du\right)\overline{\left(\int_{-\infty}^{\infty} s(t)\cdot \exp\{-j2\pi ft\}dt\right)} = G(f)\overline{S(f)}$$

המסקנה היא שגם קורלציה יכולה להיעשות בתדר במכפלה פשוטה.

שיטות מתמטיות ביישומי מחשב (234299)
פרק 10 – התמרת פורייה בשריג בדיד

נושאים שייסקרו:

1. אותות מחזוריים וטורי פורייה
2. משפט הדגימה ומשמעותו
3. עיבוד אותות דגומים – מערכות
4. עיבוד אותות דגומים – התמרת פורייה
5. טיפול באותות בעלי תמך סופי
6. חזרה למטריצות ווקטורים
7. התמרת פורייה דיסקרטית מהירה - FFT

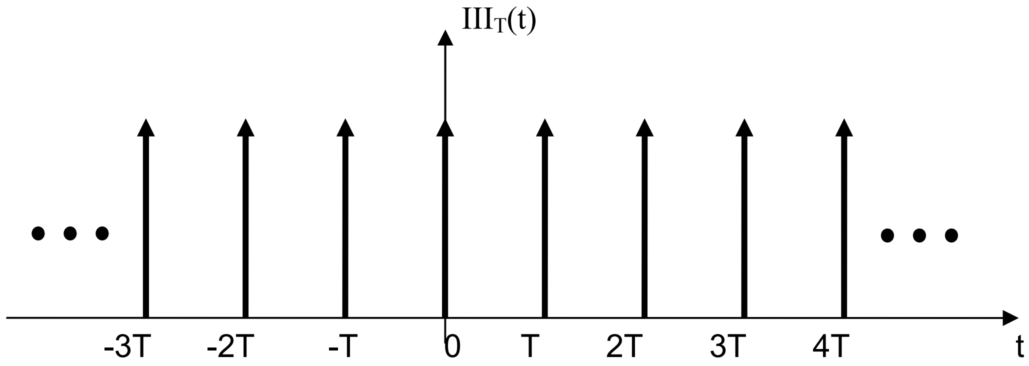
---

1. אותות מחזוריים וטורי פורייה

פונקציה מעניינת שתסייע לנו היא "רכבת הלמים", או כפי שמתמטיקאים קוראים לה, פונקצית שה (Shah) - שם זה נובע מהדמיון של הסימון לאות "ש" בעברית, וגרסאות דומות לה בקירילית ובשפת חרטומים מצריים. פונקציה זו נתונה כטור

$$\mathrm{III}_T(t) = \sum_{k=-\infty}^{\infty} \delta(t-kT),$$

והיא נראית במתואר בציור הבא:



התמרת הפורייה של פונקציה זו הינה גם היא רכבת הלמים. שימוש בהגדרה נותן

$$F\{\mathrm{III}_T(t)\} = F\left\{\sum_{k=-\infty}^{\infty} \delta(t-kT)\right\} = \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \delta(t-kT)\exp\{-j2\pi ft\}dt =$$

$$= \sum_{k=-\infty}^{\infty}\left(\int_{-\infty}^{\infty} \delta(t-kT)\exp\{-j2\pi ft\}dt\right) =$$

$$= \sum_{k=-\infty}^{\infty} \exp\{-j2\pi fkT\}$$

עבור תדרים המהווים מכפלה שלמה של $1/T$ סכימה זו תצטבר לאינסוף כיוון שכל אחד מהאקספוננטים יהיה 1, בעוד שלכל בחירת תדר אחרת סכימה זו תתאפס. לכן,

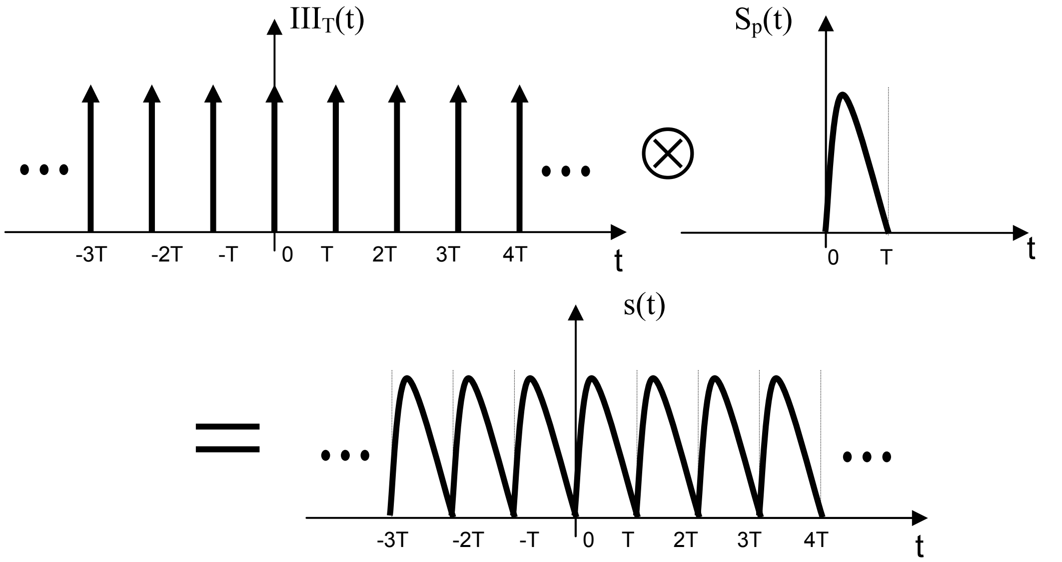
$$F\{III_T(t)\} = F\left\{\sum_{k=-\infty}^{\infty}\delta(t-kT)\right\} = \sum_{k=-\infty}^{\infty}\exp\{-j2\pi fkT\} = \frac{1}{T}\sum_{n=-\infty}^{\infty}\delta\left(f-\frac{n}{T}\right).$$

עניין המקדם $1/T$ נובע מתכונת שימור האנרגיה של פרסבל. בשני המקרים יש סכימה על אינסוף הלמים, אך בתדר יש הלם כל $1/T$ יחידות תדר, בעוד שבזמן הם מופיעים כל T. לכן יש פי T^2 יותר הלמים בתדר. עם הכנסת המקדם הנ"ל סך האנרגיה בשני האותות משתווה (יש לזכור כי בפרסבל אנו מעלים את האות בריבוע, ולכן נחוץ שורש היחס הנ"ל).

תוצאה זו מעניינת מכמה סיבות – קודם כל, מעניין לדעת מיהן הפונקציות שהתמרת הפורייה לא משנה את צורתן. ראינו קודם כי גאוסייאן הוא כזה, ועתה מצאנו פונקציה אחרת בעלת אותה תכונה. מעבר לכך שנה תכונה חשובה יותר לתוצאה הנ"ל, והיא מובילה למושג טורי פורייה שאותם נציג כעת.

כל אות מחזורי $s(t)$ במרווח T ניתן לתיאור כקונבולוציה של האות במחזור אחד בקטע $[0,T]$ עם הפונקציה $III_T(t)$. נסמן את המחזור היחיד באינטרוול $[0,T]$ ע"י $s_p(t)$ (p בשביל period – מחזור יחיד). אזי אנו בעצם קובעים כי

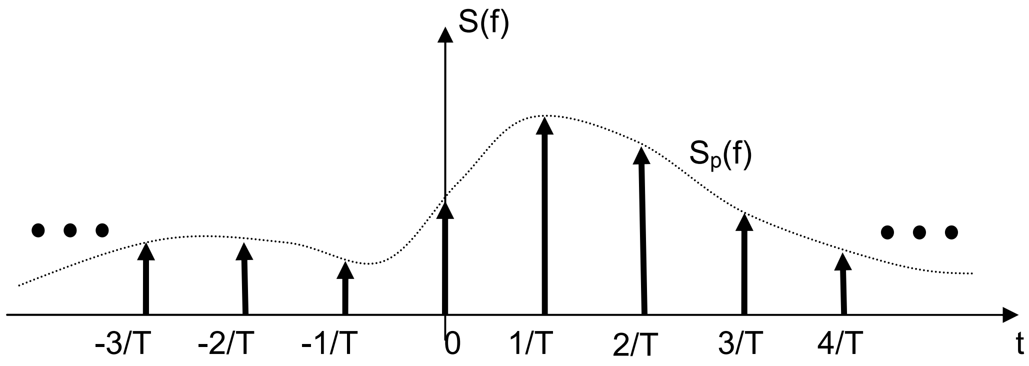
$$s(t) = s_p(t) \otimes III_T(t).$$



תכונה זו חשובה כיוון שכעת נוכל לקבוע את התמרת הפורייה של אות מחזורי כזה. התמרה זו נתונה בשל תכונת הקונבולוציה כמכפלה בין רכבת הלמים בתדר במרווחים 1/T והתמרת הפורייה של מחזור יחיד. לכן, התמרה זו תהיה

$$F\{s(t)\} = F\{s_p(t) \otimes III_T(t)\} = F\{s_p(t)\} \cdot F\{III_T(t)\} =$$
$$= S_p(f) \cdot \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right) = \frac{1}{T}\sum_{n=-\infty}^{\infty} S_p\left(\frac{n}{T}\right) \cdot \delta\left(f - \frac{n}{T}\right)$$

קיבלנו כי התמרת הפורייה של אות מחזורי אינה קיימת בכל תדר, אלא רק בנקודות תדר שהינן מכפלות שלמות של 1/T. כלומר, אות מחזורי בנוי רק מהרמוניות הקשורות באורך המחזור.



אם נבצע התמרת פורייה הפוכה על התוצאה נקבל כמובן את אות המקור, אך נוכל גם להציגו אחרת – וזוהי הנקודה הראשית בסיפורנו:

$$s(t) = F^{-1}\left\{\frac{1}{T}\sum_{n=-\infty}^{\infty} S_p\left(\frac{n}{T}\right) \cdot \delta\left(f - \frac{n}{T}\right)\right\} = \frac{1}{T}\int_{-\infty}^{\infty}\sum_{n=-\infty}^{\infty} S_p\left(\frac{n}{T}\right) \cdot \delta\left(f - \frac{n}{T}\right)\exp\{j2\pi ft\}df =$$
$$= \frac{1}{T}\sum_{n=-\infty}^{\infty} S_p\left(\frac{n}{T}\right) \cdot \int_{-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right)\exp\{j2\pi ft\}df =$$
$$= \frac{1}{T}\sum_{n=-\infty}^{\infty} S_p\left(\frac{n}{T}\right) \cdot \exp\left\{\frac{j2\pi nt}{T}\right\} = \sum_{n=-\infty}^{\infty} c_n \cdot \exp\left\{\frac{j2\pi nt}{T}\right\}$$

התקבל כי אות המקור ניתן לתיאור כטור של פונקציות הרמוניות מחזוריות בתדרים n/T לכל n שלם, עם מקדמים הניתנים לחישוב ישיר ממחזור יחיד של הפונקציה, לפי

$$.\ c_n = \frac{1}{T}S_p\left(\frac{n}{T}\right) = \frac{1}{T}\int_{-\infty}^{\infty} s_0(t)\exp\{-j2\pi ft\}dt\Big|_{f=n/T} = \frac{1}{T}\int_0^T s(t)\exp\left\{-\frac{j2\pi nt}{T}\right\}dt$$

משפט 1: בהינתן אות מחזורי s(t) בעל מחזור T, ניתן להציגו אלטרנטיבית כסכום אינסופי של פונקציות הרמוניות בתדרים n/T (לכל n שלם):

$$s(t) = \sum_{n=-\infty}^{\infty} c_n \cdot \exp\left\{\frac{j2\pi nt}{T}\right\}.$$

מקדמי טור זה נתונים ע"י

$$c_n = \frac{1}{T}\int_0^T s(t)\exp\left\{-\frac{j2\pi nt}{T}\right\}dt.$$

לדוגמה, עבור האות הפשוט $s(t) = \sin(2\pi t/T)$, אנו יודעים לפרקו לשתי הרמוניות – מעניין לראות אם הגישה הרחבה יותר כאן תצליח להגיע לאותו פירוק. מקדמי טור הפורייה יהיו

$$c_n = \frac{1}{T}S_p\left(\frac{n}{T}\right) = \frac{1}{T}\int_0^T s(t)\exp\left\{-\frac{j2\pi nt}{T}\right\}dt =$$

$$= \frac{1}{T}\int_0^T \sin\left(\frac{2\pi t}{T}\right)\exp\left\{-\frac{j2\pi nt}{T}\right\}dt =$$

$$= \frac{1}{2jT}\left[\int_0^T \exp\left\{\frac{j2\pi t}{T}\right\}\exp\left\{-\frac{j2\pi nt}{T}\right\}dt - \int_0^T \exp\left\{-\frac{j2\pi t}{T}\right\}\exp\left\{-\frac{j2\pi nt}{T}\right\}dt\right] =$$

$$= \frac{1}{2jT}\left[\int_0^T \exp\left\{-\frac{j2\pi(n-1)t}{T}\right\}dt - \int_0^T \exp\left\{-\frac{j2\pi(n+1)t}{T}\right\}dt\right]$$

ברור שעבור כל n מלבד (+1) ו-(-1) ביטוי זה מתאפס זהותית, כיוון שהוא מבצע סכימה על אות מחזורי במחזור שלם בו הממוצע הא אפס. עבור n=1 יתקבל

$$c_1 = \frac{1}{2Tj}\left[\int_0^T dt - \int_0^T \exp\left\{-\frac{j4\pi t}{T}\right\}dt\right] = \frac{1}{2j}.$$

באופן דומה, עבור n=-1 יתקבל

$$c_{-1} = \frac{1}{2Tj}\left[\int_0^T \exp\left\{\frac{j4\pi t}{T}\right\}dt - \int_0^T dt\right] = \frac{-1}{2j}.$$

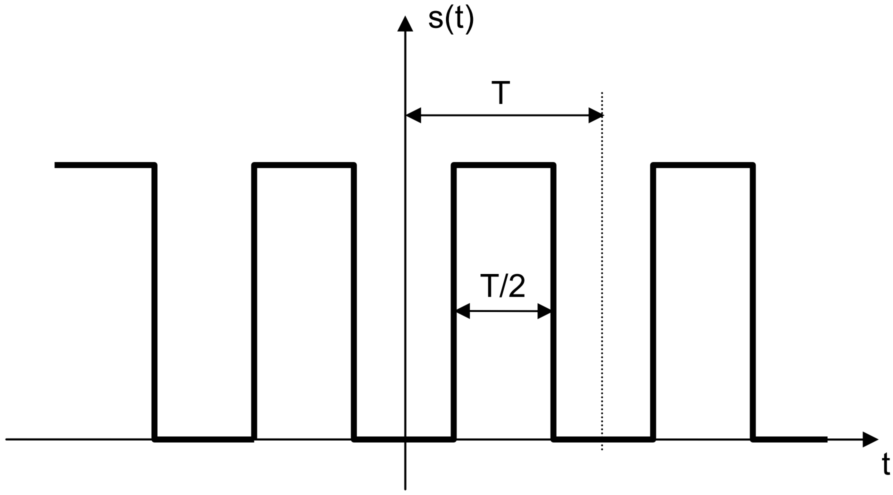
ולכן טור הפורייה של האות $s(t) = \sin(2\pi t/T)$ הוא

$$s(t) = \sum_{n=-\infty}^{\infty} c_n \cdot \exp\left\{\frac{j2\pi nt}{T}\right\} = \frac{1}{2j}\left(\exp\left\{\frac{j2\pi t}{T}\right\} - \exp\left\{-\frac{j2\pi t}{T}\right\}\right).$$

ואמנם זהו הפירוק שלו ציפינו.

כדוגמה שנייה, נתייחס לאות ריבועי מחזורי המתואר בציור הבא. ניתן יהיה להציג אות זה ע"י הטור

$$,s(t)=\sum_{n=-\infty}^{\infty} c_n \cdot \exp\left\{\frac{j2\pi nt}{T}\right\}$$



כמקדמי טור הפורייה יהיו

$$
\begin{aligned}
c_n &= \frac{1}{T}S_p\left(\frac{n}{T}\right)=\frac{1}{T}\int_0^T s(t)\exp\left\{-\frac{j2\pi nt}{T}\right\}dt= \\
&= \frac{1}{T}\int_{T/4}^{3T/4}\exp\left\{\frac{-j2\pi nt}{T}\right\}dt=\left[\frac{-1}{j2\pi n}\exp\left\{\frac{-j2\pi nt}{T}\right\}\right]_{T/4}^{3T/4}= \\
&= \frac{-1}{j2\pi n}\left(\exp\left\{\frac{-j3\pi n}{2}\right\}-\exp\left\{\frac{-j\pi n}{2}\right\}\right)= \\
&= \frac{\exp\{-j\pi n\}}{j2\pi n}\left(\exp\left\{\frac{j\pi n}{2}\right\}-\exp\left\{\frac{-j\pi n}{2}\right\}\right)=\frac{1}{2}\exp\{-j\pi n\}\operatorname{sinc}\left\{\frac{\pi n}{2}\right\}
\end{aligned}
$$

כלומר, האות הריבועי ניתן גם לתיאור כ:

$$.\,s(t)=\frac{1}{2}\sum_{n=-\infty}^{\infty}\exp\{-j\pi n\}\operatorname{sinc}\left\{\frac{\pi n}{2}\right\}\cdot\exp\left\{\frac{j2\pi nt}{T}\right\}$$

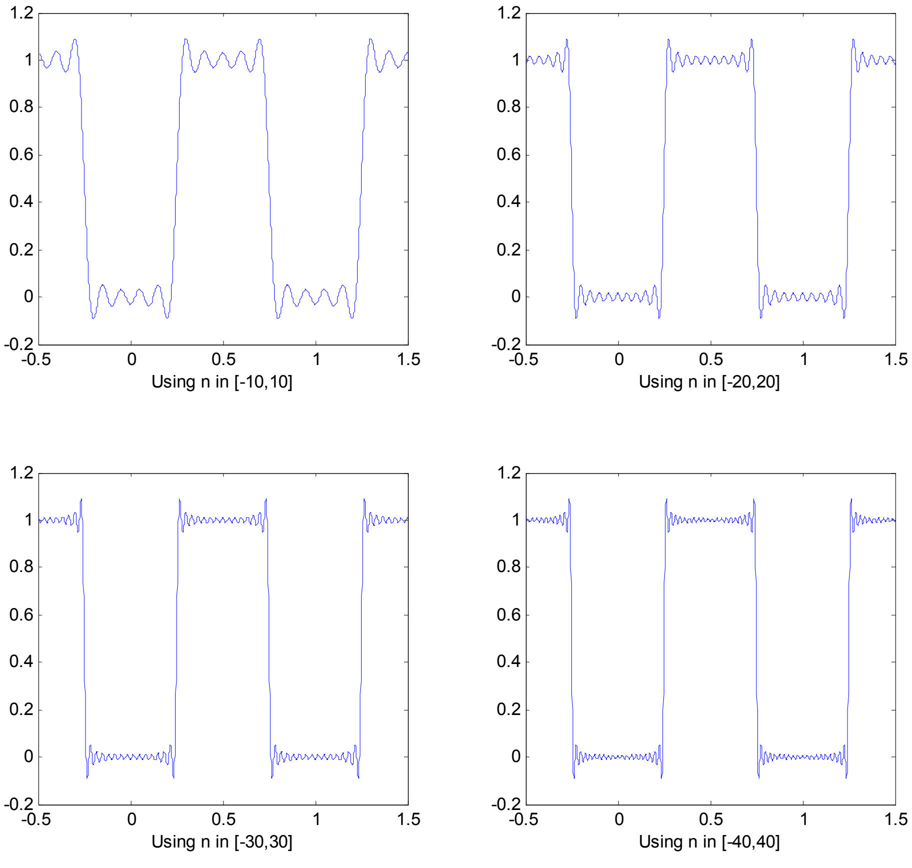
משמעות הדבר היא שבאות ריבועי יש תדרים גבוהים עד לאינסוף, אך משקלם בסכימה הכוללת דועכת כמו sinc (כלומר כמו 1/n – נשים לב כי ניתן לפשט הלאה את מקדם טור הפוריה כיוון של-sin מוזנת זווית שתמיד תהיה מכפלה שלמה של 90°). בעניין הפאזה בביטוי שהתקבל, ביטוי זה הוא

$$.\,\exp\{-j\pi n\}=\cos(-\pi n)+j\sin(-\pi n)=\cos(-\pi n)=(-1)^n$$

מעניין לציין כי מקדמי טור הפורייה סימטריים, $c_n=c_{-n}$, ולכן הגורם המדומה מתבטל ונקבל

$$s(t) = 2\sum_{n=-\infty}^{\infty} \exp\{j\pi n\} \operatorname{sinc}\left\{\frac{\pi n}{2}\right\} \cdot \exp\left\{\frac{j2\pi n}{T}\right\} = \sum_{n=-\infty}^{\infty} \frac{(-1)^n \operatorname{sinc}\left\{\frac{\pi n}{2}\right\}}{2} \cdot \cos\left\{\frac{2\pi n t}{T}\right\}.$$

הציור הבא מראה את הצטברות הטור עבור סכומים חלקיים (T=1). כפי שניתן לראות, הטור מתקרב והולך לעבר המדרגה המקורית. תופעה מעניינת ומקוללת של טורי (והתמרת) פורייה היא האפקט בסמוך למדרגה בפונקציה. תופעה זו הקרויה "תופעת גיבס" אומרת כי ליד אי-רציפות בפונקציה תהיה קפיצה של כ-9% בגובה אשר לא דועכת, גם כאשר הסכימה בטור גודלת והולכת. עם עליית מספרי האיברים בסכום התופעה נהיית בעלת תמך הולך וקטן, אך גובהה נותר. קיים ניתוח תיאורטי המסביר תופעה זו אך אנו לא נתעמק בו.



למה חשוב לנו כל כך עניין הטיפול באותות מחזוריים? האם הם באמת כה נפוצים? כלל וכלל לא! למרות זאת הטיפול שהוצג חשוב משלוש סיבות:

1. במקרים רבים נרצה לטפל באות בעל תמך סופי (כלומר כזה שקיים רק במקטע זמן סופי). במקרים אלה מקובל להניח כי אות זה "הומשך מחזורית" כדי להתקיים בכל נקודת זמן, וכדי לנצל את המבנה הנוח הנ"ל של טורי פורייה.

2. ישנה דואליות מעניינת בין ציר הזמן וציר התדר, ואנו ראינו מספר תופעות שנובעות מדואליות זו. נשים לב שההגדרה להתמרה קדימה ואחורה מאוד דומות, אם ההתמרה של פונקצית הלם היא קבוע, אז התמרה של קבוע היא הלם, ועוד. כמו שהתמרה של אות מחזורי היא דגומה, אנו נראה כי עבור אות דגום ההתמרה מחזורית, ולכך חשיבות עצומה בדיון באותות ע"י מחשב.
3. מסתבר כי כשפורייה פיתח את תורתו, הוא החל באותות בעלי תמך סופי עם המשכה מחזורית, ולכן לטורי פורייה חשיבות היסטורית.

2. משפט הדגימה ומשמעותו

נתון לנו אות $s(t)$, ורצוננו לדגום אותו במרווחים T שווים על מנת להמירו לאות דיסקרטי היכול להישמר בזיכרון מחשב (ברור כי יש לדאוג שהתמך – תחום הקיום – של האות יהיה סופי וכל ערך דגימה בו ייוצג בכמות סיביות סופית על מנת שנקבל באמת קובץ בגודל סופי של סיביות). פעולת הדגימה שנציע היא

$$s[n] = s(t)\Big|_{t=nT} = s(nT)$$

השאלה המרכזית בה נעסוק במסגרת זו היא אילו תנאים יש להציב על המרווח T כך שניתן יהיה לשחזר את $s(t)$ מתוך דגימותיו $s[n]$ ללא אובדן? לצורך מענה לשאלה זו, נעשה שימוש בתיאור מתמטי נוח יותר לפעולת הדגימה – דגימה ע"י מכפלה באותה פונקצית "רכבת הלמים" שכבר פגשו. נקבע כי האות נדגם ע"י המכפלה

$$\tilde{s}(t) = s(t) \cdot \mathrm{III}_T(t) = s(t) \cdot \sum_n \delta(t - nT)$$

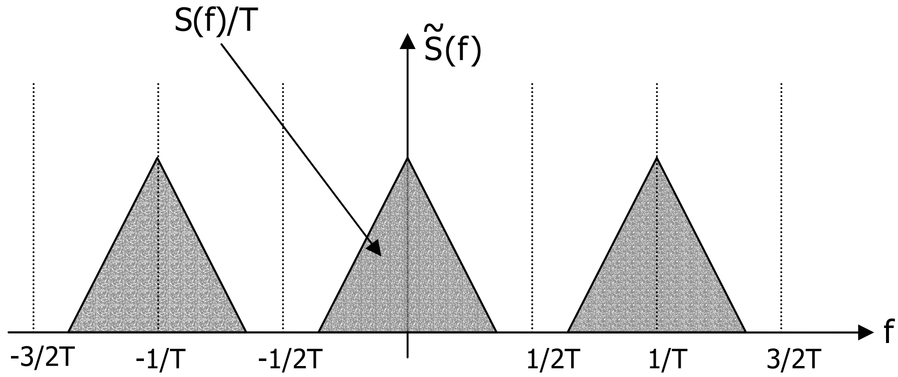
היתרון שבייצוג זה הוא שהאות הדגום יכול להיות מטופל בכלי שבנינו עבור הרצף, כי הוא אות רציף לכל דבר.

מעניין לציין כי אות זה נראה כמו פונקציות הלם הפזורות במרווחים אחידים על הציר הממשי עם גבהים משתנים. מבנה זה מוכר לנו מהדיון בטורי פורייה כהתמרת הפורייה של אות מחזורי. לכן, בשל הדואליות הקיימת בהתמרת הפורייה, נצפה שהתמרת הפורייה של אותות דגומים תהיה מחזורית, ואנו נראה שאמנם זה כך.

כיוון שהאות בנוי ממכפלה של שני אותות רציפים בעקרון, התמרת הפורייה שלו היא קונבולוציה של ההתמרות. לכן

$$\tilde{S}(f) = F\{\tilde{s}(t)\} = F\left\{s(t) \cdot \sum_n \delta(t - nT)\right\} = F\{s(t)\} \otimes F\left\{\sum_n \delta(t - nT)\right\} =$$
$$= \frac{1}{T} S(f) \otimes \sum_n \delta\left(f - \frac{n}{T}\right) = \frac{1}{T} \sum_n S\left(f - \frac{n}{T}\right)$$

השתמשנו בעובדה שהתמרה על רכבת ההלמים נתונה כרכבת הלמים עם מרווחים 1/T ועם נקדם מתאים. מתקבל כי הפונקציה S(f) משוכפלת שוב ושוב במרווחים קצובים 1/T ובהכפלה בקבוע 1/T. הציור הבא ממחיש תופעה זו.



ואמנם, כפי שציפינו, התמרת הפורייה של האות הדגום היא מחזורית. מציור זה גם ניכר כי כאשר S(f) מתאפס מחוץ לתחום [-1/2T,1/2T], מתקבל כי השכפולים לא רומסים זה על זה. נניח כי אמנם S(f) מקיים

$$S(f)=\begin{cases} S(f) & |f|\le f_{max} \\ 0 & \text{otherwise} \end{cases}$$

תכונה זו קרויה "אות חסום פס". אזי, על מנת שהשכפולים לא יעלו זה על זה נדרוש

$$f_{max}<\frac{1}{2T} \;\Rightarrow\; T<\frac{1}{2f_{max}}$$

הדרישה הזו קובעת שעלינו לדגום מספיק צפוף כדי לכלול כמות מספקת של דגימות לייצוג מלא של האות.

מדוע חשוב לנו למנוע מהשכפולים לעלות זה על זה? ובכן, בהינתן האות הדגום $\tilde{s}(t)$ (שכל דגימה בו היא פונקצית הלם), זהו אות רציף הניתן להעברה דרך מערכת. ע"י הפעלת מערכת שהתמרת הפורייה של תגובתה להלם הוא מסנן ריבועי מעביר נמוכים (LPF) מהצורה

$$H(f)=\begin{cases} T & |f|\le \dfrac{1}{2T} \\ 0 & \text{otherwise} \end{cases}$$

על אות זה נקבל בדיוק את אות המקור, כיוון שפעולת המערכת הינה מכפלה בתדר, וע"י ההכפלה במסנן הנ"ל אנו מבטלים את כל השכפולים למעט זו הראשית, וכיוון שהיא בדיוק ההתמרה של אות המקור, הרי שביצענו בשלמות פעולה הפוכה לפעולת הדגימה וקיבלנו שחזור.

מבחינה מעשית, לא נתון לנו האות $\tilde{s}(t)$, אלא אות דגום s[n]. לכן השאלה המתבקשת היא כיצד בפועל מבצעים את הסינון המשחזר? נענה לשאלה זו ע"י כיתוב מפורש של פעולת השחזור

$$\begin{aligned}\hat{s}(t) &= F^{-1}\{H(f)F\{\tilde{s}(t)\}\} = \tilde{s}(t)\otimes h(t) = \\ &= \int_x \sum_n s[n]\delta(x-nT)h(t-x)dx = \sum_n s[n]h(t-nT)\end{aligned}$$

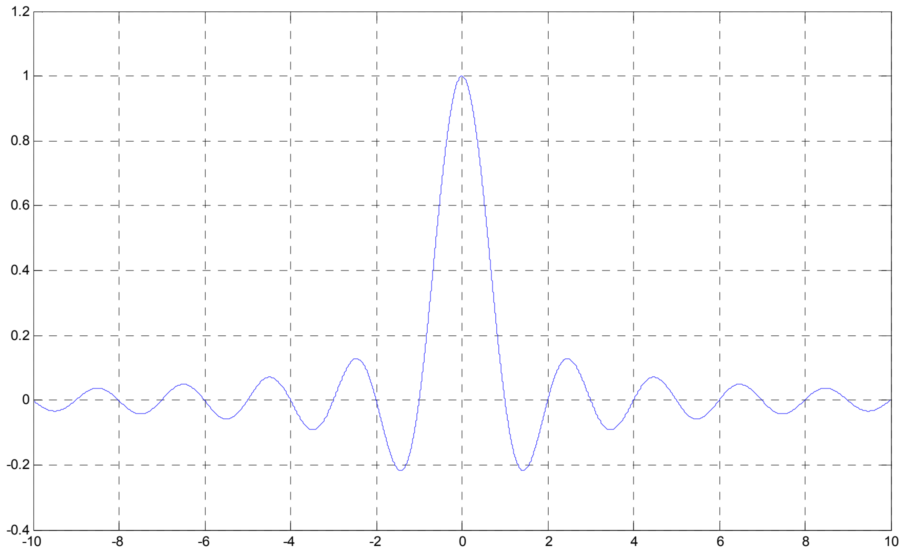
ההתמרה ההפוכה (מהתדר לזמן) של המסנן H(f) תהיה (ראה פרק קודם)

$$\mathrm{sinc}\{\pi t/T\} = \frac{\sin\{\pi t/T\}}{\pi t/T}$$

ולכן השחזור ייעשה ע"י הנוסחה

$$\hat{s}(t) = \sum_n s[n]h(t-nT) = \sum_n s[n]\,\mathrm{sinc}\left(\frac{\pi(t-nT)}{T}\right).$$

מה משמעות נוסחה זו? ננסה להבין ע"י הצגת פונקצית ה-sinc תחילה. פונקציה זו מתוארת בציור הבא עבור T=1.



ציור זה נבנה ע"י פקודות ה-Matlab :

```
T=1;t=-10.0005:0.001:10;ss=sin(pi*t/T)./(pi*t/T);
plot(t,ss); axis([-10 10 -0.4 1.2]); grid
```

נשים לב כי פונקציה זו דועכת באופן מתנדנד כשמתרחקים מהראשית, והיא מתאפסת זהותית בנקודות $t = nT$ חוץ מאשר עבור n=0, שם ערכה הוא 1. משמעות הדבר היא שכאשר נרצה להשתמש בפונקציה זו לשחזור של ערך הפונקציה בנקודת זמן שהינה מכפלה שלמה של מרווח הדגימה, $t = k_0T$, נקבל

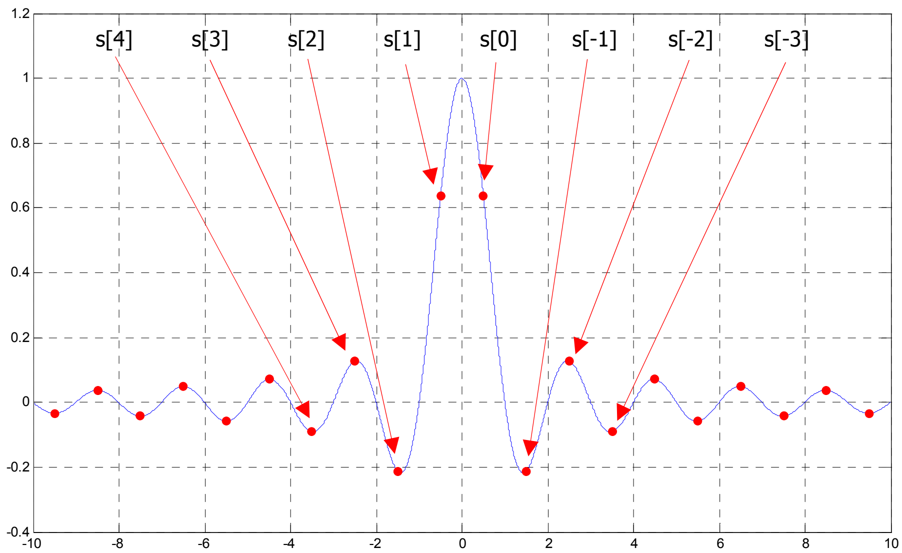
$$\hat{s}(k_0T) = \sum_n s[n]\operatorname{sinc}\left(\pi(k_0 - n)T/T\right) = s(k_0T).$$

תוצאה זו רצויה וצפויה. ערכי הדגימה ידועים במדויק ולכן טבעי שיתקבלו גם כערכי הפונקציה הרציפה בנקודות זמן אלה.

נבחן כעת כיצד מתנהגת משוואת השחזור כאשר נרצה לשחזר את ערך הפונקציה בנקודת הזמן $t = T/2$. נקבל

$$\begin{aligned}\hat{s}(T/2) &= \sum_n s[n]\operatorname{sinc}\left(\pi\frac{(T/2 - nT)}{T}\right) = \\ &= \cdots + s[-1]\operatorname{sinc}(3\pi/2) + s[0]\operatorname{sinc}(\pi/2) + s[1]\operatorname{sinc}(-3\pi/2) + \cdots\end{aligned}.$$

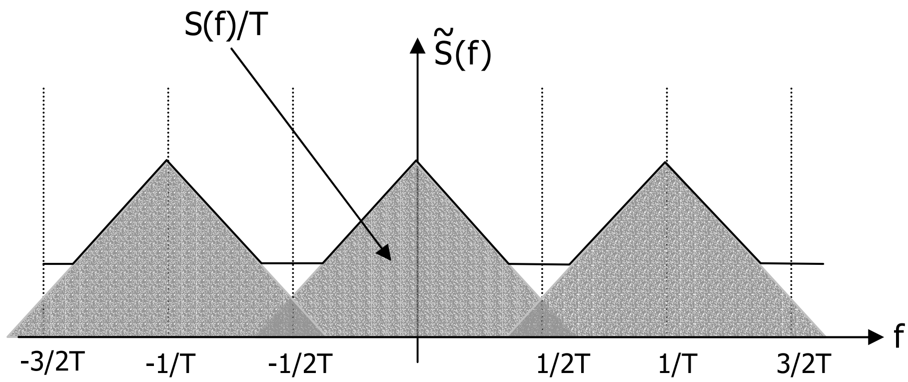
הציור הבא ממחיש את משמעות המשוואה (עבור T=1).



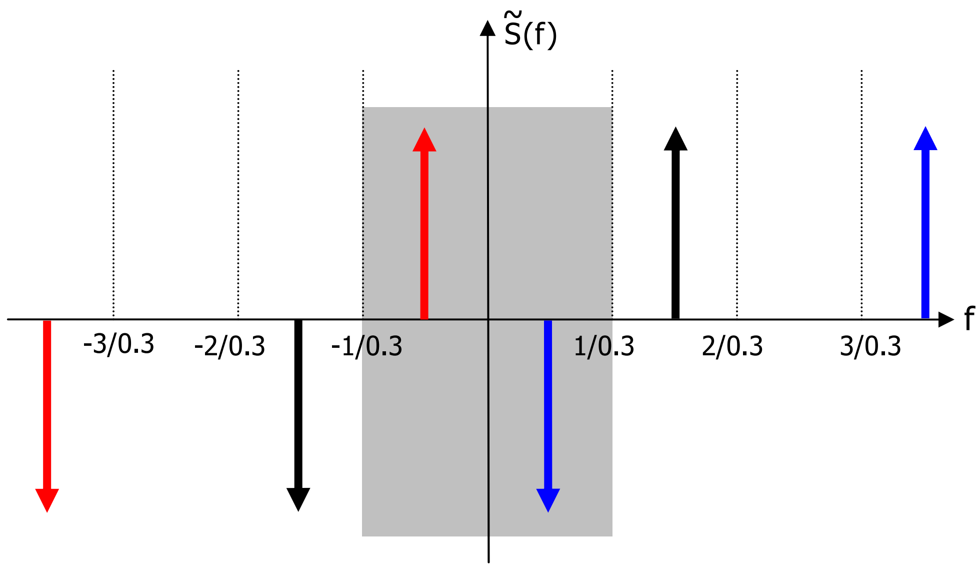
אנו רואים כי הדומיננטיים בחישוב הערך הם באופן טבעי s[0] ו-s[1], אך לשם חישוב מלא יש להשתמש בכל ערכי הדגימות מ-$-\infty$ ועד ∞, כשלכל דגימה יש השפעה פרופורציונית הפוכה למרחקה מנקודת היעד.

דרשנו כי תדר הדגימה (אם T הוא מרווח הדגימה, תדרה הוא 1/T כדי לומר כי יש 1/T דגימות בשנייה) יהיה גדול מפעמיים התדר המרבי באות, וזה יבטיח יכולת תיאורטית של שחזור. מה קורה אם מרווח הדגימה T אינו מקיים דרישה זו? במקרה זה, הרפליקות המופיעות באופן מחזורי תדרוכנה זו על זו ותסתכמנה, כמתואר בציור הבא. מציור זה ברור כי לא ניתן לשחזר את האות המקורי – זנבות המשולשים נפגשים ומסתכמים ולא נוכל לדעת את ערך הזנב האמיתי בתדרים אלו. רואים גם כי לא רק

שנגרם נזק למידע שפולש מתחום התדרים [-1/2T,1/2T] אלא נזק גם נגרם לתדרים סמוכים להם בתוך האינטרוול, בשל אותה חפיפה. תופעה זו מוכרת בשם תופעת הקיפול (Aliasing).



מה משמעות תופעת הקיפול (Aliasing)? השם aliasing אינו קיפול כי אם התחזות, ואמנם ישנה כאן תופעה של התחזות. נראה זאת בדרך של דוגמה. נניח כי בידינו האות $s(t) = \sin(2\pi \cdot 5t)$. התדר המרבי בו הוא כמובן 5 ולכן מרווח הדגימה צריך להיות 0.1 שניות ומטה. נניח כי "פשענו" כנגד אות זה ודגמנו אותו במרווחים של 0.15 שניות (גדול מהמותר). הייצוג בתדר ייראה כך

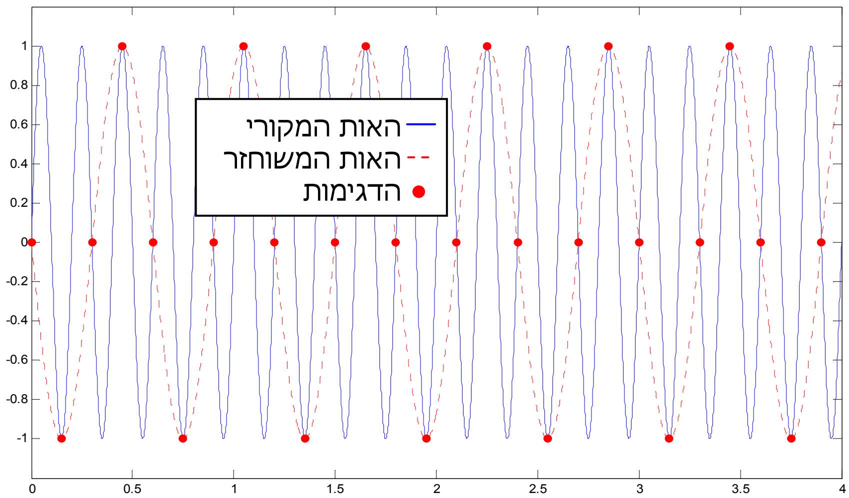


כל פונקצית הלם בציור הנ"ל בעלת גובה 1/2j. הרפליקה המקורית מתוארת בשחור, זו שמימינה בכחול, וזאת שמשמאלה באדום. בבואנו לשחזר את האות, אם נבחר את התוכן בתחום [-1/2T,1/2T] ונאפס את היתר כפי ששחזור צריך להיעשות, נקבל שתי

פונקציות הלם בתדרים $\pm 5/3$, ובהתמרה לזמן יהיה זה האות הסינוסי $s(t)=\sin(2\pi \cdot 5t/3)$ - קיבלנו אות מתחזה עם תדר שלא היה במקור. תיאור האות המקורי הרציף, הדגום והמשוחזר מתוארים בציור הבא. למעשה, התדר שהתקבל הוא תמונת המראה של התדר המקורי, כשציר הסימטריה הוא תדר הסף הנובע מהדגימה. תדר המקור הוא 5, תדר הסף הוא 3.333, והתדר שהתקבל הוא $3.333-(5-3.333)=1.667$.

קוד ה-Matlab שנתן גרף זה הוא

```
clf; t=0:0.001:4; s=sin(2*pi*5*t); plot(t,s);
axis([0 4 -1.2 1.2]); hold on; t1=0:0.15:4; s1=sin(2*pi*5*t1);
h=plot(t1,s1,'ro'); set(h,'MarkerFaceColor','r');
s2=-sin(2*pi*5*t/3); plot(t,s2,'r:');
```



נסכם, אם כן, את המסר המרכזי מהדיון הנ"ל כמשפט:

משפט 2: (דגימה ושחזור) בהינתן פונקציה $s(t)$ חסומת פס, כלומר בעלת התמרת פורייה המקיימת

$$S(f)=\begin{cases} S(f) & |f|\le f_{max} \\ 0 & \text{otherwise} \end{cases}$$

דגימת הפונקציה במרווחים קצובים T, $s[n]=s(nT)$, המקיימים

$$T<\frac{1}{2f_{max}}$$

מבטיחה יכולת שחזור הפונקציה בשלמות. נוסחת השחזור הינה

$$\hat{s}(t)=\sum_n s[n]\,\mathrm{sinc}\left(\frac{\pi(t-nT)}{T}\right).$$

3. עיבוד אותות דגומים

דגמנו את האות $s(t)$ והתקבל $s[n]$. לרוב אף לא נצטרך לטרוח ונקבל את האות כשהוא כבר דגום. נוכל להתחיל לחשוב על עיבוד דיגיטאלי של אותות, כשכלי היסוד הם התמרת פורייה של האות הדגום, וסינונו ע"י קונבולוציה. נתחיל את הדיון בתורת המערכות לאותות כאלה.

ממש כשם שהפעלנו מערכות על אותות הנראות כפונקציה ברצף, נרצה לעשות זאת על אותות בזמן בדיד (דיסקרטיים – כאלה המתוארים כסדרת מספרים). יהיה עלינו להראות כיצד מערכת מתנהגת, וכיצד סינון מומר לקונבולוציה. למעשה, אנו אמורים לחזור על כל הסיפור של מהי מערכת הפועלת על אותות כאלה, מהי ליניאריות, מהי קביעות בזמן, ועוד. נעשה זאת בתמציתיות משום שכל רעיונות אלו חוזרים על עצמם, וניגש מיד למערכות ליניאריות וקבועות בזמן.

מערכת מוגדרת כקבועה בזמן אם לכל אות כניסה מתקבל כי השהיית הכניסה יוצרת השהייה דומה ביציאה:

$$\forall s[n], d: \quad g[n] = T\{s[n]\} \Rightarrow g[n-d] = T\{s[n-d]\}.$$

מערכת מוגדרת כליניארית אם היא מקיימת את התכונה הבאה - לכל שני אותות $s_1[n]$ ו-$s_2[n]$ ולכל שני סקלרים a ו-b, יתקיים כי

$$T\{a \cdot s_1[n] + b \cdot s_2[n]\} = a \cdot T\{s_1[n]\} + b \cdot T\{s_2[n]\}.$$

והציור שהראינו קודם בהקשר לכך ביחס למערכות בזמן רציף תקף באותה מידה. ניתן מספר דוגמאות להמחיש זאת:

- המערכת $g[n] = s[n] + 3s[n-1]$ היא מערכת ליניארית וקבועה בזמן. למעשה זו מערכת סיבתית, ועם זיכרון.
- המערכת $g[n] = s[n] + \sqrt{n} \cdot s[n-1]$ תלויה בזמן, אך היא עדיין ליניארית, כי היא בונה את מוצאה כצירוף ליניארי של אות הכניסה.
- המערכת $g[n] = 2s[n] + 3$ אינה ליניארית, אך היא קבועה בזמן. קל לראות זאת לפי ההגדרה.
- המערכת $g(t) = \sum_{j=0}^{6} w_j \cdot s[n-j]$ היא מערכת ליניארית וקבועה בזמן. זוהי מערכת שמבצעת ממוצע משוקלל של 7 הדגימות האחרונות באות.
- המערכת $g[n] = s[n] + 3s[n-0.5]$ לא מוגדרת! ברגע שעברנו לאות דגום, אין זמן שברי.
- המערכת $g[n] = s[n]s[n-1]$ אינה ליניארית אך היא קבועה בזמן.

ממש כשם שעשינו זאת לרצף, יש להגדיר פונקצית הלם דרכה נוכל לנתח התנהגות מערכות ליניאריות וקבועות בזמן. פונקצית ההלם הדיסקרטית היא

$$.\ \delta_D[n] = \begin{cases} 1 & n=0 \\ 0 & \text{אחרת} \end{cases}$$

בדומה להתנהגות פונקצית ההלם ברצף, קיימת התכונה הבאה: עבור אות $s[n]$ כלשהו ופונקצית ההלם הנ"ל, יתקבל

$$\sum_{k=-\infty}^{\infty} \delta_D[k] s[n-k] = \sum_{k=-\infty}^{\infty} \delta_D[n-k] s[k] = s[n]\ ,$$

וקל לראות נכונות נוסחה זו, כי פונקצית ההלם קיימת רק באחד האינדקסים ומתאפסת לכל האחרים. משמעות משוואה זו שהאות $s[n]$ יכול להיות מתואר כצירוף ליניארי של אותות הלם מוזזות $\delta[n-k]$ כשהן משוקללות ע"י ערכי האות עצמו $s[k]$.

אם נזין למערכת זו הלם $\delta_D[n]$, נקבל מוצא אשר ייקרא התגובה להלם ויסומן $h[n]$, כלומר $T\{\delta_D[n]\} = h[n]$. נניח כעת כי הזנו אות כלשהו אחר $s[n]$ למערכת T. לכן, כיוון שהאות יכול להיות מתואר כסכום הלמים נקבל

$$g[n] = T\{s[n]\} = T\left\{ \sum_{k=-\infty}^{\infty} \delta_D[n-k] s[k] \right\} =$$

$$= \sum_{k=-\infty}^{\infty} T\{\delta_D[n-k]\} s[k] = \sum_{k=-\infty}^{\infty} h[n-k] s[k]$$

כאן השתמשנו בליניאריות של המערכת כשהחלפנו סדר פעולת המערכת עם הסכימה. השתמשנו בקביעות בזמן כשהנחנו שפעולת המערכת על הלם מוזז היא תגובת הלם מוזזת.

קיבלנו כי כל מערכת ליניארית ניתנת לאפיון ע"י תגובתה להלם, ובהינתן אות כניסה כלשהו, תגובתה ההלם מאפשרת לנו חישוב מוצאה. המשוואה שקיבלנו אינה אלא פעולת הקונבולוציה הדיסקרטית, אותה נסמן כ:

$$.\ g[n] = T\{s[n]\} = \sum_{k=-\infty}^{\infty} h[n-k] s[k] = \sum_{k=-\infty}^{\infty} h[k] s[n-k] = h[n] \otimes s[n]$$

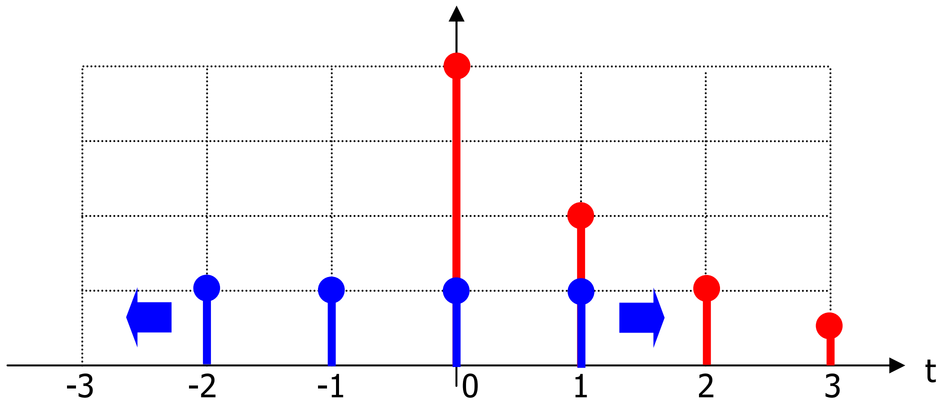
חישוב קונבולוציה זו דומה לחישוב זו הרציפה, אלא שסכימה דיסקרטית מחליפה את האינטגרל. נמחיש ביצוע חישוב זה דרך דוגמה:

$$.\ g[n] = s[n-2] + s[n-1] + s[n] \quad \Rightarrow \quad h[n] = \delta_D[n-2] + \delta_D[n-1] + \delta_D[n]$$

נניח כי אות הכניסה שלנו הוא

$$. s[n] = \begin{cases} 4 \cdot 2^{-n} & n \geq 0 \\ 0 & \text{אחרת} \end{cases}$$

הציור הבא ממחיש כיצד חישוב זה יתבצע – אנו מקפיאים את ציור האות s[n] (באדום), יוצרים תמונת מראה של h[n], ומסיעים אותו לאורך ציר הזמן תוך כדי חישוב סכום המכפלות בחפיפה שבין שני האותות.



עבור n=-1 ומטה נקבל כי המוצא הוא אפס כי אין חפיפה בין השניים. נשים לב לכך שכיוון שהמערכת סיבתית (התגובה להלם קיימת רק לזמנים מאפס ומעלה), וכיוון שהאות סיבתי, ברור כי גם המוצא יהיה סיבתי.

עבור n=0 נקבל חפיפה רק בנקודת זמן אחת והמוצא יהיה 4. עבור n=1 נקבל כי המוצא הוא $4+2=6$. באופן דומה, עבור n=2 המוצא הוא 4+2+1=7, עבור n=3 המוצא הוא 4+2+1+0.5=7.5. מנקודת זמן זו נקבל כי החפיפה מלאה תמיד, ועם כל תנועה בנקודת זמן אחת המוצא הוא חצי מקודמו. לכן התוצאה היא

$$g[n] = \begin{cases} 0 & n \leq -1 \\ 4 & n = 0 \\ 6 & n = 1 \\ 7 & n = 2 \\ 7.5 \cdot 2^{n-3} & n \geq 3 \end{cases}$$

משפט 3: כל תכונות הקונבולוציה שתוארו למקרה הרציף תקפות גם להגדרה הדיסקרטית:

$$\sum_{k=-\infty}^{\infty} g[n-k]s[k] = \sum_{k=-\infty}^{\infty} g[k]s[n-k] = g[n] \otimes s[n]$$

- קומוטטיביות - $. s[n] \otimes g[n] = g[n] \otimes s[n]$
- אסוציטיביות - $. s[n] \otimes (g[n] \otimes f[n]) = (s[n] \otimes g[n]) \otimes f[n]$

- דיסטריביוטיות - $s[n]\otimes[g[n]+f[n]]=s[n]\otimes g[n]+s[n]\otimes f[n]$.
- מכפלה בסקלר - $a\cdot[s[n]\otimes g[n]]=[as[n]]\otimes g[n]=[ag[n]]\otimes s[n]$
- הזזה - $s[n]\otimes g[n-k]=s[n-k]\otimes g[n]$.

כשדנו במקרה הרציף ציינו שתי תכונות של מערכות ליניאריות וקבועות בזמן – אלה תקפות גם כאן:

משפט 4: מערכות ליניאריות וקבועות בזמן מקיימות:

1. הן משתחלפות (נובע ישירות מהתכונות של הקונבולוציה).
2. מערכת סיבתית אם תגובתה להלם סיבתית –

$$g[n]=T\{s[n]\}=\sum_{k=-\infty}^{\infty}h[k]s[n-k] \underset{\forall n<0,h[n]=0}{\underset{\uparrow}{=}} \sum_{k=0}^{\infty}h[k]s[n-k]$$

4. עיבוד אותות דגומים – התמרת פורייה

נדבר כעת על התמרת פורייה לאות בזמן בדיד – כיצד זו תתבצע? נוכל פשוט להשתמש בהגדרת ההתמרה המקורית כשזו מופעלת על האות הדגום (המוכפל ברכבת הלמים). זה ייתן

$$\begin{aligned} F\{\tilde{s}(t)\} &= \int_{-\infty}^{\infty} s(t)\cdot\sum_n \delta(t-nT)\exp(-j2\pi ft)dt = \\ &= \sum_n \int_{-\infty}^{\infty} s(t)\delta(t-nT)\exp(-j2\pi ft)dt = \\ &= \sum_n s(nT)\exp(-j2\pi nTf)=\sum_n s[n]\exp(-j2\pi n\theta) \end{aligned}$$

כשהשתמשנו כאן בהגדרה $\theta=Tf$. הבחירה בהתמרה זו חשובה על מנת להוציא מההגדרה את הצורך בידיעת מרווח הדגימה – כשנתון לנו $s[n]$ אין לנו עניין יותר בשאלת מקורו הרציף.

דרך אחרת להגיע לאותה הגדרה היא הבאה: בדומה לנעשה עבור הרצף, נגדיר את האותות העצמיים שחולפים דרך המערכת ללא שינוי. "ננחש" שאותות אלו הם $s[n]=\exp\{j2\pi n\theta\}$. הזנת אות כזה למערכת ליניארית וקבועה בזמן המאופיינת ע"י תגובתה להלם $h[n]$ תיתן

$$g[n]=\sum_{k=-\infty}^{\infty}h[k]s[n-k]=\exp\{j2\pi n\theta\}\sum_{k=-\infty}^{\infty}h[k]\exp\{-j2\pi k\theta\}=H(\theta)\cdot\exp\{j2\pi n\theta\}$$

וכך קיבלנו הגדרה של התמרת פורייה לאות בזמן בדיד, ע"י

$$S(\theta)=F\{s[n]\}=\sum_{n}s[n]\cdot\exp\{-j2\pi n\theta\}.$$

נשים לב כי $S(\theta)=S(\theta+k)$, $\forall k\in Z$, כלומר התמרת הפורייה הזו מחזורית, ואורך מחזורה הוא 1. לכן נתעניין בתחום $-0.5\le\theta\le 0.5$ בלבד.

כדוגמה, עבור האות הדיסקרטי $s[n]=\exp\{-\alpha|n|\}$ עבור $\alpha>0$ התמרת הפורייה היא

$$\begin{aligned}S(\theta)=F\{s[n]\}&=\sum_{n=-\infty}^{\infty}s[n]\cdot\exp\{-j2\pi n\theta\}=\sum_{n=-\infty}^{\infty}\exp\{-j2\pi n\theta-\alpha|n|\}=\\&=\sum_{n=0}^{\infty}\exp\{-j2\pi n\theta-\alpha|n|\}+\sum_{n=-\infty}^{-1}\exp\{-j2\pi n\theta-\alpha|n|\}=\\&=\sum_{n=0}^{\infty}\exp\{-(j2\pi\theta+\alpha)n\}+\sum_{n=1}^{\infty}\exp\{(j2\pi\theta-\alpha)n\}\end{aligned}$$

נשתמש בנוסחה של טור גיאומטרי

$$\forall|x|<1,\quad \sum_{n=0}^{\infty}x^{n}=\frac{1}{1-x}$$

קל להוכיח קשר זה כיוון שע"י הכפלה ב- (1-x) נקבל זהות:

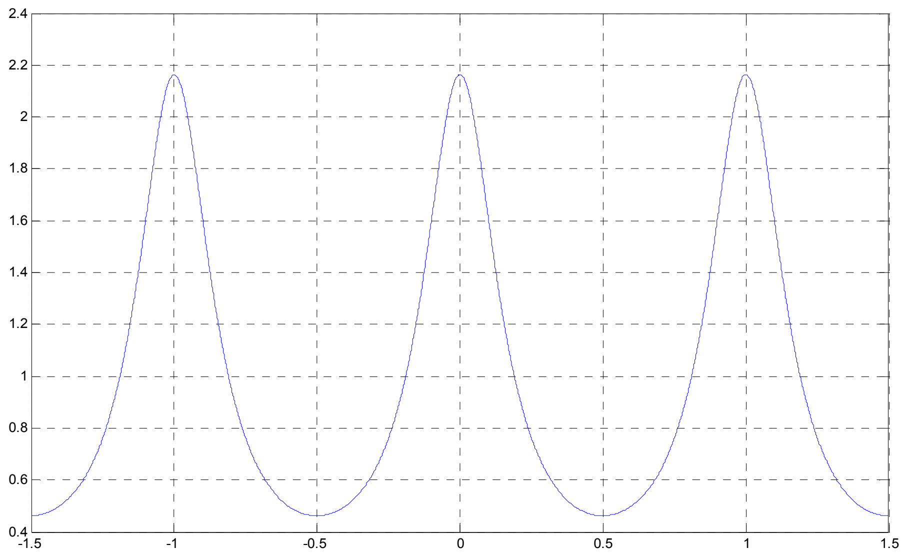
$$(1-x)\sum_{n=0}^{\infty}x^{n}=\sum_{n=0}^{\infty}x^{n}-\sum_{n=1}^{\infty}x^{n}=1.$$

לכן,

$$\begin{aligned}S(\theta)&=\sum_{n=0}^{\infty}\exp\{-(j2\pi\theta+\alpha)n\}+\sum_{n=1}^{\infty}\exp\{(j2\pi\theta-\alpha)n\}=\\&=\frac{1}{1-\exp\{-j2\pi\theta-\alpha\}}+\frac{1}{1-\exp\{j2\pi\theta-\alpha\}}-1=\\&=\frac{1-\exp\{j2\pi\theta-\alpha\}+1-\exp\{-j2\pi\theta-\alpha\}}{1-\exp\{-j2\pi\theta-\alpha\}-\exp\{j2\pi\theta-\alpha\}+\exp\{-2\alpha\}}-1=\\&=\frac{2-2\exp\{-\alpha\}\cos(2\pi\theta)}{1-2\exp\{-\alpha\}\cos(2\pi\theta)+\exp\{-2\alpha\}}-1=\\&=\frac{1-\exp\{-2\alpha\}}{1-2\exp\{-\alpha\}\cos(2\pi\theta)+\exp\{-2\alpha\}}\end{aligned}$$

זהו הגרף של פונקציה זו, כשהוא מתואר בכוונה בתחום [-1.5,1.5] על מנת להראות 3 מחזורים שלמים שלו. נשים לב כי הפונקציה ממשית בשל היות האות המקורי סימטרי.

אנו לא נטרח לחזור על תכונות התמרת הפורייה הדיסקרטית כיוון שהם מאוד דומים לתכונות התמרת הפורייה המקורית.



גרף זה התקבל ע"י פקודות ה-Matlab הבאות:

```
a=1; theta=-1.5:0.001:1.5;
F=(1-exp(-2))./(1-2*exp(-1)*cos(2*pi*theta)+exp(-2));
plot(theta,F);
```

מה באשר להתמרה ההפוכה? בהינתן $S(\theta)$, נרצה לחזור לאות המקור $s[n]$. זה ייעשה ע"י

$$s[n]=\int_{-1/2}^{1/2} S(\theta)\exp\{j2\pi\theta n\}d\theta .$$

פיתוח זה מתאפשר ע"י הקשר אל הרצף. אנו נוכיח את נכונות פעולה זו ע"י הצבת ההתמרה קדימה בתוך זו לקבלת אותו אות כניסה. מתקבל:

$$\begin{aligned} s[n] &= \int_{-1/2}^{1/2} S(\theta)\exp\{j2\pi\theta n\}d\theta = \int_{-1/2}^{1/2}\left(\sum_{k=-\infty}^{\infty} s[k]\exp\{-j2\pi\theta k\}\right)\exp\{j2\pi\theta n\}d\theta = \\ &= \sum_{k=-\infty}^{\infty} s[k]\int_{-1/2}^{1/2}\exp\{j2\pi\theta(n-k)\}d\theta = \\ &= \sum_{k=-\infty}^{\infty} s[k]\left[\frac{\exp\{j\pi(n-k)\}-\exp\{-j\pi(n-k)\}}{j2\pi(n-k)}\right] = \\ &= \sum_{k=-\infty}^{\infty} s[k]\,\mathrm{sinc}\big(\pi(n-k)\big) = \sum_{k=-\infty}^{\infty} s[k]\delta_D[n-k] = s[n] \end{aligned}$$

השתמשנו כאן בתכונה שפונקצית ה-sinc מקבלת ערכי אפס בכל נקודה למעט בראשית, שם היא 1.

להתמרה שהוגדרה יש את מרבית התכונות שהצגנו להתמרה הרציפה. חשוב להבהיר כי אין זה במקרה – ההתמרה לאות הדגום אינה אלא מקרה פרטי של ההתמרה הכללית הרציפה. במסגרת זו נציג מספר תכונות חשובות:

משפט 5: התמר פורייה לאותות בזמן בדיד:

- הגדרה - עבור האות $g[n]$ התמרת הפורייה נתונה ע"י

$$G(\theta) = F\{g[n]\} = \sum_{n=-\infty}^{\infty} g[n]\exp\{-j2\pi n\theta\},$$

וההתמרה חזרה ע"י $g[n] = F^{-1}\{G(\theta)\} = \int_{-0.5}^{0.5} G(\theta)\exp\{j2\pi n\theta\}d\theta$.

- ליניאריות – התמרת הפורייה ליניארית, כלומר, עבור כל שני אותות $s[n]$ ו-$g[n]$, ושני סקלרים a ו-b, מתקיים כי

$$F\{a\cdot s[n] + b\cdot g[n]\} = a\cdot F\{s[n]\} + b\cdot F\{g[n]\}.$$

- ממוצע – ממוצע האות $g[n]$ אינו אלא $G(0)$:

$$G(0) = F\{g[n]\}\big|_{\theta=0} = \sum_{n=-\infty}^{\infty} g[n].$$

- ממשיות וסימטריות – בהינתן לנו אות $g[n]$ ממשי טהור (כלומר ללא מרכיב מדומה), להתמרת הפורייה שלו מבנה סימטרי צמוד,

$$G(\theta) = \sum_{n=-\infty}^{\infty} g[n]\exp\{-j2\pi n\theta\}$$

$$\overline{G(\theta)} = \overline{\sum_{n=-\infty}^{\infty} g[n]\exp\{-j2\pi n\theta\}} = \sum_{n=-\infty}^{\infty} g[n]\exp\{j2\pi n\theta\} = G(-\theta)$$

אם האות הנתון לנו $g[n]$ גם ממשי וגם סימטרי יתקבל כי $G(\theta)$ ממשי וסימטרי אף הוא.

- שימור מכפלה פנימית – עבור שני אותות $g[n]$ ו-$s[n]$, והתמרתם $G(\theta)$ ו-$S(\theta)$. מכפלה פנימית בין שתי פונקציות בזמן מוגדרת ע"י

$$\langle g[n], s[n]\rangle_n = \sum_{n=-\infty}^{\infty} g[n]\cdot\overline{s[n]}.$$

מכפלה פנימית בתדר של פונקציות ההתמרה תוביל לאותה תוצאה:

$$\langle G(\theta),S(\theta)\rangle_\theta = \int_{-0.5}^{0.5} G(\theta)\overline{S(\theta)}d\theta =$$

$$= \int_{-0.5}^{0.5}\left(\sum_{n=-\infty}^{\infty} g[n]\exp\{-j2\pi n\theta\}\right)\left(\sum_{k=-\infty}^{\infty}\overline{s[k]}\exp\{j2\pi k\theta\}\right)d\theta$$

$$= \sum_{n=-\infty}^{\infty} g[n]\sum_{k=-\infty}^{\infty}\overline{s[k]}\int_{-0.5}^{0.5}\exp\{-j2\pi(n-k)\theta\}d\theta$$

כבר ראינו כי האינטגרל הפנימי אינו אלא פונקצית הלם דיסקרטית (זוהי פונקצית sinc הדגומה בנקודות ההתאפסות של פונקציה זו).

$$\langle G(\theta),S(\theta)\rangle_\theta = \sum_{n=-\infty}^{\infty} g[n]\sum_{k=-\infty}^{\infty}\overline{s[k]}\int_{-0.5}^{0.5}\exp\{-j2\pi(n-k)\theta\}d\theta =$$

$$= \sum_{n=-\infty}^{\infty} g[n]\sum_{k=-\infty}^{\infty}\overline{s[k]}\delta_D[n-k] = \sum_{n=-\infty}^{\infty} g[n]\overline{s[n]}$$

כמקרה פרטי תיתן את תוצאת פרסבל:

$$\int_{-0.5}^{0.5}|G(\theta)|^2 d\theta = \sum_{n=-\infty}^{\infty}|g[n]|^2 .$$

- תכונת הקונבולוציה – עבור שני אותות g[n] ו-s[n] עם התמרתם G(θ) ו-S(θ). הגדרנו את הקונבולוציה בין שתי הפונקציות בזמן ע"י

$$g[n] = \sum_{k=-\infty}^{\infty} h[n-k]s[k]$$

נראה כי בתדר חישוב הקונבולוציה הופך למכפלה רגילה:

$$F\{g[n]\otimes s[n]\} = F\left\{\sum_{k=-\infty}^{\infty} h[n-k]s[k]\right\} =$$

$$= \sum_{n=-\infty}^{\infty}\left(\sum_{k=-\infty}^{\infty} h[n-k]s[k]\right)\cdot\exp\{-j2\pi n\theta\} =$$

$$= \sum_{k=-\infty}^{\infty} s[k]\exp\{-j2\pi k\theta\}\cdot\sum_{n=-\infty}^{\infty} h[n-k]\exp\{-j2\pi(n-k)\theta\}$$

$$= G(\theta)S(\theta)$$

משמעות תוצאה זו היא שבמקום לחשב קונבולוציה נוכל לעבור להתמרות הפורייה של שני האותות המעורבים, להכפילם זה בזה סקלרית, ואז לבצע התמרה הפוכה חזרה לזמן.

5. טיפול באותות בעלי תמך סופי

נניח כי בידנו אות s[n] שקיים רק עבור $n \in [0,N-1]$. הרי מרבית האותות בהם נעסוק יהיו כאלה, מכיוון שלא ניתן לטפל בקבצי אותות בגודל אינסופי. מה נוכל לומר על אותות אלו ביחס לסינונם, התמרתם, ועוד? קל יהיה להניח כי נטפל באות זה כאילו היה בעל תמך אינסופי, כשנניח התאפסות מחוץ לתמך $n \in [0,N-1]$. טכניקה זו קרויה המשכת האות באפסים. בדרך זו נוכל להציע קונבולוציה בין מערכת שתגובתה להלם h[n] ובין אות זה ע"י

$$h[n] \otimes s[n] = \sum_{k=0}^{N-1} h[n-k]s[k] = \sum_{k=0}^{N-1} h[k]s[n-k].$$

כל שהשתנה כאן הוא תחום הסכימה הלוקח בחשבון את התאפסות האות מחוץ לתחום מסוים. מכיוון שכך, נוכל לעשות שימוש בתכונת הקונבולוציה של התמרת הפורייה, להמיר את האות כאילו היה קיים בכל זמן, להמיר את תגובת ההלם, להכפיל, ולבצע התמרה הפוכה.

פגם מטריד בביצוע התמרת הפורייה על s[n] הוא העובדה שאנו מתחילים עם N מספרים, ומוצאים עצמנו מאפיינים אותם בתדר באמצעות ערכים ברצף בתחום $[-0.5,0.5]$ לפי:

$$S(\theta) = F\{s[n]\} = \sum_{n=-\infty}^{\infty} s[n]\exp\{-j2\pi n\theta\} = \sum_{n=0}^{N-1} s[n]\exp\{-j2\pi n\theta\}$$

וזה נראה כבזבוז. נוכל להקל זאת ולהציע לדגום את ציר התדר ב-N הנקודות $\theta = k/N$ עבור $k = -N/2, -N/2+1, \cdots -1,0,1, \cdots N/2-1$, או, כפי שיותר מקובל, להשתמש בערכי k בתחום [0,N-1]. בצורה זו נקבל את התמרת הפורייה הדיסקרטית:

$$S[k] = DFT\{s[n]\} = \sum_{n=0}^{N-1} s[n]\exp\left\{-\frac{j2\pi nk}{N}\right\} \quad k = 0,1,2, \cdots ,N-1.$$

פעולת ה-Discrete Fourier Transform (DFT) מקבלת N מספרים המתארים אות כזמן, ומחזירה N מספרים המתארים את התנהגות אות זה בתדר, כשהרעיון הוא לדגום את הפונקציה $S(\theta)$.

כיצד נצדיק את הדגימה? מדוע שלא ייווצר מצב של אובדן כתוצאה ממנה? התשובה לכך, מסתבר, פשוטה. אם כתחליף להנחה שהאות הינו אפס מחוץ לתחום [0,N-1] נניח כי הוא הומשך מחזורית, נקבל כי התמרת הפורייה הרגילה הייתה מובילה ל-$S(\theta) = F\{s[n]\}$ שאינה קיימת אלא בנקודות הדגימה, וערכיה שם היו בדיוק הערכים של

ההתמרה תחת הנחת ההמשכה באפסים! לכן, בדגימה המוצעת אנו בסך הכל משנים את כלל ההמשכה מחוץ לתחום הקיום של האות.

האם ה-DFT הפיכה? מסתבר שכן! מסתבר כי ההתמרה ההפוכה נתונה ע"י

$$s[n] = IDFT\{S[k]\} = \frac{1}{N}\cdot\sum_{k=0}^{N-1} S[k]\exp\left\{\frac{j2\pi nk}{N}\right\} \quad n = 0,1,2,\cdots,N-1$$

קל לראות נכונות התמרה הפוכה זו לפי ההגדרה:

$$IDFT\{DFT\{s[m]\}\} = \frac{1}{N}\cdot\sum_{k=0}^{N-1} S[k]\exp\left\{\frac{j2\pi nk}{N}\right\} =$$

$$= \frac{1}{N}\cdot\sum_{k=0}^{N-1}\left(\sum_{n=0}^{N-1} s[n]\exp\left\{-\frac{j2\pi nk}{N}\right\}\right)\exp\left\{\frac{j2\pi mk}{N}\right\} =$$

$$= \frac{1}{N}\cdot\sum_{n=0}^{N-1} s[n]\left(\sum_{k=0}^{N-1}\exp\left\{\frac{j2\pi(m-n)k}{N}\right\}\right)$$

הסכימה הפנימית ניתנת לחישוב לפי נוסחת טור גיאומטרי סופי,

$$\sum_{k=0}^{N-1} q^k = \frac{1-q^N}{1-q},$$

ולכן נקבל

$$\sum_{k=0}^{N-1}\exp\left\{\frac{j2\pi(m-n)k}{N}\right\} = \frac{1-\exp\{j2\pi(m-n)\}}{1-\exp\left\{\frac{j2\pi(m-n)}{N}\right\}} = N\cdot\delta_D[m-n].$$

לכל בחירה $n \neq m$ (וכן $1-N \le m-n \le N-1$ כפי שאמנם קורה) המנה הנ"ל תתאפס כי המונה מתאפס. עבור n=m מנה זו היא אפס חלקי אפס, ולפי פיתוח של פונקצית האקספוננט לפי טור טיילור ($\exp\{x\} \cong 1+x+0.5x^2$ עבור $x \to 0$) ערך המנה הוא N. לכן,

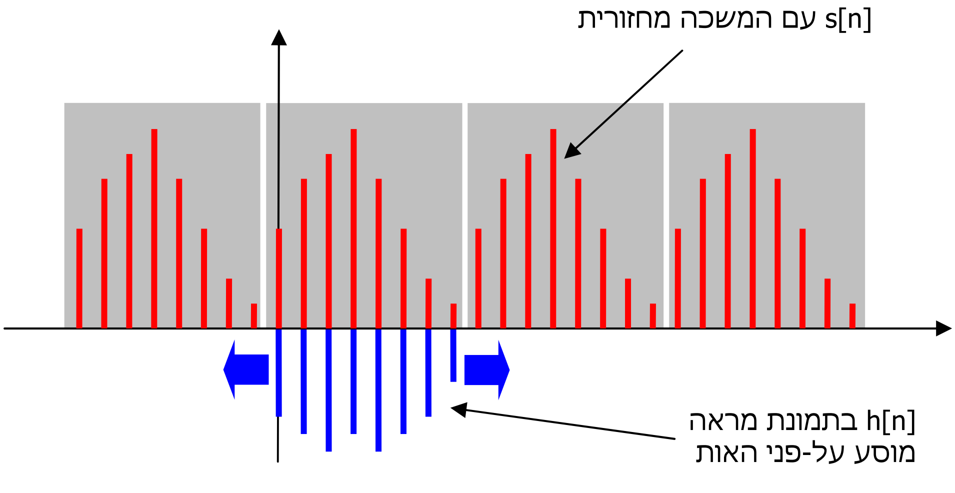
$$IDFT\{DFT\{s[m]\}\} = \frac{1}{N}\cdot\sum_{n=0}^{N-1} s[n]\left(\sum_{k=0}^{N-1}\exp\left\{\frac{j2\pi(m-n)k}{N}\right\}\right) = \sum_{n=0}^{N-1} s[n]\delta_D[m-n] = s[m]$$

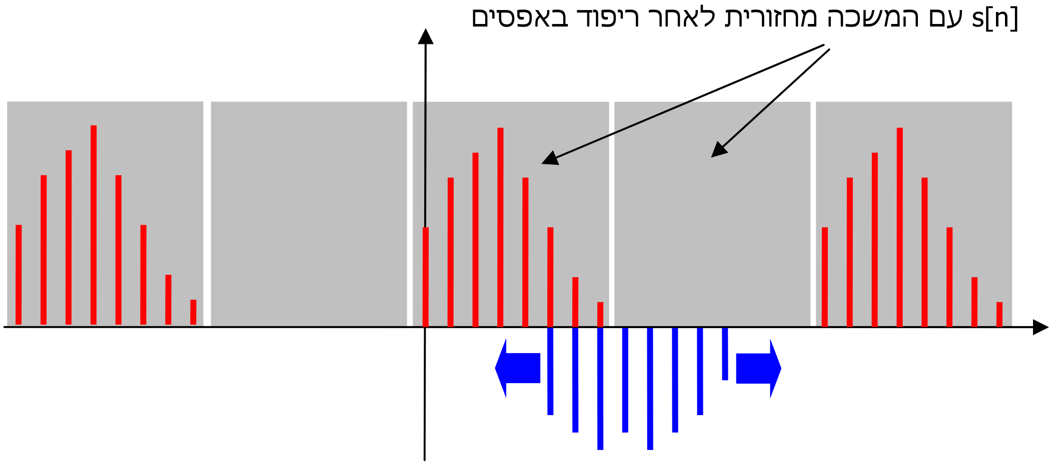
מסתבר כי ניתן לחזור ולהציג את כל אותן תכונות שהצגנו קודם (ליניאריות, הזזה ואפנון, ממוצע, ממשיות, פרסבל, קונבולוציה) גם להתמרה זו. ישנו הבדל אחד מהותי ואנחנו נבהיר אותו בקצרה – כבר ראינו בעבר כי דגימה של אות בתדר פירושה מחזוריות בזמן (ראינו זאת בהקשר לטורי פורייה). ראינו גם כי דגימה בזמן מתבטאת במחזוריות בתדר. בשל כך, בהגדרת ה- DFT דגמנו את פונקצית התדר, ולכן צפוי כי

תתקבל מחזוריות בזמן. ובאמת, אם נתבונן בהגדרת ה- IDFT ונציב לתוכה ערכי n מחוץ ל-$[0,N-1]\in n$, האות המקורי יחזור על עצמו במחזוריות של N.

תופעה זו פירושה שתכונת הקונבולוציה או הקורלציה עם DFT לא מתייחסות להמשכה של האות באפסים, אלא להמשכה מחזורית. לכן, אם בידינו שני אותות s[n] ו-h[n] – שניהם מוגדרים לתחום [0,N-1], כשנבצע DFT על כל אחד מהם, נכפיל איבר-איבר (N מכפלות) ונבצע IDFT, האפקט יהיה של קונבולוציה של h[n] עם אות מחזורי הנבנה מ-s[n] (או להיפך). קונבולוציה זו קרויה קונבולוציה ציקלית, והיא מתוארת בציור הבא.

ומה אם נרצה קונבולוציה רגילה (ליניארית) בין שני אותות אלו? עלינו להאריך את שני האותות באפסים לאורך כפול מאורכם המקורי, ואז לעשות את הקונבולוציה בעזרת DFT. למרות היותה ציקלית, ה-"ריפוד" באפסים ייתן את התוצאה הנכונה, כפי שמתואר בציור הבא.





6. חזרה למטריצות ווקטורים

כשבידינו אות $s[n]$ שקיים רק עבור $n \in [0, N-1]$, נוכל להתייחס אליו כווקטור $\underline{s}$ באורך N. מערכת ליניארית הפועלת עליו הינה מטריצה **M** בגודל N-על-N. נראה דוגמה להמחיש עניין זה. נניח כי המערכת שלנו פועלת על אותות בגודל של 8 דגימות, ומבצעת סיכום פשוט מהצורה

$$g[n] = as[n] + bs[n+1] + cs[n-1] \,.$$

תחת הנחה של טיפול בקצוות ע"י המשכה באפסים, המטריצה המתארת את המערכת תהיה

$$\mathbf{M} = \begin{bmatrix} a & b & 0 & 0 & 0 & 0 & 0 & 0 \\ c & a & b & 0 & 0 & 0 & 0 & 0 \\ 0 & c & a & b & 0 & 0 & 0 & 0 \\ 0 & 0 & c & a & b & 0 & 0 & 0 \\ 0 & 0 & 0 & c & a & b & 0 & 0 \\ 0 & 0 & 0 & 0 & c & a & b & 0 \\ 0 & 0 & 0 & 0 & 0 & c & a & b \\ 0 & 0 & 0 & 0 & 0 & 0 & c & a \end{bmatrix}.$$

המבנה הנ"ל נקרא מטריצת Toeplitz תלת אלכסונית. במטריצת טופליץ יאוכלס אותו ערך לאורך האלכסונים. מטריצה כזו מתארת בהכרח מערכת ליניארית קבועה בזמן, כיוון שהפעולה על האיבר ה-k וה-j צריכה להיות זהה עם הזזה מתאימה – עובדה היוצרת את ההזזות של השורות לקבלת מבנה הטופליץ.

אותה מערכת תחת הנחה כי ישנה המשכה מחזורית תתואר כ:

$$\mathbf{M} = \begin{bmatrix} a & b & 0 & 0 & 0 & 0 & 0 & c \\ c & a & b & 0 & 0 & 0 & 0 & 0 \\ 0 & c & a & b & 0 & 0 & 0 & 0 \\ 0 & 0 & c & a & b & 0 & 0 & 0 \\ 0 & 0 & 0 & c & a & b & 0 & 0 \\ 0 & 0 & 0 & 0 & c & a & b & 0 \\ 0 & 0 & 0 & 0 & 0 & c & a & b \\ b & 0 & 0 & 0 & 0 & 0 & c & a \end{bmatrix}.$$

גם זו מטריצת טופליץ, כי לאורך כל אלכסון יש את אותו ערך. למעשה זו מטריצה מיוחדת יותר המוכרת כמטריצה סיבובית (circulant) – כל שורה בה מתקבלת מהשורה הקודמת ע"י הזזה ימינה, והערך שנופל מימין נכנס משמאל. גם מטריצות סיבוביות מתארות מערכות ליניאריות וקבועות בזמן, אלא שאלה מתייחסות להמשכה מחזורית מחוץ לתמך האות.

פעולת ה- DFT יכולה להיות מיוצגת בצורה מטריצית, כמטריצה ריבועית בגודל N-על-N המקבלת וקטור באורך N ומכפילה אותו לקבלת וקטור מוצא באותו אורך. המספרים המאכלסים את המטריצה הזו הם $\exp\{-j2\pi nk/N\}$,

$$\mathbf{F}=\begin{bmatrix} \exp\left\{-\frac{j2\pi 0\cdot 0}{N}\right\} & \exp\left\{-\frac{j2\pi 1\cdot 0}{N}\right\} & \exp\left\{-\frac{j2\pi 2\cdot 0}{N}\right\} & \cdots & \exp\left\{-\frac{j2\pi (N-1)\cdot 0}{N}\right\} \\ \exp\left\{-\frac{j2\pi 0\cdot 1}{N}\right\} & \exp\left\{-\frac{j2\pi 1\cdot 1}{N}\right\} & \exp\left\{-\frac{j2\pi 2\cdot 1}{N}\right\} & \cdots & \exp\left\{-\frac{j2\pi (N-1)\cdot 1}{N}\right\} \\ \exp\left\{-\frac{j2\pi 0\cdot 2}{N}\right\} & \exp\left\{-\frac{j2\pi 1\cdot 2}{N}\right\} & \exp\left\{-\frac{j2\pi 2\cdot 2}{N}\right\} & \cdots & \exp\left\{-\frac{j2\pi (N-1)\cdot 2}{N}\right\} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \exp\left\{-\frac{j2\pi 0\cdot (N-1)}{N}\right\} & \exp\left\{-\frac{j2\pi 1\cdot (N-1)}{N}\right\} & \exp\left\{-\frac{j2\pi 2\cdot (N-1)}{N}\right\} & \cdots & \exp\left\{-\frac{j2\pi (N-1)\cdot (N-1)}{N}\right\} \end{bmatrix}.$$

מטריצה זו בעלת תכונות מיוחדות – היא יוניטרית עד כדי קבוע (כלומר, היפוכה מתקבל ע"י שחלוף צמוד), ותכונה זו מסבירה את נכונות תכונת פרסבל.

תכונה מעניינת של מטריצת התמרת הפורייה **F** היא היותה מלכסנת יוניטרית את כל המטריצות הסיבוביות.

משפט 6: בהינתן מטריצה **C** סיבובית, מטריצת ה-DFT מלכסנת אותה יוניטרית, כלומר, $\mathbf{FCF}^H=\mathbf{D}$ היא מטריצה אלכסונית. איברי האלכסון הינם ה- DFT על השורה הראשונה ב-**C**.

קל להוכיח משפט זה. נזכור כי מטריצה סיבובית ניתנת גם להתייחסות כמטריצה בה העמודות מתקבלות ע"י הזזה מטה ומילוי למעלה של האיבר הנפלט. נתייחס לביטוי $\underline{v}_1=\mathbf{F}\underline{c}_1$ אשר מבצע DFT על העמודה הראשונה של **C**. נניח כי $\underline{v}_1$ ידוע. מה נוכל לומר על $\underline{v}_2=\mathbf{F}\underline{c}_2$? זהי התמרה על וקטור מוזז לפי

$$\begin{aligned} v_2[k] &= \sum_{n=0}^{N-1} c_2[n]\exp\left\{-\frac{j2\pi nk}{N}\right\}= \\ &= c_1[N-1]\exp\left\{-\frac{j2\pi 0\cdot k}{N}\right\}+\sum_{n=1}^{N-1} c_1[n-1]\exp\left\{-\frac{j2\pi nk}{N}\right\}= \\ &\underset{m=n-1}{\overset{\uparrow}{=}} c_1[N-1]\exp\left\{-\frac{j2\pi 0\cdot k}{N}\right\}+\sum_{m=0}^{N-2} c_1[m]\exp\left\{-\frac{j2\pi mk}{N}\right\}\exp\left\{-\frac{j2\pi k}{N}\right\}= \\ &= \exp\left\{-\frac{j2\pi k}{N}\right\}\cdot\sum_{m=0}^{N-1} c_1[m]\exp\left\{-\frac{j2\pi mk}{N}\right\}=\exp\left\{-\frac{j2\pi k}{N}\right\}\cdot v_1[k] \end{aligned}$$

ברישום מרוכז של וקטור כניסה מול וקטור מוצא נקבל

$$\underline{v}_2=\begin{bmatrix}\exp\left\{-\frac{j2\pi 0}{N}\right\}\cdot v_1[0]\\ \exp\left\{-\frac{j2\pi 1}{N}\right\}\cdot v_1[1]\\ \vdots\\ \exp\left\{-\frac{j2\pi (N-1)}{N}\right\}\cdot v_1[N-1]\end{bmatrix}=\begin{bmatrix}v_1[0] & 0 & \cdots & 0\\ 0 & v_1[1] & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & v_1[N-1]\end{bmatrix}\underline{f}_1,$$

כאשר $\underline{f}_1$ היא העמודה הראשונה מ-**F**. באופן דומה, התמרת העמודה ה-i תיתן

$$\underline{v}_i=\mathbf{F}\underline{c}_i=\begin{bmatrix}\exp\left\{-\frac{j2\pi i\cdot 0}{N}\right\}\cdot v_1[0]\\ \exp\left\{-\frac{j2\pi i\cdot 1}{N}\right\}\cdot v_1[1]\\ \vdots\\ \exp\left\{-\frac{j2\pi i\cdot (N-1)}{N}\right\}\cdot v_1[N-1]\end{bmatrix}=\begin{bmatrix}v_1[0] & 0 & \cdots & 0\\ 0 & v_1[1] & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & v_1[N-1]\end{bmatrix}\underline{f}_i.$$

ברישום מרוכז נקבל כי

$$\mathbf{FC}=\mathbf{F}\begin{bmatrix}| & & |\\ \underline{c}_1 & \cdots & \underline{c}_N\\ | & & |\end{bmatrix}=\begin{bmatrix}v_1[0] & 0 & \cdots & 0\\ 0 & v_1[1] & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & v_1[N-1]\end{bmatrix}\mathbf{F},$$

כטענת המשפט.

7. התמרת פורייה דיסקרטית מהירה - FFT

כשכופלים מטריצה בגודל N-על-N בוקטור באורך N יש N^2 מכפלות וסיכומים. תכונה חשובה של ה- DFT היא היכולת לבצע ההתמרה במורכבות של NlogN, והאלגוריתם המבצע זאת קרוי Fast Fourier Transform (FFT). אלגוריתם זה נפוץ והשימוש בו רווח. במסגרת זו נראה את העיקרון הבסיסי עליו מושתת הרווח החישובי הזה.

ההתמרה המקורית המלאה נעשית ע"י

$$S[k]=DFT\{s[n]\}=\sum_{n=0}^{N-1}s[n]\exp\left\{-\frac{j2\pi nk}{N}\right\}\quad k=0,1,2,\cdots,N-1.$$

נפריד סכימה זו לאיברים זוגיים ואי-זוגיים של הסכימה בהנחה ש-N הינו חזקה שלמה של 2, ונקבל

$$\begin{aligned}S[k]&=\sum_{n=0}^{N/2-1}s[2n]\exp\left\{-\frac{j2\pi 2nk}{N}\right\}+\sum_{n=0}^{N/2-1}s[2n+1]\exp\left\{-\frac{j2\pi(2n+1)k}{N}\right\}=\\ &=\sum_{n=0}^{N/2-1}s_{even}[n]\exp\left\{-\frac{j2\pi nk}{N/2}\right\}+\exp\left\{-\frac{j2\pi k}{N}\right\}\sum_{n=0}^{N/2-1}s_{odd}[n]\exp\left\{-\frac{j2\pi nk}{N/2}\right\}\end{aligned}.$$

קיבלנו כי יש לקחת שני אותות - $\underline{s}_{even}$ ו- $\underline{s}_{odd}$ כשכל אחד באורך N/2 ולבצע עליהם DFT, כשהפעם עלינו לעשות זאת ל- $k = 0,1,\dots, N/2-1$:

$$\underline{S}_{even} = \mathbf{F}_{N/2} \cdot \underline{s}_{even}$$
$$\underline{S}_{odd} = \mathbf{F}_{N/2} \cdot \underline{s}_{odd}$$

בהינתן שני וקטורים אלה, מיזוגם ע"י

$$S[k] = S_{even}[k] + \exp\left\{-\frac{j2\pi k}{N}\right\} S_{odd}[k]$$

ייתן את התוצאה הרצויה. חשוב להבהיר כי נשתמש בנוסחא הנ"ל ב- k בתחום $k = 0,1,2,\cdots, N-1$ עם ניצול המחזוריות של שתי הסדרות הבסיסיות. באופן זה, פעולה שבמקור דרשה N^2 מכפלות וסיכומים מרוכבים תדרוש כעת $2N^2/4$ פעולות כנ"ל להשגת שתי ההתמרות, ועוד N פעולות מכפלה וסיכום לשם מיזוגם.

כעת ניקח את כל אחת מההתמרות המוקטנות המייצרות את $\underline{S}_{even}, \underline{S}_{odd}$ ונפרק גם אותה לזוגיים ואי-זוגיים. נידרש כעת ל- $4N^2/16$ פעולות לביצוע ההתמרות (כל התמרה קטנה תדרוש $N^2/16$ פעולות, בניית $\underline{S}_{even}$ דורשת שתי התמרות כאלה, וכך גם $\underline{S}_{odd}$), ו-N למיזוגן (נדרשים N/2 לשם מיזוג וקבלת $\underline{S}_{even}$, וכנ"ל לבניית $\underline{S}_{odd}$). אם נמשיך בפירוק זה נקבל כי סך הפעולות הנדרש בסופו של דבר נוגע רק למיזוגים, וזה נתון ע"י N כשהוא מוכפל ב-$\log_2 N$ שלבי הפירוק (עד שמגיעים להתמרת מספר יחיד שהיא אופרטור יחידה פשוט). קיבלנו באופן זה את התמרת הפורייה הדיסקרטית המהירה – .Fast Fourier Trans