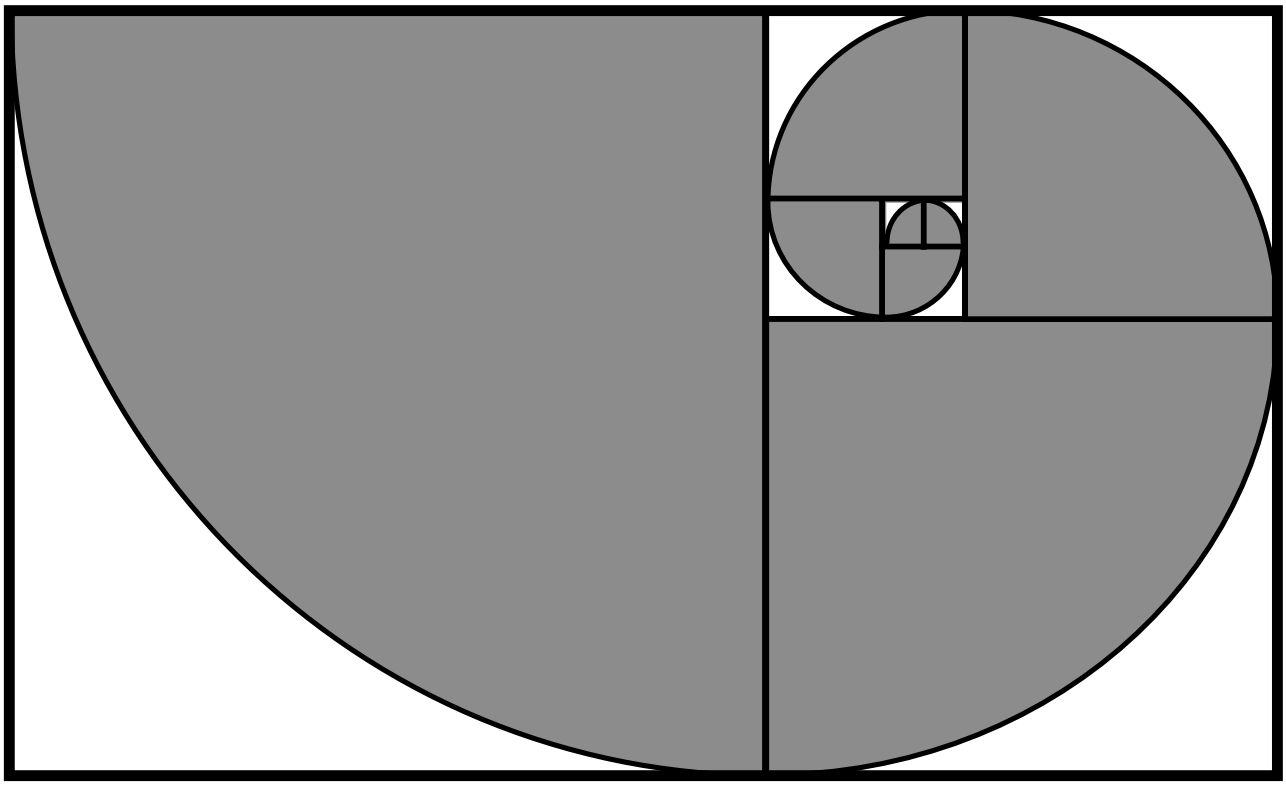
מיכאל אלעד

שיטות חישוביות באופטימיזציה



הטכניון

1998

הקדמה

ספר זה נכתב ממערכי שיעור עבור הקורס "שיטות חישוביות באופטימיזציה", אשר ניתן במסגרת הפקולטה להנדסת חשמל בטכניון. הבסיס לחומר המתואר בספר זה הוא סיפרו של David G. Luenberger הנקרא - Linear and Non-Linear Programming. עם זאת, במהלך כתיבת הפרקים ראיתי לנכון לשנות סדר הופעת הנושאים, שינוי הוכחות שונות, וכן הכנסת נושאים נוספים ודוגמאות מיוחדות אשר אינם מופיעים בספר זה. לצורך כך נעזרתי במספר ספרים ומקורות נוספים (רשימתם מצורפת). ספר זה הינו מהדורה שלישית הכוללת תיקונים ותוספות - לפי לקחים מהוראה על-פיו בשנים 1997 ו- 1998.

מבחינת כיסוי הפרקים מתורת האופטימיזציה, ספר זה מספק סקירה נרחבת של שיטות איטרטיביות המיועדות לפתרון בעיות מינימיזציה ללא אילוצים. חלקו השני של הספר אשר דן בבעיות אופטימיזציה עם אילוצים מתמקד במשפחת בעיות התכנות הליניארי ואלגוריתם ה-Simplex אשר נועד לפתרונן. ישנה סקירה של שיטות כלליות לפתרון בעיות אופטימיזציה עם אילוצים אך סקירה זו שטחית בהתחשב בכך שבסבירות גבוהה לא ניתן לתת ניתוח מעמיק יותר במסגרת קורס סמסטריאלי המנסה להקיף את הכל.

מקורות

1. David G. Luenberger, **Linear and Non Linear Programming**, Addison-Wesley Pub. Co. , Menlo-Park, California, Second Edition, 1987.
2. Dimitri P. Bertsekas, **Nonlinear Programming**, Athena Scientific, Belmont, Massachusetts, 1995.
3. W. H. Press, S. A. Teukolsky, W. T. Vatterling & B. P. Flannery**, Numerical recepies In C**, Cambridge University Press, 1987.
4. R. Fletcher, **Practical Methods of Optimization**, Jhon Wiley & Sons, Second Edition, 1987.
5. R. Meir, Lecture Notes in Optimization Theory
6. L.A. Hageman & D. M. Young, **Applied Iterative Methods**, Academic Press, 1981.
7. Donald A. Pierre, **Optimization Theory with Applications**, Dover Publications, Inc., New York, 1986.

תוכן העניינים

פרק 1 - תורת האופטימיזציה - כללי 1

1.1 כללי 1

1.2 דוגמאות לבעיות אופטימיזציה 3

1.3 חזרה מתמטית - אלגברה ליניארית 8

1.4 חזרה מתמטית - חשבון דיפרנציאלי ואינטגרלי 11

1.5 חזרה מתמטית - תורת הקבוצות 13

פרק 2 - תנאים הכרחיים ומספיקים לנקודת מינימום 15

2.1 נקודת מינימום לוקלית ונקודת מינימום גלובלית 15

2.2 תנאים הכרחיים מסדר ראשון 16

2.3 תנאים הכרחיים מסדר שני 19

2.4 תנאים מספיקים מסדר שני לנקודת מינימום 22

2.5 תנאים הכרחיים ומספיקים לבעיות מעל תחום הגדרה קומפקטי 24

2.6 סיכום 28

פרק 3 - פונקציות וקבוצות קמורות 31

3.1 פונקציות קמורות - הגדרות ותכונות בסיסיות 31

3.2 תכונות של פונקציות קמורות גזירות 33

3.3 מינימיזציה ומקסימיזציה של פונקציות קמורות 35

3.4 סיכום 38

פרק 4 - עקרונות של אלגוריתמים למינימיזציה של פונקציה 39

4.1 מבנה עקרוני של אלגוריתם נומרי למינימיזציה של פונקציה 39

4.2 פונקצית ירידה 41

4.3 מיפויים סגורים 42

4.4 משפט ההתכנסות הגלובלית 43

4.5 אומדנים למהירות התכנסות 46

4.6 סיכום 50

פרק 5 - אלגוריתמי חיפוש על ישר 51

5.1 מדוע חיפוש על ישר ? 51

5.2 סידרת פיבונצ'י וחיפוש הזהב 54

5.3 התאמת עקומים - גישת ניוטון 60
5.4 התאמת עקומים - גישת האינטרפולציה 63
5.5 התאמת עקומים - גישת ההתאמה הריבועית 65
5.6 קריטריוני עצירה לאלגוריתמי החיפוש על ישר, וחיפוש מקורב 67
5.7 סיכום 67

פרק 6 - אלגוריתמי מינימיזציה ללא חישוב נגזרות 69

6.1 אלגוריתם דעיכה על קואורדינטות 69
6.2 אלגוריתם ה- Simplex הכללי 69
6.3 קירובי נגזרות 73
6.4 סיכום 74

פרק 7 - אלגוריתמי מינימיזציה מבוססי גרדיאנט 75

7.1 אלגוריתמי מינימיזציה מבוססי גרדיאנט 75
7.2 הפונקציה הריבועית ושימושיה באופטימיזציה 77
7.3 ההתכנסות של אלגוריתם ה- SD עבור הבעיה הריבועית 79
7.4 ההתכנסות של אלגוריתם ה- NSD עבור הבעיה הריבועית 84
7.5 ההתכנסות של אלגוריתם ניוטון 87
7.6 וריאציות של אלגוריתם ניוטון 89
7.7 ההתכנסות של אלגוריתם ה- DNSD עבור הבעיה הריבועית 92
7.8 סיכום 94

פרק 8 - שיטות האצה לאלגוריתם ה- SD 95

8.1 גישת PARTAN לניטרול תופעת הזיג-זג 95
8.2 האצה פולינומיאלית של אלגוריתמים 98
8.3 האצת ה- SD בגישת המומנט 102
8.4 סיכום 106

פרק 9 - אלגוריתם הגרדיאנטים הצמודים - CG 107

9.1 שיטת הכיוונים הצמודים 107
9.2 אלגוריתם הגרדיאנטים הצמודים - CG 112
9.3 התכנסות אלגוריתם ה- CG 115
9.4 אלגוריתם ה- CG החלקי 122

9.5 אלגוריתם ה- CG לבעיות שאינן ריבועיות 124
9.6 סיכום 125

פרק 10 - השיטות הניוטוניות למחצה 127

10.1 כללי 127
10.2 בניית ההסיין ע"י עדכונים מדרגה 1 127
10.3 אלגוריתם ה- DFP 131
10.4 אלגוריתמי Broyden 136
10.5 תכונות התכנסות של האלגוריתמים הניוטוניים למחצה 139
10.6 שיטות ניוטוניות למחצה חסרות זכרון 142
10.7 סיכום 143

פרק 11 - בעיות LS ודרכים לפתרון מערכות של משוואות 145

11.1 כללי 145
11.2 פתרון בעיות LS ליניאריות 147
11.3 בעיית ה- LS הכללית - אלגוריתם Gauss-Newton 149
11.4 בעיית LS לא ליניארית - מבנים ייחודיים 154
11.5 בעיות LS והקשר לפתרון מערכות של משוואות 157
11.6 סיכום 166

פרק 12 - בעיות תכנות ליניארי ופתרונן 167

12.1 מבנה קלאסי של בעיות תכנות ליניארי 167
12.2 המשפט היסודי של התכנות הליניארי 170
12.3 הקשר בין אילוצי בעיית תכנות ליניארי ותחום קמור 173
12.4 אלגוריתם ה- Simplex 178
12.5 איתחול אלגוריתם ה- Simplex 187
12.6 אלגוריתם ה- Simplex עבור ערכים חסומים מלמעלה 190
12.7 הגדרת הבעיה הדואלית ומאפייניה 191
12.8 סיכום 194

פרק 13 - תנאים הכרחיים ומספיקים בבעיות עם אילוצים 195

13.1 כללי 195
13.2 תנאים הכרחיים מסדר ראשון - אילוצי שיויון 196
13.3 תנאים הכרחיים ומספיקים מסדר שני - אילוצי שיויון 201

13.4 רגישות והקשר לכופלי לגרנז' 208
13.5 אילוצי אי-שיויון - הכללת התוצאות הקודמות 212
13.6 סיכום 217

פרק 14 - בעיות אופטימ' עם אילוצים - גישות לפתרון 219

14.1 כללי 219
14.2 גישת המחיר - The Penalty Method 219
14.3 גישת המחסום - The Berrier Method 224
14.4 שיטת ההטלה למרחב האילוצים 227
14.5 שיטת ה- Interior Point לפתרון בעיות תכנות ליניארי 231
14.6 שיטות מבוססות לגרנז' 233
14.7 סיכום 236

פרק 1 - תורת האופטימיזציה - כללי

1.1 כללי

נתחיל בשאלה המתבקשת - מהי בעיית אופטימיזציה? נגדיר את הבעיה הבאה - נתונה פונקציה $f(x): \mathbf{R}^n \to \mathbf{R}^1$, כלומר פונקציה המקבלת וקטור של n ערכים המסומן x, ומניבה ערך סקלרי. בעיית אופטימיזציה קלאסית תהיה - מצא את הוקטור x אשר מגשים את המינימום של הפונקציה f, דהיינו - $\hat{x} = \text{ArgMin}\, f(x)$. 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- 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 - $x \in \Omega$.

ב. אילוצי שיויון - $h(x) = 0$.

ג. אילוצי אי-שיויון - $g(x) = 0$.

במסגרת קורס 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 המיחשוב גבר העניין בדרכים נומריות ובתחום זה מרוכז הקורס "שיטות חישוביות באופטימיזציה".

המעבר מפתרון אנליטי לגישות נומריות אינו נובע רק מהתפתחות עולם המיחשוב. 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 מסויימים, 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.

"שיטות חישוביות באופטימיזציה" הוא בעצם פרק אחד מיני רבים בתחום הקרוי אנליזה נומרית. אנליזה נומרית היא פרק מרתק במתמטיקה שימושית בו עוסקים בשאלה - כיצד (אלו אלגוריתם) פותרים בעיות מתמטיות שונות ע"י מחשב. לדוגמה, חישוב של נורמה של מטריצה או ערכים עצמיים עלולים להיות פעולות מורכבות למדי עבור מטריצות גדולות - אך ישנם משפטים שונים אשר מספקים קירובים לגודל זה, וניתן בעזרתם להציע אלגוריתמים איטרטיביים אשר יתכנסו לערך הנורמה הנכון. סוגיה אחרת אשר זכתה לטיפול נרחב היא פתרון נומרי של משוואות דיפרנציאליות חלקיות בהן הנעלם היא פונקציה. יש קשר ואף חפיפה בין מגוון התחומים באנליזה נומרית. כך למשל - ידיעת הפתרון לבעיית אופטימיזציה שקול לידיעת פתרון מערכת משוואות ולהיפך, כיוון שבהינתן מערכת משוואות מהצורה:

$$k=1,2,\ldots,n \qquad h_k(x_1,x_2,\ \ldots\ ,x_n)=0$$

אנו יכולים להגדיר בעיית מינימיזציה באופן הבא:

$$f(x_1,x_2,\ \ldots\ ,x_n)=\sum_{k=1}^{n}h_k^2(x_1,x_2,\ \ldots\ ,x_n)$$

וברור 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.

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 קורס זה היא אחרת. בתחילה נתייחס לטיפול בבעיות אופטימיזציה נטולות אילוצים - על תחום זה נתעכב כשני שלישים ראשונים של הקורס. בשליש האחרון של הקורס נעסוק בבעיות 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.

שיטת העבודה במהלך קורס זה תהיה הבאה - אנו נפתח בדיון אנליטי ונספק תוצאות אנליטיות בסיסיות בתחום האופטימיזציה. על יסוד תוצאות אלו נציע אלגוריתמים נומריים, ננתח אותם ונחשוף את חולשותיהם. באופן זה נפתח אבולוציה של אלגוריתמים משתפרים והולכים. התהליך הנ"ל יקרה פעמיים - פעם לטיפול בבעיות ללא אילוצים, ופעם לטיפול בבעיות עם אילוצים.

מהו העניין שלנו כמהנדסים בפרק זה של אלגוריתמי אופטימיזציה נומריים? מסתבר כי אין כמעט תחום בהנדסת חשמל, תעשיה וניהול, מכונות, הנדסה אזרחית, מדעי המחשב, כימיה ועוד, אשר אינו נדרש לפרק זה. מהנדסי חשמל ימצאו צורך לתחום זה בעיקר בבקרה, עיבוד אותות ותמונות, ראיה ממוחשבת, תקשורת, מבנה מחשבים, אופטיקה ואלקטרו-אופטיקה. מהנדסי תעשיה וניהול יישרו לתחום זה בכלכלה, ניהול, סטטיסטיקה, חקר-ביצועים, חשבונאות ועוד. וכך גם ניתן להגדיר לפירקי הנדסה אחרים תחומים בעלי ענין באופטימיזציה.

בכדי ליצור מוטיבציה לדיון בבעיות אופטימיזציה נציג במסגרת פרק זה מספר דוגמאות לבעיות קלאסיות המתעוררות במדעים, ואשר הופכות להיות בעיות אופטימיזציה. לעניין זה יוקדש הסעיף הבא. במהלך הקורס נידרש לידע בסיסי באלגברה ליניארית, חשבון דיפרנציאלי ואינטגרלי, ובתורת הקבוצות. בסיומו של פרק זה ניתן חזרה מהירה על נושאים ומושגים שונים אשר יעלו במהלך הסמסטר.

1.2 דוגמאות לבעיות אופטימיזציה

דוגמה 1.1: (הקצאת משאבים). לרשותנו 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 נאמר את הדבר הבא - 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 -

$$\sum_{n=1}^{N} X_{m,n} = S_m \quad m = 1,2,\ldots,M \qquad \sum_{m=1}^{M} X_{m,n} = C_n \quad n = 1,2,\ldots,N$$

$$\Rightarrow \sum_{n=1}^{N} C_n = \sum_{m=1}^{M} S_m \qquad X_{m,n} \geq 0 \quad \forall m,n$$

M אילוצים ראשונים מתייחסים לכך שכל תחנה תספק את מלוא כוח היצור שלה. N אילוצים נוספים מבטיחים כי כל צרכן יקבל את הדרוש לו. שיויון בין הצריכה והייצור הכוללים הינו אילוץ על הנתונים ואינו חלק מהבעיה. תחת האילוצים הנ"ל, רצוננו להביא למינימום את הביטוי:

$$\text{Cost} = \sum_{m=1}^{M} \sum_{n=1}^{N} X_{m,n} P_{m,n}$$

הבעיה אותה הגדרנו היא בעיית אופטימיזציה עם אילוצים בה כל הקשרים בין הנעלמים לנתונים הם קשרים ליניאריים. בעיה זו קרויה בעיית **תכנות ליניארי** והיא עולה באופן תכוף בכלכלה, בתכנון משאבים, ניהול, תקשורת מחשבים, רשתות ומעגלים חשמליים, תורת הגרפים ועוד. אנו נעסוק במשפחה זו של בעיות לקראת סופו של הקורס, ונכיר דרך ייחודית לפתרון בעיות אלו - גישת ה- Simplex.

דוגמה 1.2: (צ'יקן באגט) 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 צ'יקן באגט 8.5 שקלים מתוך המטבעות הנ"ל באופן שיוריד מעלינו משקל גבוה ככל האפשר. נסמן ב - $[nn_1, nn_2, nn_3, nn_4]$ (שלמים בלבד!) את מספר המטבעות מכל סוג בהתאמה בהן נשתמש לתשלום. נוכל לכתוב את האילוצים הבאים:

$$0 \leq nn_k \leq n_k \quad 1 \leq k \leq 4$$
$$10nn_1 + 50nn_2 + 100nn_3 + 500nn_4 = 850$$

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 -

$$\text{Weight} = \sum_{k=1}^{4} m_k \cdot nn_k$$

גם הפעם התקבלה בעיית תכנות ליניארי, אלא שהפעם נכללים אילוצי אי-שוויון, ובנוסף, אנו מחפשים את הפתרון רק מעל שדה השלמים. אנו נראה בהמשך הסמסטר כיצד ממירים בעיית תכנות ליניארי בעלת אילוצי אי-שיוויון לתבנית סטנדרטית של בעיה עם אילוצי שוויון בלבד. במסגרת קורס זה לא נעסוק בבעיה האחרת המתעוררת פה - עניין היות הפתרון מעל השלמים.

דוגמה 1.3: (שחזור תמונה מהיטלים). בהדמיה רפואית רצוננו להפיק תמונה של חתך הגוף במקום כלשהו. התמונה הרצויה היא X, והיא לא ידועה וגם אינה ניתנת לצילום - אלא אם כן נחתוך את היעד המצולם (- ואנו נצא מנקודת הנחה שהוא לא יאהב את זה). כתחליף, נשיג היטלים של התמונה שלענייננו הם בעלי הקשר המתמטי הבא אל תמונת הידע - $Y_k = H_k X + N_k$. Y_k הוא תוצאת ההיטל ה- k של התמונה. H_k הוא אופרטור ההטלה ה - k אותו אנו מניחים כידוע. N_k הוא רעש מדידה בהיטל ה- k. כעת, מתוך M היטלים רצוננו למצוא (או למעשה - לשערך) את X. נציע את קריטריון הטיב הבא:

$$\varepsilon^2(X) = \sum_{k=1}^{M} \left\| Y_k - H_k X \right\|^2 = \sum_{k=1}^{M} \left[Y_k - H_k X \right]^T \left[Y_k - H_k X \right]$$

ונחפש את X אשר יביא ביטוי זה למינימום.

התבנית שהוצגה לעיל הינה תבנית ריבועים פחותים קלאסית בה עושים שימוש תכוף בעיבוד אותות, עיבוד תמונות, תקשורת ועוד. בהנחה שהתמונה הנדרשת נבנית באלפים רבים של נקודות, זוהי גם כמות הנעלמים, ודרושות גישות מיוחדות לפתרון נומרי של בעיות כאלה עם מורכבות חישובית סבירה. בחלקו הראשון של הקורס נכיר גישות איטרטיביות לפתרון בעיות אופטימיזציה במבנה זה.

דוגמה 1.4: (בקרה). נניח כי נתונה מערכת אשר אמורה לנעול על מטרה. המשתנה x ייצג את שגיאת העקיבה. נניח כי מצבה ההתחלי של המערכת הוא x_0, ורצוננו להביאה תוך m צעדים לשגיאת עקיבה מינימלית (קרובה ככל האפשר ל- 0). לצורך כך אנו מפעילים בכל נקודת זמן k כוח תיקון u_k. ידוע כי הקשר בין התיקון ושגיאת העקיבה נתון ע"י: $x_{k+1} = x_k + u_k$. נתון נוסף הידוע לנו הוא שהפעלת תיקון כרוכה בבזבוז אנרגיה הפרופורציונית לריבוע כוח התיקון.

נשים לב לעובדה כי עד כה לא הוגדר קריטריון טיב אשר לפיו יש לתכנן את הבקרה האופטימלית, ובלעדי קריטריון זה - הפתרון אינו יחיד. נציע את הקריטריון הבא:

$$\varepsilon^2(u_1,u_2,\cdots,u_m)=\alpha\cdot\sum_{k=0}^{m}x_k^2+\beta\cdot\sum_{k=0}^{m}u_k^2$$

הגורם הראשון בפונקציה הנ"ל מייצג את מידת הקירבה אל היעד -אפס שגיאה. הגורם השני סוכם את האנרגיה אשר בוזבזה בתהליך תיקון העקיבה. ניתן להציב את הקשר הבא:

$$x_k=x_{k-1}+u_{k-1}=x_{k-2}+u_{k-2}+u_{k-1}=\cdots=x_0+\sum_{j=0}^{k-1}u_j$$

בתוך הפונקציה שהוגדרה לעיל ולקבל:

$$\varepsilon^2(u_0,u_1,\cdots,u_{m-1})=\alpha\cdot\sum_{k=0}^{m-1}\left[x_0+\sum_{j=0}^{k-1}u_j\right]^2+\beta\cdot\sum_{k=0}^{m-1}u_k^2$$

בדומה לבעיה הקודמת, גם בעיה זו הינה בתבנית קוודרטית (ריבועית), וגישות דומות משמשות פתרונה. תחום שלם בבקרה אשר נגזר מאלגוריתמים מתורת האופטימיזציה קרוי "בקרה אופטימלית", בו עוסקים באיפיון מערכת בקרה ע"י סידרת פרמטרים, ומחפשים את ערכיהם אשר יביאו קריטריון טיב כלשהו לאופטימום. כזו היא גם הבעיה אותה הצגנו כאן.

דוגמה 1.5: (קירוב מודל AR). נתון וקטור של מדידות $\{x_k\}_{k=1}^{N}$ של תופעה כלשהי (לדוגמה - אות מכ"מי כלשהו). רצוננו להתאים לאות זה מודל אוטורגרסיבי מסדר I, דהיינו להניח כי מדידות אלו התקבלו מאות אחר $\{u_k\}_{k=1}^{N}$ אשר עבר דרך מערכת בעלת קוטב יחיד מהצורה:

$$x_k=a\cdot x_{k-1}+u_k$$

השאלה שעל הפרק היא מהו ערכו של a אם ידוע כי $\{u_k\}_{k=1}^{N}$ צריך להיות בעל מינימום אנרגיה?

על מנת לפתור בעיה זו נשים לב כי רצוננו להביא למינימום את הביטוי:

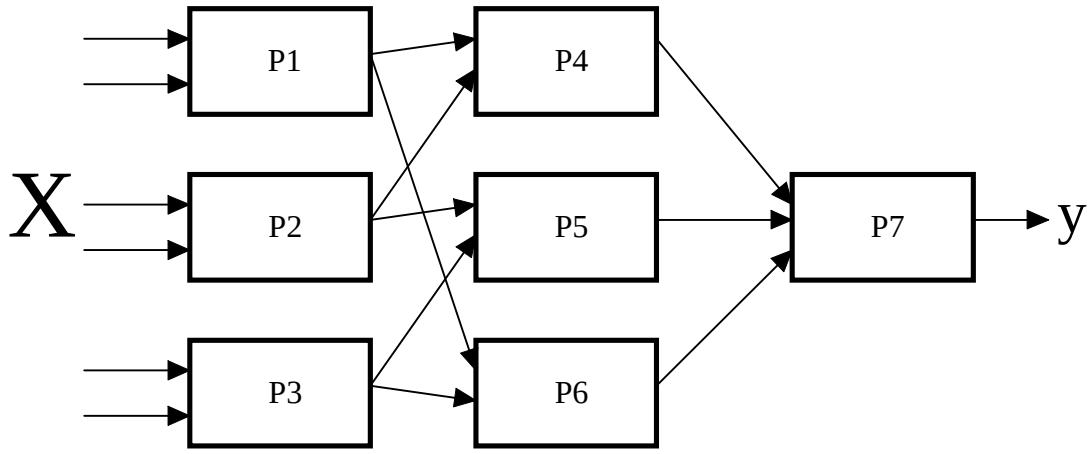
$$E^2=\sum_{k=1}^{N}[x_k-a\cdot x_{k-1}]^2=[x_1-a\cdot x_0]^2+\sum_{k=2}^{N}[x_k-a\cdot x_{k-1}]^2$$

כשהנעלמים הם a , x_0.

דוגמה 1.6: (רשת נוירונים). רשת נוירונים בנויה ממרכיבים אטומיים (נוירונים או פרספטרונים) אשר מקבלים n כניסות - $\{x_k\}_{k=1}^{n}$, ולהם יציאה אחת, מהצורה:

$$y = \phi\left(c + \sum_{k=1}^{n} w_k x_k \right)$$

כאשר c ו- w_k הם פרמטרי הנוירון הבודד, ו- ϕ היא פונקציה לא ליניארית כלשהי (מקובל להשתמש בפונקציה - $\phi(x) = \sinh(x)$ או דומות לה). רשת נוירונים מטיפוס Multilayered Perceptron נראית כך:



ציור 1.1 - רשת נוירונים

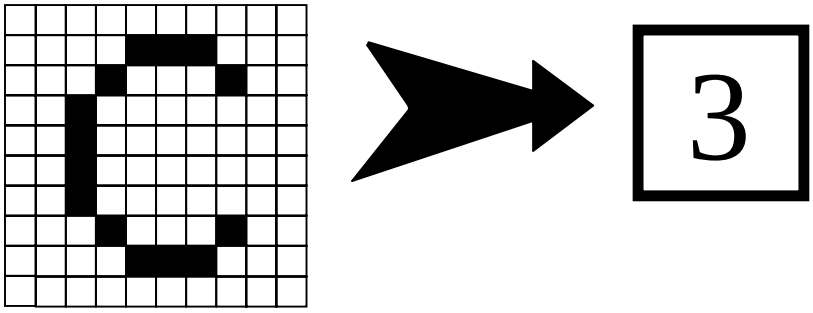
כשמספר הנוירונים, מספר כניסותיהם וקשריהם לשכניהם הינם גורמים הנקבעים באופן כלשהו. נניח כי ברשותנו רשת כזו על קשריה. עבור סידרת כל הפרמטרים המאפיינים את כל הנוירונים ברשת, אותה נסמן P, נקבל קשר פונקציונלי בין כניסה ליציאה של הרשת. קשר זה יהיה מהצורה $y=h(P,X)$, כאשר y הוא מוצא הרשת, ו- X הוא וקטור כניסותיו.

כעת נניח כי רצוננו "ללמד" את הרשת לעשות דבר מה (פעולת הרשת היא יצירת התאמה כלשהי). לדוגמה - זיהוי אותיות - אנו מספקים בכניסותיה קבוצה של 100 פיקסלים ובהם ציור של אות ב- ABC, והמוצא צריך להיות מספרה הסידורי של האות. יש בידנו סידרת אימון הכוללת N

דוגמאות של אותיות ומספרן הסידורי - צמדים אלו נסמן כ- $\left\{X^k, y_k\right\}_{k=1}^{N}$. רצוננו לאמן את הרשת לביצוע משימת זיהוי התוים. לצורך זה נגדיר את מידת ההתאמה הבאה:

$$\varepsilon^2\{P\} = \sum_{k=1}^{N}\left[y_k - h\left(P, X^k\right)\right]^2$$

לפנינו בעיית אופטימיזציה בה הנעלמים הם הפרמטרים P. מציאתם, ומציאת ערך השגיאה הכוללת על סידרת האימון יתנו אמצעי לקבוע האם קונפיגורציית הרשת שנבחרה מתאימה, וכיצד יש להפעיל את הרשת לקבלת זיהוי אותיות. בעיית המינימיזציה שלפנינו היא בעיה במבנה LS בו נדון בהמשך.



<u>ציור 1.2 - כניסה ויציאה מרשת הנוירונים בדוגמה</u>

לסיכום סעיף זה, ראינו מגוון בעיות ממקורות שונים אשר לאחר הגדרת קריטריון טיב הופכות להיות בעיות אופטימיזציה. ככלל, בפתרונן של בעיות רבות במדעים מוצעת גישה של שני שלבים - בשלב הראשון נבנות משוואות לתיאור הבעיה עפ"י הנחות מודל וכדומה, ובשלב השני - ע"י הגדרת קריטריון טיב מומרת הבעיה לבעיית אופטימיזציה.

<u>1.3 חזרה מתמטית - אלגברה ליניארית</u>

בהינתן קבוצת m וקטורים $\left\{V_k\right\}_{k=1}^{m}$ במימד $n(\geq m)$, נאמר כי וקטורים אלו תלויים ליניארית אם קיים סט של סקלרים שונים מאפס $\left\{\alpha_k\right\}_{k=1}^{m}$ אשר נותנים כי:

$$\sum_{k=1}^{m}\alpha_k V_k = \mathbf{0}$$

מכפלה סקלרית של שני וקטורים x ו- y נתונה ע"י $\sum_{k=1}^{n} x_i y_i = x^T y$. וקטורים ייחשבו אורתוגונליים אם מכפלתם הפנימית שווה אפס. נורמה אויקלידית של וקטור אשר מסומנת $\|x\|$ נתונה ע"י שורש של מכפלתו הפנימית בעצמו - דהיינו - שורש על סכום ריבועי איבריו. אי-שוויון קושי-שוורץ קובע כי - $x^T y < \|x\| \cdot \|y\|$.

מטריצה **A** היא מערך דו-מימדי של ערכים סקלריים הקרוים אברי המטריצה. האיבר בשורה ה- i והעמודה ה- j יסומן ע"י $a_{i,j}$. במקרה הכללי מטריצה יכולה להיות מלבנית עם m שורות ו- n עמודות. מטריצה שכל איבריה הם אפס תסומן ע"י **0**. מטריצה ריבועית בה כל איברי האלכסון הראשי 1 והיתר אפס תיקרא מטריצת הזהות ותסומן I.

בהכפלת מטריצה בקבוע, כל איבר בה מוכפל בקבוע זה. חיבור מטריצות אפשרי אם הן באותו גודל, ואז חיבור איברים תואמים יוצר את מטריצת התוצאה. בהכפלת מטריצות זו בזו **C=AB**, קיימת המכפלה אם מספר העמודות ב- **A** שווה למספר השורות ב- **B**. האיבר $c_{i,j}$ הוא הסכימה על פני המכפלות של איברי השורה ה- i ב- **A** בעמודה ה- j ב- **B**, דהיינו:

$$c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}$$

בהפעלת שחלוף (Transpose) על מטריצה, האיבר ה - i,j הופך להיות האיבר ה - j,i. פעולה זו תסומן ע"י $\mathbf{A}^T$. פעולה מורכבת יותר של שחלוף וביצוע צמוד קומפלקסי לכל איבר נקראת שחלוף הרמיטי ותסומן - $\mathbf{A}^H$. חשיבות מיוחדת יש למטריצות סימטריות (אשר חייבות להיות ריבועיות) - אשר מקיימות כי $\mathbf{A} = \mathbf{A}^T$. לכל מטריצה **A** ריבועית ניתן לחשב דטרמיננט אשר יסומן det(**A**). משמעותו הגיאומטרית של גודל זה - נפח המקבילית הרב מימדית המוגדרת משורות (או עמודות) המטריצה. מטריצה אשר לה דטרמיננט שווה לאפס נקראת סינגולרית. פירוש הדבר שאחת משורותיה לפחות תלויה ליניארית באחרות. למטריצה שאינה סינגולרית ניתן להגדיר מטריצה הפוכה אשר תסומן - $\mathbf{A}^{-1}$, ומתקיים כי $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. למטריצה ניתן גם לקבוע מספר הקרוי דרגת המטריצה. ערך זה הינו מספר השורות (או העמודות - זה זהה) במטריצה אשר הינן בלתי תלויות. מטריצה מדרגה מלאה היא מטריצה שכל שורותיה או עמודותיה בלתי תלויות (הקטן בין השניים).

שתי מטריצות ריבועיות בגודל זהה **A** ו- **B** תיחשבנה דומות אם קיימת מטריצה לא סינגולרית **S** כך ש: $\mathbf{B} = \mathbf{S}\mathbf{A}\mathbf{S}^{-1}$. מטריצה ריבועית **U** תיקרא יוניטרית אם שורותיה מהוות בסיס

אורתונורמלי למרחב ה- n מימדי. במקרה זה מתקיים כי $\mathbf{U}^{-1} = \mathbf{U}^{H}$. מטריצות ריבועיות בגודל זהה $\mathbf{A}$ ו- $\mathbf{B}$ תיקראנה דומות יוניטרית אם קיימת מטריצה יוניטרית $\mathbf{U}$ כך ש: $\mathbf{B} = \mathbf{U}\mathbf{A}\mathbf{U}^{H}$.

למטריצה ריבועית מוגדרים ערכים עצמיים λ ווקטורים עצמיים $\mathbf{x}$ ע"י הקשר - $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.
לא תמיד קיימים ערכים עצמיים - דרישה הכרחית ומספיקה לקיומם היא - $\det[\mathbf{A}\mathbf{x} - \lambda\mathbf{x}] = 0$.
קשר זה מניב משוואה פולינומיאלית אשר קרויה המשוואה האופיינית של המטריצה, ושורשיה הם הערכים העצמיים. ליכסון מטריצה מתייחס לפעולת דמיון אשר מניבה מטריצה אלכסונית. עבור מטריצה סימטרית כל הערכים העצמיים ממשיים, והוקטורים העצמיים התואמים להם מהווים בסיס אורתונורמלי (כלומר - כל מטריצה הרמיטית ניתנת לליכסון יוניטרי). כל מטריצה סימטרית יכולה להיות מיוצגת בעזרת ערכיה ווקטוריה העצמיים ע"י הקשר:

$$Au_k = \lambda_k u_k \quad \Rightarrow \quad A = \sum_{k=1}^{n} \lambda_k u_k u_k^T$$

מטריצה סימטרית תיקרא חיובית חצי מוגדרת אם כל ערכיה העצמיים אי שליליים, וחיובית ממש אם כל ערכיה העצמיים חיוביים ממש. למטריצה כזו $\mathbf{A}$ מתקיים כי לכל וקטור $\mathbf{x}$, הגודל $\mathbf{x}^{H}\mathbf{A}\mathbf{x}$ הינו אי שלילי (או חיובי בהתאמה).
נורמה של מטריצה מוגדרת ע"י הפעלתה כאופרטור על כל הוקטורים בעלי אורך יחידה וחישוב נורמת המוצא. המקסימום של ערכים אלו היא נורמת המטריצה. ברישום מתמטי:

$$\|A\| \overset{\Delta}{=} \max_{\|x\|=1} \|Ax\|$$

נורמה זו שווה גם לשורש של הערך העצמי הגדול ביותר של המטריצה $A^{T}A$. עבור מטריצה בעלת נורמה קטנה ממש מ- 1 קיים הקשר הבא:

$$\sum_{k=0}^{L} Q^{k} = \left[I - Q^{L+1}\right][I - Q]^{-1} \xrightarrow[L\to\infty]{} [I - Q]^{-1}$$

בהינתן מרחב ליניארי M אשר הינו תת-מרחב של R^{n} במימד m, ניתן למצוא m וקטורים בלתי תלויים ליניארית אשר יהוו בסיס למרחב זה. המשלים האורתוגונלי של M, אשר יסומן כ- $M^{\perp}$ הוא תת-מרחב בו כל וקטור ניצב לכל וקטור המצוי ב- M.

1.4 חזרה מתמטית - חשבון דיפרנציאלי ואינטגרלי

פונקציה $f(x)$ המקבלת ערכים ממשיים תיקרא רציפה בנקודה $x=\begin{bmatrix}x_1 & \cdots & x_n\end{bmatrix}$ אם לכל סידרת נקודות $x_k \to x$ מתקבל כי $f(x_k)\to f(x)$. באופן שקול ניתן להגדיר רציפות בנקודה x אם לכל $\varepsilon>0$ קיים $\delta>0$ כך שלכל y המקיים $\|x-y\|<\delta$, מתקבל כי $|f(x)-f(y)|<\varepsilon$. תוצאה חשובה המתייחסת לעניין הרציפות היא המשפט של Weierstrass, לפיו לכל פונקציה רציפה מעל תחום קומפקטי קיימות נקודות אקסטרמום בתחום ההגדרה.

הסימון $f(x)\in C^p$ פירושו שנגזרות חלקיות של הפונקציה עד סדר p רציפות ב- x.
לפונקציה ממשית סקלרית $f(x)\in C^1$ נגדיר את וקטור הגרדיאנט כאוסף נגזרותיו החלקיות לפי כל אחד מאברי x, דהיינו:

$$\nabla f(x)=\begin{bmatrix}\dfrac{\partial f(x)}{\partial x_1} & \cdots & \dfrac{\partial f(x)}{\partial x_n}\end{bmatrix}$$

באופן דומה, לפונקציה ממשית סקלרית $f(x)\in C^2$ נגדיר את מטריצת ההסיין כמטריצה המכילה את הנגזרות החלקיות השניות של הפונקציה:

$$\nabla^2 f(x)=\mathbf{F}(x)=\begin{bmatrix}\dfrac{\partial^2 f(x)}{\partial x_1\partial x_1} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_1\partial x_n}\\ \vdots & \ddots & \vdots\\ \dfrac{\partial^2 f(x)}{\partial x_n\partial x_1} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_n\partial x_n}\end{bmatrix}$$

וכיוון ש- $f(x)\in C^2$ סדר הנגזרות לא משנה, ומטריצת ההסיין היא מטריצה סימטרית.

לפונקציה ממשית סקלרית $f(x)\in C^1$ נגדיר את הנגזרת הכיוונית בכיוון d במקום x ע"י הגדרת ישר מהצורה $x(\alpha)=x+\alpha d$ (α הוא סקלר), והגדרת הפונקציה החד מימדית הבאה:

$$g(\alpha)=f(x(\alpha))=f(x+\alpha d)$$

נגזרת כיוונית של $f(x)$ בנקודה x ובכיוון d נתון ע"י:

$$\left.\frac{\partial f(x)}{\partial x}\right|_d = \frac{dg(\alpha)}{d\alpha} = \frac{df(x+\alpha d)}{d\alpha} = \frac{\partial f(x)}{\partial x} d$$

כשעשינו שימוש בכלל השרשרת בנגזרות. התוצאה לעיל הינה למעשה היטלו של הגרדיאנט על וקטור הכיוון d - תוצאה סבירה גם מבחינה אינטואיטיבית. באופן דומה, לפונקציה ממשית סקלרית $f(x) \in C^2$ נגדיר את הנגזרת הכיוונית השניה בכיוון d ע"י אותה בניה של פונקציה חד מימדית, ונקבל כי:

$$\frac{d^2 g(\alpha)}{d\alpha^2} = \frac{d^2 f(x+\alpha d)}{d\alpha^2} = \frac{d}{d\alpha}\frac{\partial f(x+\alpha d)}{\partial x} d = d^T \nabla^2 f(x) d$$

לפי משפט טיילור, בהינתן פונקציה ממשית סקלרית $f(x) \in C^1$ ובהינתן שתי נקודות x ו- y בתחום הגדרת הפונקציה, אזי ניתן לכתוב את שני הקשרים הבאים:

$$f(y) = f(x) + \nabla f(x)(y-x) + o(\|x-y\|)$$
$$f(y) = f(x) + \nabla f(\theta x + (1-\theta)y)(y-x) \quad \theta \in [0,1]$$

שפירושו שהערך בנקודה y ניתן לקירוב ע"י הערך בנקודה x, גורם עדכון הקשור לנגזרת הפונקציה בנקודה x (או בנקודה בין x ל- 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:

$$f(y) = f(x) + \nabla f(x)(y-x) + \frac{1}{2}(x-y)^T \nabla^2 f(x)(x-y) + o\left(\|x-y\|^2\right)$$
$$f(y) = f(x) + \nabla f(x)(y-x) + \frac{1}{2}(x-y)^T \nabla^2 f(\theta x + (1-\theta)y)(x-y)$$

1.5 חזרה מתמטית - תורת הקבוצות

סידרת וקטורים $\{x_k\}_{k=1}^{\infty}$ מתכנסת לוקטור הגבול x אם לכל $\varepsilon > 0$ קיים N גדול דיו כך שלכל $k \geq N$ מתקבל כי $\|x_k - x\| < \varepsilon$. נקודה כלשהי x נקראת נקודת גבול של סידרה $\{x_k\}_{k=1}^{\infty}$ אם בסדרה זו קיימת תת-סידרה אינסופית אשר מתכנסת לנקודה זו.

כדור (sphere) סביב הנקודה הרב-מימדית x נתון ע"י $\{y \mid \|x-y\| < \varepsilon\}$. כדור כזה מוגדר כסביבתה של הנקודה x ברדיוס ε. קבוצה S תיקרא קבוצה פתוחה אם לכל נקודה בקבוצה זו קיים כדור קטן דיו אשר מוכל כולו בקבוצה. באופן אינטואיטיבי נאמר כי קבוצה פתוחה היא קבוצה שאינה כוללת את שפותיה. הפנים של כל קבוצה הוא אוסף הנקודות אשר יכולות לשמש מרכז לכדורים אשר מוכלים בקבוצה - פנים של קבוצה הוא תמיד קבוצה פתוחה.

קבוצה S תיקרא קבוצה סגורה אם היא מכילה את נקודות הגבול של כל סידרת נקודות בקבוצה אשר מתכנסת. באופן מקביל נוכל לומר כי קבוצה סגורה כוללת את שפותיה (גבולותיה). סגור של קבוצה S הוא הוספת נקודות הגבול של כל סידרת נקודות בקבוצה - וזו תמיד קבוצה סגורה שהינה הקבוצה הסגורה הקטנה ביותר אשר מכילה את S.

קבוצה תיקרא קומפקטית אם היא סגורה וחסומה. דהיינו - אם קיים כדור ברדיוס סופי אשר כולה מוכלת בתוכו. משפט נוסף של Weierstrass קובע כי לכל סידרת נקודות אינסופית בקבוצה קומפקטית קיימת תת-סידרה אשר מתכנסת לנקודה בקבוצה זו.

פרק 2 - תנאים הכרחיים ומספיקים לנקודת מינימום

הבעיה בה נטפל במסגרת ההרצאות הקרובות הינה מציאת x אשר יביא למינימום את הפונקציה $f(x)$ - בשלב זה לא נניח כי קיימים אילוצים כלשהם על x. המשתנה x הוא גודל וקטורי במימד n (דהיינו - קיימים n נעלמים). במסגרת קורסים קודמים הכרנו קשרים אשר מתקיימים בין נגזרות הפונקציה $f(x)$ ובין מיקום נקודת המינימום (או מקסימום) עבור פונקציות במשתנה סקלרי. במסגרת הרצאה זו נחדד את הבנתנו את קשרים אלו, ונכלילם למימד הנדון.

2.1 נקודת מינימום לוקלית ונקודת מינימום גלובלית

שאלה ראשונה שצריכה להישאל עבור הבעיה הנ"ל הינה האם בכלל קיים פתרון. מענה לשאלה זו מספק המשפט של Weierstras, לפיו אם f היא פונקציה רציפה ותחום הערכים של x מצוי בתחום קומפקטי, אזי קיימות נקודות מינימום ומקסימום לפונקציה. בהנחה כי ניתן לתחם את ערכי ה- x המותרים לתחום קומפקטי Ω גדול דיו, הרי שממשפט זה נובע כי קיים פתרון. מהמשפט לא נובעת דרך למציאתן של נקודות אלה, ועיקר הקורס יעסוק בבניית כלים לצורך חיפושן. לצורך כך נגדיר שתי פתרונות אפשריים לפי ההגדרות הבאות:

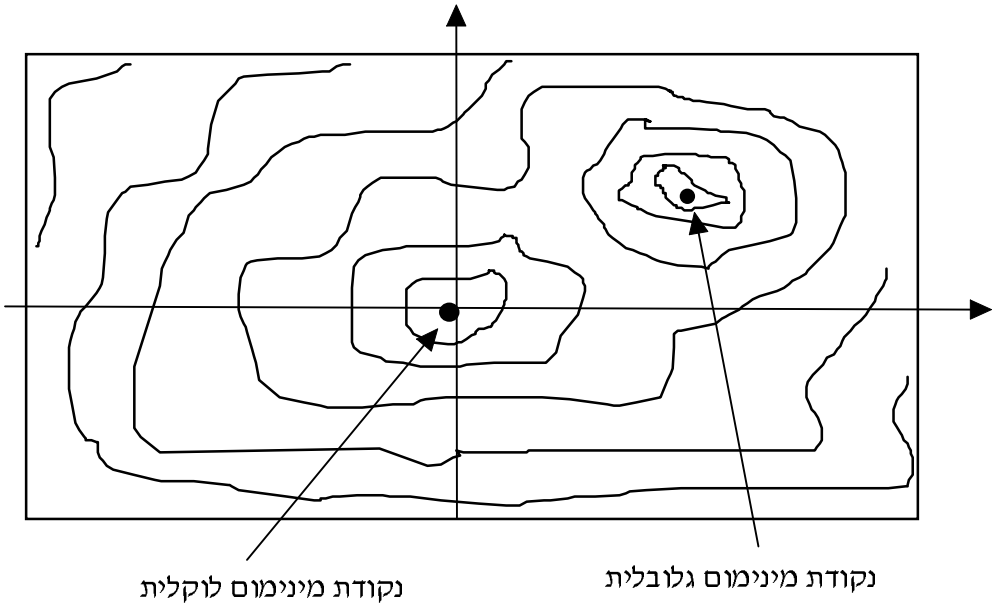
הגדרה 2.1: נקודה כלשהי x^* תיקרא נקודת מינימום לוקלית (או נקודת מינימום יחסית) אם: $\exists \varepsilon > 0 | \quad \forall \left\| x - x^* \right\| < \varepsilon \quad f(x) \geq f\left(x^*\right)$. אם סימן הגדול-שווה מוחלף בסימן גדול, תיקרא נקודה זו נקודת מינימום לוקלית ממש.

הגדרה 2.2: נקודה x^* תיקרא נקודת מינימום גלובלית אם קיים כי $\forall x \quad f(x) \geq f\left(x^*\right)$. אם סימן הגדול-שווה מוחלף בסימן גדול, תיקרא נקודה זו נקודת מינימום גלובלית ממש.

למרות שרצוננו בד"כ במינימום גלובלי, אנו נעסוק במשך כל קורס 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. לכן, בבואנו לפתור את הבעיה אשר הוזכרה בתחילת ההרצאה, גם אם לא ייאמר כך, כוונתנו לאיתור נקודות מינימום לוקליות.

2.2 תנאים הכרחיים מסדר ראשון

בהימצאנו בנקודה כלשהי x, ורצוננו לנוע בכיוון d ממנה, אנו מגדירים ישר פרמטרי מהצורה $x(\alpha) = x + \alpha d$. הפונקציה $f(x(\alpha)) = g(\alpha)$ היא פונקציה בת משתנה יחיד ומתארת חתך של f בכיוון שנבחר.



ציור 2.1 - המחשת נקודות מינימום לוקלית וגלובלית

משפט 2.1: אם הפונקציה f גזירה ברציפות ($f \in C^1$), והנקודה x^* היא נקודת מינימום של f, אזי מתקבל כי $\nabla f(x^*) \equiv \mathbf{0}$.

הוכחה: לכל כיוון d שנבחר מתקבל כי הישר $x(\alpha) = x^* + \alpha d$ מגדיר פונקציה $f(x(\alpha)) = g(\alpha)$ חד מימדית לה נקודת מינימום בערך $\alpha = 0$. פונקציה זו ניתנת לתיאור ע"י משפט טיילור כ:

$$g(\alpha) - g(0) = g'(0)\alpha + o(\alpha)$$

כאשר עבור $\alpha \to 0$, $o(\alpha)$ דועך לאפס יותר מהר מאשר α. אם $g'(0) < 0$, פירוש הדבר שעבור ערכי α מספיק קטנים יתקבל כי $g(\alpha) - g(0) < 0$ אשר סותר את הנתונים. לכן ברור כי:

$$g'(0)=\left.\frac{dg(\alpha)}{d\alpha}\right|_{\alpha=0}=\left.\frac{df\left(x^*+\alpha d\right)}{d\alpha}\right|_{\alpha=0}=\nabla f\left(x^*\right)\cdot d\geq 0$$

ומכיוון שהקשר הנ"ל נכון לכל d, מתקבל כי $\nabla f\left(x^*\right)\equiv \mathbf{0}$. ■

מהמשפט הנ"ל נובע כי בנקודת המינימום x^* מתקיימת מערכת של n משוואות הנתונות ע"י $\nabla f\left(x^*\right)\equiv \mathbf{0}$. עם זאת, בד"כ מערכת המשוואות המתקבלת קשה לפתרון ישיר, ועיקר הקורס עוסק בדרכים עוקפות לפתרון ישיר מעין זה.

דוגמה 2.1: רצוננו למצוא את נקודת המינימום של הפונקציה $f\left(x_1,x_2\right)=x_1^2-x_1x_2+x_2^2-3x_2$. נחשב את גרדיאנט הפונקציה ונקבל:

$$\nabla f\left(x_1,x_2\right)=\begin{bmatrix}\dfrac{\partial f\left(x_1,x_2\right)}{\partial x_1}\\ \dfrac{\partial f\left(x_1,x_2\right)}{\partial x_2}\end{bmatrix}=\begin{bmatrix}2x_1-x_2\\ -x_1+2x_2-3\end{bmatrix}=0$$

מכאן נובע כי הנקודה היחידה אשר יכולה להיות נקודת מינימום היא $x_1=1\ ,\ x_2=2$. כיוון שאיפוס הגרדיאנט הינו תנאי הכרחי אך לא מספיק לנקודת מינימום, אין כל הבטחה כי נקודה זו אמנם נקודת מינימום.

דוגמה 2.2: (קירוב פונקציה). קיימת פונקציה לא ידועה $g(x)$ (x הוא סקלר). פונקציה זו נמדדה ב - m הנקודות $\{x_k\}_{k=1}^m$ והתקבלו הערכים $\{g_k\}_{k=1}^m=\{g(x_k)\}_{k=1}^m$. רצוננו לקרב פונקציה זו ע"י פולינום מסדר n מהצורה - $h(x)=\sum_{k=0}^{n}a_k\cdot x^k$ (כאשר $n<m$). השאלה שעל הפרק היא מהם מקדמי הפולינום המתאימים ביותר?

הפתרון המקובל הוא מציאת המקדמים אשר יביאו למינימום את הביטוי $\varepsilon^2\left(a_0,a_1,\cdots,a_n\right)=\sum_{k=1}^{m}\left[h\left(x_k\right)-g\left(x_k\right)\right]^2$. רישום מלא של הביטוי אותו רצוננו להביא למינימום נותן:

$$\varepsilon^2(a_0,a_1,\cdots,a_n)=\sum_{k=1}^{m}\left[\sum_{j=0}^{n}a_j x_k^j-g_k\right]^2$$

גזירת ביטוי זה לפי a_i נותנת:

$$\frac{\partial}{\partial a_i}\varepsilon^2(a_0,a_1,\cdots,a_n)=\sum_{k=1}^{m}2x_k^i\left[\sum_{j=0}^{n}a_j x_k^j-g_k\right]=0$$

אם נגדיר - $q_{i,j}=\sum_{k=1}^{m}x_k^i\cdot x_k^j$; $b_i=\sum_{k=1}^{m}x_k^i\cdot g_k$ נקבל כי הביטוי הנ"ל ניתן לרישום כ:

$$\sum_{j=0}^{n}q_{i,j}\cdot a_j-b_i=0$$

משוואה זו ניתנת לרישום לכל i בתחום 0 עד n. קיבוץ כל משוואות אלו לרישום מטריצי נותן:

$$\begin{bmatrix} q_{0,0} & q_{0,0} & \cdots & q_{0,n} \\ q_{1,0} & q_{1,1} & \cdots & q_{w,n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n,0} & q_{n,1} & \cdots & q_{n,n} \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}=\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{Qa}=\mathbf{b}$$

כך שבעיית האופטימיזציה שבנינו מסתכמת בפתרון מערכת משוואות ליניארית. יש לשים לב לכך שהפתרון לא הוכח כנקודת מינימום.

דוגמה 2.3: (בקרה). נניח כי נתונה מערכת אשר אמורה לנעול על מטרה. המשתנה x ייצג את שגיאת העקיבה. נניח כי מצבה ההתחלי של המערכת הוא x_0, ורצוננו להביאה תוך m צעדים לשגיאת עקיבה מינימלית (קרובה ככל האפשר ל- 0). לצורך כך אנו מפעילים בכל נקודת זמן k כוח תיקון u_k. ידוע כי הקשר בין התיקון ושגיאת העקיבה נתון ע"י: $x_{k+1}=x_k+u_k$. נתון נוסף הידוע לנו הוא שהפעלת תיקון כרוכה בבזבוז אנרגיה הפרופורציונית לריבוע כוח התיקון.

בעיה זו הוצגה כבר במסגרת הפרק הקודם. קריטריון הטיב לפיו תתוכנן הבקרה הוצע להיות הביטוי:

$$\varepsilon^2(u_1,u_2,\cdots,u_m)=\alpha\cdot\sum_{k=0}^{m}x_k^2+\beta\cdot\sum_{k=0}^{m-1}u_k^2$$

ולאחר הצבת - $x_k = x_{k-1} + u_{k-1} = x_{k-2} + u_{k-2} + u_{k-1} = \cdots = x_0 + \sum_{j=0}^{k-1} u_j$ בתוך הפונקציה שהוגדרה לעיל, מתקבל:

$$\varepsilon^2(u_0, u_1, \cdots, u_{m-1}) = \alpha \cdot \sum_{k=0}^{m-1} \left[x_0 + \sum_{j=0}^{k-1} u_j \right]^2 + \beta \cdot \sum_{k=0}^{m-1} u_k^2$$

גזירה לפי u_i תיתן:

$$\frac{\partial \varepsilon^2}{\partial u_i} = 2\alpha \cdot \sum_{k=i+1}^{m-1} \left[x_0 + \sum_{j=0}^{k-1} u_j \right] + 2\beta u_i = 0$$

$$\Rightarrow \alpha[m-1-i]x_0 + \alpha \sum_{k=i+1}^{m-1} \sum_{j=0}^{k-1} u_j + \beta u_i = 0$$

גזירה זו נתנה שוב משוואה ליניארית במשתנים, וכמקודם נקבל כי בסיכומו של דבר, איפוס הגרדיאנט של הפונקציה שהוגדרה שקול לפתרון מערכת של m משוואות ליניאריות ב- m נעלמים.

2.3 תנאים הכרחיים מסדר שני

את המשפט מהסעיף הקודם קיבלנו ע"י שימוש בפיתוח לטור טיילור עד סדר ראשון. באופן דומה ניתן להציע המשך הפיתוח עד לסדר שני, ולגזור מכך דרישות על הנגזרות השניות של הפונקציה $f(x)$. זוהי הגישה בה ננקוט בסעיף זה.

הגדרה 2.3: מטריצת ההסיין של פונקציה סקלרית $f(x)$ מוגדרת כמטריצת הנגזרות השניות בנקודה x באופן הבא:

$$\mathbf{F}(x) = \nabla^2 f(x) \stackrel{\Delta}{=} \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

הסימון $\mathbf{A}>\mathbf{0}$ כאשר $\mathbf{A}$ מטריצה ריבועית פירושו שהמטריצה חיובית מוגדרת. כאשר הסימון מוחלף בגדול-שווה, פירוש הדבר שהמטריצה חיובית חצי מוגדרת.

משפט 2.2: אם הפונקציה f גזירה פעמיים ברציפות ($f \in C^2$), והנקודה x^* היא נקודת מינימום של f, אזי מתקבל כי $\nabla f(x^*) \equiv \mathbf{0}$, וכן כי $\nabla^2 f(x^*) \geq \mathbf{0}$.

הוכחה: למעשה את הטענה הראשונה כבר הוכחנו במסגרת משפט 2.1. לכל כיוון d שנבחר מתקבל כי הישר $x(\alpha) = x^* + \alpha d$ מגדיר פונקציה $f(x(\alpha)) = g(\alpha)$ חד מימדית לה נקודת מינימום בערך $\alpha = 0$. פונקציה זו ניתנת לתיאור ע"י משפט טיילור כ:

$$g(\alpha) - g(0) = g'(0)\alpha + \frac{1}{2}g''(0)\alpha^2 + o(\alpha^2) = \frac{1}{2}g''(0)\alpha^2 + o(\alpha^2)$$

הגודל $g'(0)\alpha$ הופל כיוון שהוא שווה אפס לפי המשפט הקודם.

אם $g''(0) < 0$, פירוש הדבר שעבור ערכי α מספיק קטנים יתקבל כי $g(\alpha) - g(0) < 0$ אשר סותר את הנתונים. לכן ברור כי:

$$g''(0) = \left.\frac{d^2 g(\alpha)}{(d\alpha)^2}\right|_{\alpha=0} = \left.\frac{d^2 f(x^* + \alpha d)}{(d\alpha)^2}\right|_{\alpha=0} = d^T \nabla^2 f(x^*) \cdot d \geq 0$$

ומכיוון שהקשר הנ"ל נכון לכל d, מתקבל כי $\nabla^2 f(x^*) \geq \mathbf{0}$. ■

קיבלנו אם כך כי בנקודת מינימום של הפונקציה f הגרדיאנט חייב להתאפס, וההסיין חייב להיות חיובי חצי מוגדר. תנאים אלו הינם הכללה של עובדות אותן כבר ידענו בעבר עבור פונקציות של משתנה יחיד - נקודת מינימום מחייבת כי הנגזרת הראשונה מתאפסת והנגזרת השניה חיובית. כאן אנו דורשים רק אי-שליליות של הנגזרת השניה כיוון שמדובר בתנאים הכרחיים (ברור שבמקרה כזה של התאפסות הנגזרת השניה, על הנגזרת השלישית ל-"הצילי" את המצב). התנאים הנ"ל אינם מספיקים לקיומה של נקודת מינימום - הם הכרחיים!

דוגמה 2.4: עבור דוגמה 2.1, נחשב את ההסיין:

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x_1,x_2)}{\partial x_2 \partial x_2} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

לכיוון d שרירותי נקבל כי:

$$\begin{bmatrix} d_1 & d_2 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 2d_1 - d_2 & -d_1 + 2d_2 \end{bmatrix}\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} =$$
$$= 2\left[d_1^2 - d_1 d_2 + d_2^2\right] = \left[(d_1 - d_2)^2 + d_1^2 + d_2^2\right] \geq 0$$

כך שאמנם ההסיין היא מטריצה חיובית מוגדרת. זה נכון לכל נקודה ובפרט לנקודה $x_1 = 1\ ,\ x_2 = 2$. כך שהתנאי החדש המתייחס לנגזרות שניות תומך בקביעת הנקודה הנ"ל כנקודת מינימום - תומך אך לא מבטיח עדיין.

הדרך בה נקטנו על מנת להוכיח כי ההסיין חיובי מוגדר אינה דרך יעילה, וניתן להשתמש בה רק למטריצות מסדר 2. במקרה הכללי, הגישה הנוחה ביותר להראות כי מטריצה היא חיובית מוגדרת היא עיון בערכים העצמיים - אם כולם אי-שליליים אזי המטריצה חיובית חצי מוגדרת. על מנת להוכיח כי מטריצה אינה חיובית חצי מוגדרת ניתן לחשב את הדטרמיננט -אם הוא שלילי - המטריצה אינה חיובית. אלטרנטיבה פשוטה יותר המתאפשרת לפעמים היא שימוש באברי האלכסון הראשי - אם לא כולם חיוביים - המטריצה אינה חיובית.

דוגמה 2.5: עבור דוגמה 2.2 בסעיף הקודם, ניתן להראות ע"י גזירה פשוטה כי המטריצה **Q** היא ההסיין של הפונקציה. אם היא חיובית מוגדרת, אזי פתרון מערכת המשוואות **Qa=b** יכול להיות נקודת מינימום. נראה כי אמנם מטריצה זו חיובית מוגדרת - נגדיר את המטריצה X:

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix}$$

אזי ניתן לראות כי $Q = X^T X$, ולכן Q לפחות חיובית חצי מוגדרת. המטריצה X בעלת מבנה הקרוי מטריצת Vandemonde אשר עבורה ידוע כי אם כל הנקודות $\left[x_k\right]_{k=1}^n$ שונות זו מזו ומספרן קטן ממש מ- m, המטריצה בעלת דרגה מלאה. לכן - Q חיובית מוגדרת ממש, כנדרש.

דוגמה 2.6: 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:

$$\nabla f(x_1,x_2)=\begin{bmatrix}3x_1^2-2x_1x_2\\ -x_1^2+4x_2\end{bmatrix}=\mathbf{0}$$

נקודות אפשריות למינימום הינן לפיכך: $x_1=0\ ,\ x_2=0$ וכן - $x_1=6\ ,\ x_2=9$. ההסיין בנקודות אלו נתון ע"י:

$$\nabla^2 f(x_1,x_2)=\begin{bmatrix}6x_1-2x_2 & -2x_1\\ -2x_1 & 4\end{bmatrix}\Rightarrow\begin{cases}x_1=0\ ,\ x_2=0 \quad\Rightarrow\quad F=\begin{bmatrix}0&0\\0&4\end{bmatrix}\geq 0\\ x_1=6\ ,\ x_2=9 \quad\Rightarrow\quad F=\begin{bmatrix}18&-12\\-12&4\end{bmatrix}\end{cases}$$

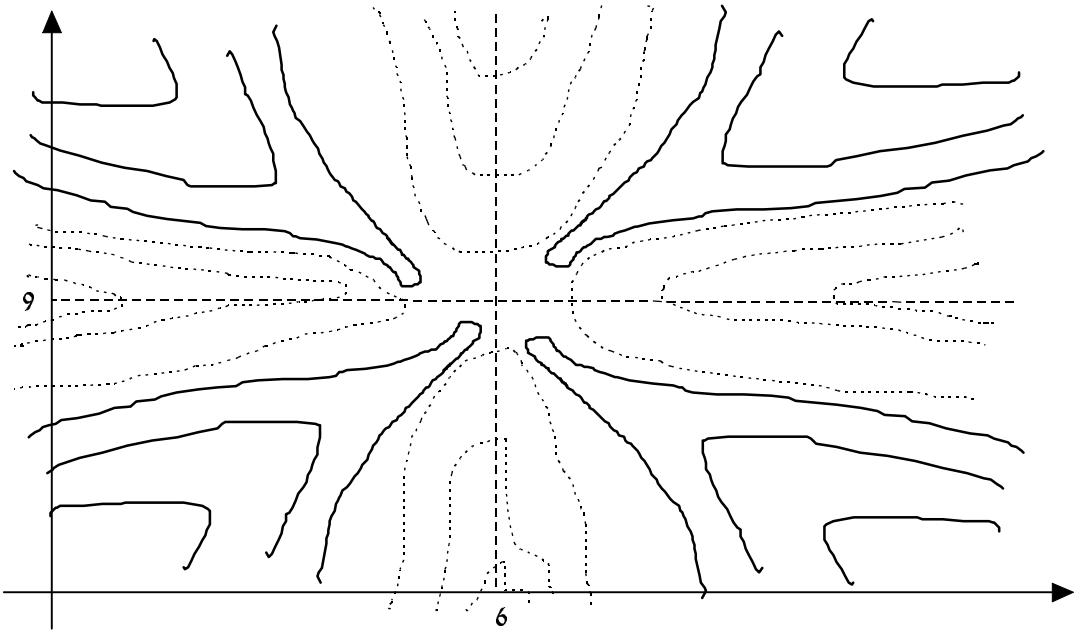
כלומר - הנקודה $x_1=0\ ,\ x_2=0$ יכולה להיות נקודת מינימום כיוון שהגרדיאנט בה מתאפס, וההסיין אי שלילי. הנקודה $x_1=6\ ,\ x_2=9$ היא בודאי לא נקודת מינימום כיוון שדטרמיננט ההסיין בנקודה זו הוא שלילי והערכים העצמיים של מטריצה זו הם 24 ו- (-2)! נקודה זו גם לא נקודת מקסימה כיוון שלצורך זה נדרש שההסיין יהיה שלילי חצי מוגדר (כלומר שמינוס מטריצת ההסיין תהיה חיובית חצי מוגדרת). גם דרישה זו לא מתקיימת כיוון שהכפלת ההסיין בנקודה זו ב- 1- נותנת מטריצה בה אברי האלכסון שליליים!

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-מימדית. עובדות אלו ניכרות מהנגזרות שחושבו עד כה - הגרדיאנט מתאפס בנקודה, ונגזרות שניות בכוונים האנכיים הם 18 ו- 4 בהתאמה - מה שמבטיח (כפי שנראה בהמשך) מינימום. כדי להבין כיצד אפשרי הדבר נתאר את קווי הגובה של הפונקציה בציור הבא. עקומים בקו מלא מתארים גובה גבוה יותר, ועקומים בקו מקווקו מתארים עליה.

2.4 תנאים מספיקים מסדר שני לנקודת מינימום

ע"י הקשחת התנאים שהתקבלו במשפט הקודם נוכל לקבל תנאים מספיקים לקבלתה של נקודת מינימום.

משפט 2.3: תהי הפונקציה f גזירה פעמיים ברציפות ($f\in C^2$), ותהי הנקודה x^* נקודה כלשהי בתחום ההגדרה של הפונקציה f. נניח כי ידוע שבנקודה x^* מתקיים כי $\nabla f(x^*)\equiv\mathbf{0}$, וכן כי $\nabla^2 f(x^*)>\mathbf{0}$. אזי - הנקודה x^* היא נקודת מינימום לוקלית ממש של הפונקציה f.



ציור 2.2 - קווי גובה של הפונקציה

הוכחה: כיוון ש- $\mathbf{F}\left(x^*\right)=\nabla^2 f\left(x^*\right)>\mathbf{0}$, פירוש הדבר הוא שקיים $\beta>0$ כך שלכל וקטור d יתקבל כי $d^T\mathbf{F}\left(x^*\right)d\geq\beta\|d\|^2$. למעשה - ערך של β שיתן קיום של קשר זה הוא הערך העצמי הנמוך ביותר של , $\mathbf{F}$ או כל ערך חיובי קטן מזה - נראה זאת בקצרה. נייצג את $\mathbf{F}(x^*)$ כ - $\mathbf{F}(x^*)=U^H\Delta U$, כאשר U יוניטרית ו- Δ אלכסונית. לכן -

$$d^T\mathbf{F}(x^*)d=d^TU^H\Delta Ud \underset{\substack{\uparrow \\ Ud=d_1}}{=} d_1^H\Delta d_1=\sum_k\lambda_k\cdot d_1^2(k)\geq$$

$$\geq\lambda_{min}\sum_k d_1^2(k)=\lambda_{min}\|d_1\|^2=\lambda_{min}\|d\|^2$$

נוכל לפתח את הפונקציה f סביב הנקודה x^* לטור טיילור מהצורה:

$$f\left(x^*+d\right)=f\left(x^*\right)+\nabla f\left(x^*\right)\cdot d+\frac{1}{2}d^T\mathbf{F}\left(x^*\right)d+o\left(\|d\|^2\right)$$

נשתמש בידיעה שבנקודה x^* מתקיים כי $\nabla f\left(x^*\right)\equiv\mathbf{0}$, וכן כי $\nabla^2 f\left(x^*\right)>\mathbf{0}$, ונקבל כי -

$$f\left(x^*+d\right)-f\left(x^*\right)=\frac{1}{2}d^T\mathbf{F}\left(x^*\right)d+o\left(\|d\|^2\right)\geq\frac{1}{2}\beta\cdot\|d\|^2+o\left(\|d\|^2\right)$$

עבור ערכים קטנים דיים של $\|d\|^2$ נקבל כי הגודל הנ"ל חיובי - דהיינו - לכל d קיבלנו כי הנקודה x^* בעלת ערך נמוך משכנותיה - לכן ברור כי היא נקודת מינימום לוקלית ממש. ■

נשים לב כי ההבדל היחיד בין תנאי משפט זה והתנאים של משפט 2.2 הקודם מתמקד בחיוביות ממש של מטריצת ההסיין - כאשר זה מתקבל - הנקודה היא מינימום ממש. אחרת - על נגזרות מסדר גבוה יותר לקבוע את אופיה של נקודה זו.

דוגמה 2.7: עבור דוגמה 2.4 בסעיף הקודם, ההסיין היה המטריצה $\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}$ אשר הינה חיובית מוגדרת. לכן הנקודה $x_1=1\ ,\ x_2=2$ הינה בודאות נקודת מינימום.

דוגמה 2.8: עבור דוגמה 2.6 מסעיף קודם, הפתרון $x_1=0\ ,\ x_2=0$ נתן את ההסיין $\begin{bmatrix}0 & 0\\ 0 & 4\end{bmatrix}$ אשר אינה חיובית מוגדרת ממש. לכן תנאי המשפט כאן לא מתקיימים ולא ניתן לקבוע עפ"י המידע שבידנו אם נקודה זו הינה נקודת מינימום ממש.

2.5 תנאים הכרחיים ומספיקים לבעיות מעל תחום הגדרה קומפקטי

עד כה התיחסנו לפונקציה המוגדרת מעל כל המרחב הוקטורי של x. כעת נטפל בבעיה מעט יותר מורכבת, ובה מרחב תחום ההגדרה של f(x) הוא הקבוצה הקומפקטית $x\in\Omega$. ההבדל היחיד שאמור להתקבל עבור כל המשפטים והניתוחים הקודמים מתיחס לנקודות השפה של קבוצה זו. בסעיף זה נכליל את שלושת המשפטים 2.1, 2.2 ו- 2.3 לטיפול בבעיה - מציאת נקודת מינימום מקומי לפונקציה f(x) מעל הקבוצה הקומפקטית $x\in\Omega$.

הגדרה 2.4: בהינתן נקודה כלשהי $x\in\Omega$, כיוון חוקי ייקרא וקטור d אשר נותן כי קיים $\varepsilon_0>0$ כך שלכל $0\leq\varepsilon\leq\varepsilon_0$ מתקבל כי $x(\varepsilon)=x+\varepsilon d\in\Omega$.

משמעותה המעשית של הגדרה זו היא שלנקודה שאינה על שפת התחום Ω, כל הכיוונים הם חוקיים (כי קיים כדור קטן דיו אשר יוכל כולו בתוך Ω). לנקודות השפה של התחום Ω, חלק מהכיוונים אינם חוקיים כיוון שהם מובילים אל מחוץ לתחום.

משפט 2.4: (תואם משפט 2.1) אם הפונקציה f גזירה ברציפות ($f \in C^1$), והנקודה x^* היא נקודת מינימום של f מעל Ω, אזי לכל כיוון חוקי d מתקבל כי $\nabla f\left(x^*\right)d \geq 0$.

הוכחה: ניקח את הוכחת משפט 2.1 ונכניס בה מספר שינויים מתחייבים. לכל כיוון חוקי d שנבחר קיים ε_0 כך שמתקבל כי הישר $x(\varepsilon) = x^* + \varepsilon d$ ($0 \leq \varepsilon \leq \varepsilon_0$) מגדיר פונקציה $f\left(x(\varepsilon)\right) = g(\varepsilon)$ חד מימדית לה נקודת מינימום בערך $\varepsilon = 0$. פונקציה זו ניתנת לתיאור ע"י משפט טיילור כ:

$$g(\varepsilon) - g(0) = g'(0)\varepsilon + o(\varepsilon)$$

אם $g'(0) < 0$, פירוש הדבר שעבור ערכי ε מספיק קטנים יתקבל כי $g(\varepsilon) - g(0) < 0$ אשר סותר את הנתונים. לכן ברור כי:

$$g'(0) = \left.\frac{dg(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} = \left.\frac{df\left(x^* + \varepsilon d\right)}{d\varepsilon}\right|_{\varepsilon=0} = \nabla f\left(x^*\right)\cdot d \geq 0$$

אשר מוכיח את המשפט. אם הנקודה x פנימית לתחום Ω נקבל כי כל כיוון הוא חוקי ולכן הקשר הנ"ל נכון לכל d, עובדה אשר גוררת כי מתקבל כי $\nabla f\left(x^*\right) \equiv \mathbf{0}$, כפי שקיבלנו במשפט 2.1. ■

משפט 2.5: (תואם משפט 2.2). אם הפונקציה f גזירה פעמיים ברציפות ($f \in C^2$), הנקודה x^* היא נקודת מינימום של f מעל Ω, ו- d הוא כיוון חוקי כלשהו, אזי מתקבל כי $\nabla f\left(x^*\right)d \geq 0$. אם עבור כיוון זה מתקיים $\nabla f\left(x^*\right)d = 0$ אזי $d^T\nabla^2 f\left(x^*\right)d \geq 0$.

הוכחה: למעשה את הטענה הראשונה כבר הוכחנו במסגרת משפט 2.4. לכל כיוון חוקי d שנבחר קיים ε_0 כך שמתקבל כי הישר $x(\varepsilon) = x^* + \varepsilon d$ ($0 \leq \varepsilon \leq \varepsilon_0$) מגדיר פונקציה $f\left(x(\varepsilon)\right) = g(\varepsilon)$ חד מימדית לה נקודת מינימום בערך $\varepsilon = 0$. פונקציה זו ניתנת לתיאור ע"י משפט טיילור כ:

$$g(\varepsilon) - g(0) = g'(0)\varepsilon + \frac{1}{2}g''(0)\varepsilon^2 + o\left(\varepsilon^2\right) = \frac{1}{2}g''(0)\varepsilon^2 + o\left(\varepsilon^2\right)$$

אם $g''(0) < 0$, פירוש הדבר שעבור ערכי ε מספיק קטנים יתקבל כי $g(\varepsilon) - g(0) < 0$ אשר סותר את הנתונים. לכן ברור כי:

$$g''(0)=\frac{d^2g(\varepsilon)}{(d\varepsilon)^2}\bigg|_{\varepsilon=0}=\frac{d^2f\left(x^*+\varepsilon d\right)}{(d\varepsilon)^2}\bigg|_{\varepsilon=0}=d^T\nabla^2f\left(x^*\right)\cdot d\geq 0$$

אשר מוכיח את המשפט. אם הנקודה x פנימית לתחום Ω נקבל כי כל כיוון הוא חוקי ולכן הקשר הנ"ל נכון לכל d, ומתקבל כי $\nabla^2f\left(x^*\right)\geq \mathbf{0}$, כמו במשפט 2.2. ■

דוגמה 2.9: נתונה הפונקציה $f\left(x_1,x_2\right)=\left(x_1+3\right)^2+\left(x_2+1\right)^2$, ורצוננו למצוא את נקודת המינימום שלה בתחום $\left(x_1,x_2\right)\in[0,1]\times[0,1]$ (ריבוע יחידה). ממבנה הפונקציה ניכר כי נקודת המינימום שלה ללא הגבלה על הקואורדינטות היא (-3,-1), אך נקודה זו מחוץ לתחום ההגדרה בשאלה. גרדיאנט הפונקציה נתון ע"י:

$$\nabla f\left(x_1,x_2\right)=\begin{bmatrix}\dfrac{\partial f}{\partial x_1}\\ \dfrac{\partial f}{\partial x_1}\end{bmatrix}=\begin{bmatrix}2\left(x_1+3\right)\\ 2\left(x_2+1\right)\end{bmatrix}$$

וכפי שרואים, לכל נקודה $\left(x_1,x_2\right)\in[0,1]\times[0,1]$ גרדיאנט זה חיובי בשני איבריו. ההסיין של פונקציה זו נתון ע"י:

$$\nabla^2f\left(x_1,x_2\right)=\begin{bmatrix}\dfrac{\partial^2f\left(x_1,x_2\right)}{\partial x_1\partial x_1} & \dfrac{\partial^2f\left(x_1,x_2\right)}{\partial x_1\partial x_2}\\ \dfrac{\partial^2f\left(x_1,x_2\right)}{\partial x_2\partial x_1} & \dfrac{\partial^2f\left(x_1,x_2\right)}{\partial x_2\partial x_2}\end{bmatrix}=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}>0$$

ומטריצה זו קבועה במקום, וחיובית מוגדרת.

נבחר את הנקודה $x_1=0.5,\ x_2=0.5$ ונבחן האם נקודה זו יכולה להיות נקודת מינימום. עפ"י משפט 2.4, אם קיים כיוון חוקי d אשר ייתן $\nabla f(0.5,0.5)d<0$ אזי ברור כי נקודה זו לא יכולה להיות נקודת מינימום. הגרדיאנט בנקודה זו הוא [7,3]. כיוון שכל כיוון הוא חוקי בנקודה זו, בחירת וקטור הכיוון $[-1,-1]$ יתן כי $\nabla f(0.5,0.5)d<0$, ואמנם נקודה זו לא יכולה להיות נקודת מינימום. למעשה יכולנו לדעת זאת תוך שימוש במשפט 2.1 כיוון שהנקודה שנבחנה היא נקודה פנימית והגרדיאנט בה חייב להתאפס - דבר שאינו קורה!

נבחר את הנקודה $x_1 = 0.5,\ x_2 = 0$ ונבחן האם נקודה זו יכולה להיות נקודת מינימום. זוהי נקודת שפה ועפ"י משפט 2.4, אם קיים כיוון חוקי d אשר ייתן $\nabla f(0.5,0)d < 0$ אזי ברור כי נקודה זו לא יכולה להיות נקודת מינימום. מכפלת הגרדיאנט בוקטור כיוון כללי תיתן את הביטוי -

$$\nabla f(0.5,0)d = \begin{bmatrix} 7 & 2 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = 7d_1 + 2d_2$$

קביעת וקטור הכיוון כ - $\begin{bmatrix} -1 & 1 \end{bmatrix}$ כש-$x(\varepsilon) = [0.5,\ 0] + \varepsilon[-1,\ 1]$ תיתן כי הישר המתקבל מתנועה בכיוון d נותן את הנקודות $x(\varepsilon) = \begin{bmatrix} x_1(\varepsilon) & x_2(\varepsilon) \end{bmatrix} = \begin{bmatrix} 0.5-\varepsilon & \varepsilon \end{bmatrix}$, אשר לכל $\varepsilon \le 0.5$ מצוי בתוך התחום. כלומר - הכיוון שבחרנו הוא חוקי. מתקבל כי:

$$\nabla f(0.5,0)d = \begin{bmatrix} 7 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = -5 < 0$$

כך שגם נקודה זו לא יכולה להיות נקודת מינימום. בחירת הנקודה $x_1 = 0,\ x_2 = 0$ באופן דומה תגלה שלכל כיוון חוקי שנבחר נקבל כי ישנה עליה ($\nabla f\left(x^*\right)d \ge 0$) - הסיבה לכך היא שכדי לקבל גודל שלילי לגורם זה נדרש שאחד הכיוונים בכיוון החוקי יהיה שלילי, ולא ניתן לממש כיוון כזה ללא יציאה מתחום ההגדרה. לכן, בנקודה $x_1 = 0,\ x_2 = 0$ מתקימים התנאים ההכרחיים של משפט 2.5, ולמעשה - של משפט 2.4 (כי לכל כיוון חוקי יש עליה ממש).

משפט 2.6: (תואם משפט 2.3). תהי הפונקציה f גזירה פעמיים ברציפות ($f \in C^2$), ותהי הנקודה x^* נקודה כלשהי בתחום ההגדרה של f אשר סומן ב- Ω. אזי, אם לכל כיוון חוקי d מהנקודה x^* מתקבל אחת משתי התוצאות הבאות:

1. $\nabla f\left(x^*\right)d > 0$, או
2. $\nabla f\left(x^*\right)d = 0$ וכן - $d^T\nabla^2 f\left(x^*\right)d > 0$,

אזי הנקודה x^* היא נקודת מינימום של f בתחום הגדרתה Ω.

הוכחה: נוכיח זאת באופן הבא - נעבור על כל הכיוונים החוקיים ונראה כי כל כיוון כזה נותן בהכרח רק עליה בערך הפונקציה ביחס לערכה בנקודה x^*.

כמקודם, לכל כיוון חוקי d שנבחר קיים ε_0 כך שמתקבל כי הישר $x(\varepsilon)=x^*+\varepsilon d$ $(0\le\varepsilon\le\varepsilon_0)$ מגדיר פונקציה $f(x(\varepsilon))=g(\varepsilon)$ חד מימדית. פונקציה זו ניתנת לתיאור ע"י משפט טיילור כ:

$$g(\varepsilon)-g(0)=g'(0)\varepsilon+o(\varepsilon)=\nabla f(x^*)d\cdot\varepsilon+o(\varepsilon)=f(x(\varepsilon))-f(x^*)$$

אם מתקיימת התכונה הראשונה - $\nabla f(x^*)d>0$ - פירוש הדבר שבכיוון d יש לפונקציה g נקודת מינימום ב- $\varepsilon=0$, דהיינו בחתך בכיוון d של הקואורדינטות מקבלת הפונקציה $f(x)$ נקודת מינימום בנקודה x^*.

אם לעומת זאת נותן הכיוון החוקי d שבחרנו כי $\nabla f(x^*)d=0$ וכן - $d^T\nabla^2 f(x^*)d>0$, אזי פיתוח לטור טיילור כמקודם ייתן:

$$g(\varepsilon)-g(0)=\frac{1}{2}g''(0)\varepsilon^2+o(\varepsilon^2)=f(x(\varepsilon))-f(x^*)$$

וכיוון ש- $d^T\nabla^2 f(x^*)d>0$, פירוש הדבר שעבור ערכי ε מספיק קטנים יתקבל כי שוב קיבלנו כי הנקודה x^* היא נקודת מינימום של החתך.

כיוון שהראנו כי לכל חתך חוקי שנבחר נקבל רק עליה בהתרחקות מהנקודה x^*, ברור כי נקודה זו היא נקודת מינימום יחסית. ■

2.6 סיכום

במסגרת פרק זה ראינו כי ניתן לקשור את נגזרותיה מסדר ראשון ושני של פונקציה כלשהי למיקומה של נקודת מינימום. הבעיה בה התמקדנו היא מינימיזציה של פונקציה ללא אילוצים או עם אילוץ פשוט של תחום הגדרה סופי. ישנה אבחנה בין תנאים הכרחיים אשר נגזרות אלו צריכות לקיים כדי שנקודה תוכל לשמש כנקודת מינימום, ובין תנאים מספיקים אשר די בהם לשם הבטחה כי נקודה נתונה היא נקודת מינימום.

יכולנו לדלות תנאים על הנגזרת השלישית ויותר כדי להכליל עוד ועוד את תוצאות פרק זה (למקרה בו גם הנגזרת הראשונה וגם השניה מתאפסות). בכל אופן, בבעיות מעשיות אין כל עניין

בנגזרות גבוהות מהשניה, ואף נגזרת זו לעיתים מטרידה מספיק כדי להימנע ממנה ולחפש דרכים עוקפות להשגתה.

התוצאות אשר הופקו בפרק זה הינן תאורתיות. אבל, אנו נראה בהמשך כי על-פיהן נוכל להציע אלגוריתמים מעשיים אשר יתכנסו למינימום של פונקציה באופן איטרטיבי.

פרק 3 - פונקציות וקבוצות קמורות

מקום של כבוד שמור בתורת האופטימיזציה לפונקציות קמורות (קונווקסיות), ובמסגרת פרק זה נבין מדוע. מסתבר כי לפונקציות אלה תכונות מאוד נוחות המבטיחות למשל כי נקודת מינימום לוקלית היא גם גלובלית - תכונה חשובה המבטיחה את איתור התוצאה הטובה ביותר. במסגרת פרק זה נדון בפונקציות קמורות - תכונותיהן, והשפעת תכונות אלה על מינימיזציה ומקסימיזציה על הפונקציה הקמורה. למעשה, בשל התכונות שיוצגו במסגרת פרק זה, ניתן לומר כי במקרים רבים בהם נדון באופטימיזציה של פונקציה כללית, נקרב אותה מקומית כפונקציה קמורה.

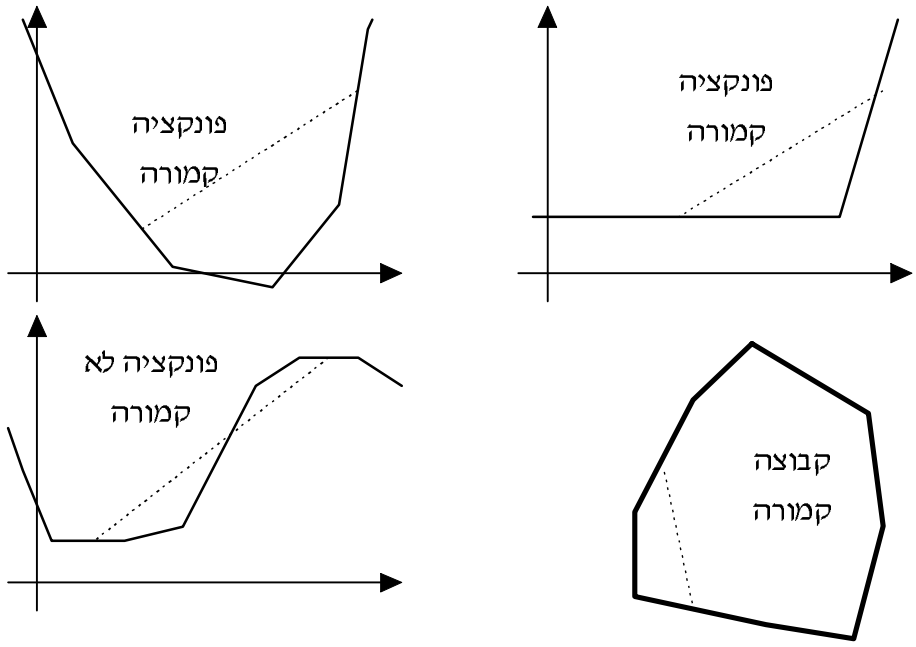
3.1 פונקציות קמורות - הגדרות ותכונות בסיסיות

הגדרה 3.1: קבוצה Ω מוגדרת כקבוצה קמורה אם לכל שתי נקודות $x_1, x_2 \in \Omega$ (בקבוצה) מתקבל כי לכל $0 \le \theta \le 1$, הנקודה $\theta x_1 + (1-\theta)x_2$ אף היא בקבוצה Ω.

הגדרה 3.2: פונקציה $f(x)$ מעל קבוצה קמורה Ω מוגדרת כפונקציה קמורה אם לכל שתי נקודות $x_1, x_2 \in \Omega$ מתקבל כי:

$$f(\theta x_1 + (1-\theta)x_2) \le \theta f(x_1) + (1-\theta) f(x_2)$$

אם סימן הקטן שווה מומר בסימן קטן ממש, תיקרא פונקציה זו פונקציה קמורה ממש.



ציור 3.1 - קבוצות ופונקציות קמורות ושאינן קמורות - דוגמאות

הגדרה 3.3: פונקציה $f(x)$ מעל קבוצה קמורה Ω מוגדרת כפונקציה קעורה אם הפונקציה $[-f(x)]$ היא פונקציה קמורה.

לעיתים קרובות ניתקל במצבים בהם אנו מקבלים צירופים ליניאריים של פונקציות קמורות. אנו נראה כעת כי כל צירוף כזה עם מקדמים חיוביים ייתן פונקציה קמורה.

למה 3.1: תהיינה הפונקציות $f_1(x)$ ו- $f_2(x)$ 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.

הוכחה: ניקח שתי נקודות כלשהן $x_1, x_2 \in \Omega$. נוכל לומר כי לכל $0 \le \theta \le 1$:

$$\begin{aligned} &a_1 f_1(\theta x_1 + (1-\theta)x_2) + a_2 f_2(\theta x_1 + (1-\theta)x_2) \le \\ &\quad \le a_1\left[\theta f_1(x_1) + (1-\theta) f_1(x_2)\right] + a_2\left[\theta f_2(x_1) + (1-\theta) f_2(x_2)\right] \end{aligned}$$

ולאחר קיבוץ איברים וסידור נקבל:

$$\begin{aligned} &a_1 f_1(\theta x_1 + (1-\theta)x_2) + a_2 f_2(\theta x_1 + (1-\theta)x_2) \le \\ &\quad \le \theta\left[a_1 f_1(x_1) + a_2 f_2(x_1)\right] + (1-\theta)\left[a_1 f_1(x_2) + a_2 f_2(x_2)\right] \end{aligned}$$

אשר מוכיח את הטענה. ■

עד כה עיננו בקבוצות קמורות ובפונקציות קמורות. כעת נמחיש קשר ישיר ביניהן, המתואר בלמה הבאה:

למה 3.2: תהי נתונה פונקציה קמורה $f(x)$ מעל קבוצה קמורה Ω. אזי הקבוצה הנתונה ע"י הביטוי $\Gamma = \{x | f(x) \le c\}$ היא קבוצה קמורה.

הוכחה: ניקח שתי נקודות כלשהן $x_1, x_2 \in \Gamma$. בשל השתייכותן של נקודות אלו לקבוצה Γ ברור כי מתקיים $f(x_1) \le c$, $f(x_2) \le c$. נוכל לומר כי לכל $0 \le \theta \le 1$:

$$f(\theta x_1 + (1-\theta)x_2) \le \theta f(x_1) + (1-\theta) f(x_2) \le c$$

לכן, גם כל נקודה על הישר המחבר שתי נקודות אלה יתקבל כי ערך הפונקציה קטן מ- c ולכן כל ישר זה מוכל בקבוצה Γ. באופן זה הראנו שלכל שתי נקודות ב- Γ, כל הישר ביניהן מוכל אף-הוא בקבוצה - הקבוצה קמורה. ■

3.2 תכונות של פונקציות קמורות גזירות

כדי לחבר את תוצאות הפרק הקודם לפרק זה, נראה כעת כי לפונקציות קמורות תכונות הנוגעות לנגזרותיהן. עפ"י תכונות אלה נוכל לדון לאחר מכן במיקומן של נקודות מינימום ומקסימום בפונקציות כאלה.

למה 3.3: תהי $f(x) \in C^1$. אזי פונקציה זו קמורה מעל תחום הגדרה קמור Ω אם ורק אם מתקיים כי לכל $x, y \in \Omega$:

$$f(y) \geq f(x) + \nabla f(x)(y - x)$$

הוכחה: נניח כי פונקציה זו קמורה. אזי, נוכל לכתוב עבור כל $0 < \alpha \leq 1$ מתקיים:

$$f(\alpha y + (1-\alpha)x) \leq \alpha f(y) + (1-\alpha) f(x)$$

ע"י סידור איברים בצורה שונה נקבל:

$$\frac{f(\alpha y + (1-\alpha)x) - f(x)}{\alpha} \leq f(y) - f(x)$$

ועם השאפת $\alpha \to 0$ מתקבלת טענת הלמה כיוון ש:

$$\lim_{\alpha \to 0} \frac{f(x + \alpha(y-x)) - f(x)}{\alpha} = \nabla f^T(x)(y-x) \leq f(y) - f(x)$$

באופן זה הוכחנו כיוון אחד של הלמה. כדי להוכיח את הכיוון ההפוך נניח כי הקשר $f(y) \geq f(x) + \nabla f(x)(y-x)$ נכון, ונראה כי נובע ממנו שהפוקציה קמורה.

תהיינה $x_1, x_2 \in \Omega$ שתי נקודות כלשהן בתחום ההגדרה, ויהיה α סקלר בעל ערך שרירותי באינטרוול $[0,1]$. קביעת הנקודה $x = \alpha x_1 + (1-\alpha)x_2$ ושימוש באי-השיוויון אשר הונח כנכון - $f(y) \geq f(x) + \nabla f(x)(y-x)$ נותן כי:

$$f(x_1) \geq f(x) + \nabla f(x)(x_1 - x)$$
$$f(x_2) \geq f(x) + \nabla f(x)(x_2 - x)$$

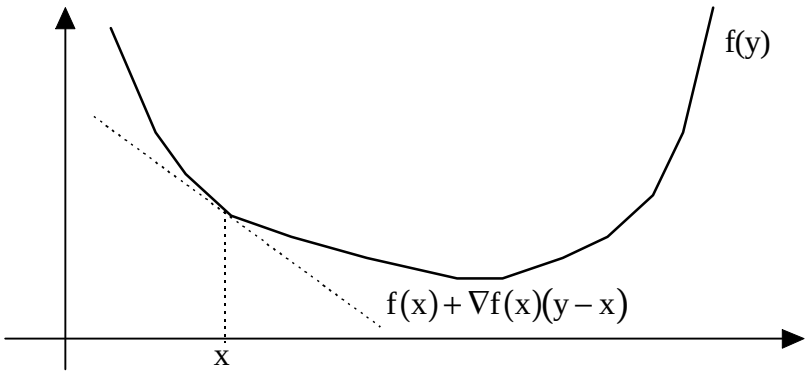
הכפלת המשוואה העליונה ב- α והשניה ב- $(1-\alpha)$ וסיכומן ייתן:

$$\alpha f(x_1) + (1-\alpha) f(x_2) \geq f(x) + \alpha \nabla f(x)(x_1 - x) + (1-\alpha)\nabla f(x)(x_2 - x)$$

$$\Rightarrow \alpha f(x_1) + (1-\alpha) f(x_2) \geq f(x) + \alpha \nabla f(x)(\alpha x_1 + (1-\alpha)x_2 - x)$$
$$\Rightarrow \alpha f(x_1) + (1-\alpha) f(x_2) \geq f(x)$$

וקיבלנו את תכונת היסוד של פונקציה קמורה, ולכן הלמה הוכחה. ■

בציור 3.2 מתוארת טענת למה זו באופן גרפי. אם קודם קבענו כי אינטרפולציה בין שתי נקודות מניבה ישר המצוי מעל הפונקציה, כאן אנו רואים כי מתיחת קו המקרב מקומית את הפונקציה ע"י ישר משיק מניבה ישר אשר תמיד מצוי מתחת לפונקציה.



<u>ציור 3.2 - המחשת למה 3.3 באופן גרפי</u>

למה 3.4: תהי $f(x) \in C^2$. אזי פונקציה זו קמורה ממש מעל תחום הגדרה קמור Ω אם ורק אם מתקיים כי ההסיין של פונקציה זו - $\mathbf{F}(x)$ - חיובי מוגדר בכל נקודה בתחום ההגדרה.

הוכחה: ע"י נוסחת טיילור נוכל לכתוב כי לכל y מתקיים:

$$f(y) = f(x) + \nabla f(x)(y-x) + \frac{1}{2}(y-x)^T \mathbf{F}(x)(y-x) + o\left\{\|x-y\|^2\right\}$$

אנו נתעניין בנקודות y המצויות על שפתו של כדור n מימדי שמרכזו בנקודה x ורדיוסו ε כה קטן עד כי מתקבל כי:

$$\left|\frac{1}{2}(y-x)^T\mathbf{F}(x)(y-x)\right| >> \left|o\left\{\|x-y\|^2\right\}\right|$$

אם אמנם ההסיין חיובי מוגדר בכל נקודה נתקבל כי $\frac{1}{2}(y-x)^T\mathbf{F}(x)(y-x)>0$. הגודל $o\left\{\|x-y\|^2\right\}$ יכול להיות שלילי או חיובי, אך בכל מקרה קטן משמעותית (בשל דרך בחירת ה- y) ביחס לגודל הביטוי - $\frac{1}{2}(y-x)^T\mathbf{F}(x)(y-x)>0$. לכן מתקבל - $f(y)>f(x)+\nabla f(x)(y-x)$ וכבר ראינו מלמה קודמת כי פירוש הדבר שהפונקציה קמורה ממש. בדרך זו הוכחנו כיוון אחד של הלמה.

לצורך הוכחת הכיוון ההפוך נניח כי הפונקציה הנתונה קמורה ממש, וקיימת נקודה x בה ההסיין אינו חיובי חצי מוגדר. בשל קמירותה ממש של הפונקציה מתקיים כי $f(y)>f(x)+\nabla f(x)(y-x)$. מהנחותינו נובע גם שקיימת נקודה y כלשהי אשר מצד אחד נותנת כי $(y-x)^T\mathbf{F}(x)(y-x)<0$, ומצד שני ע"י שימוש בנוסחאת טיילור שנרשמה בתחילת ההוכחה (וע"י בחירת y מספיק קרוב ל-x) תיתן - $\frac{1}{2}(y-x)^T\mathbf{F}(x)(y-x)+o\left\{\|x-y\|^2\right\}>0$. תוצאה זו סותרת את הגדרת הביטוי $o\left\{\|x-y\|^2\right\}$, כשואף לאפס מהר יותר מריבוע המרחק, כאשר מרחק זה קטן לאפס. באופן זה הוכחנו גם את הכיוון השני של הלמה. ■

3.3 מינימיזציה ומקסימיזציה של פונקציות קמורות

בסעיף זה נציג שלושה משפטים קלאסיים המתיחסים לבעיות אופטימיזציה על פונקציות קמורות, ואשר עושים שימוש בתכונות שפותחו בסעיפים קודמים.

משפט 3.1: תהי נתונה פונקציה קמורה f(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.

הוכחה: נניח כי ערך המינימום של הפונקציה f הוא c_0. אזי נוכל לכתוב את תת-קבוצת נקודות המינימום ע"י הביטוי: $\Gamma=\left\{y\middle|f(y)\le c_0\right\}$, ולפי למה 3.2 כבר ראינו כי זוהי קבוצה קמורה.

כעת נניח כי הנקודה x_1 היא נקודת מינימום לוקלית, וקיימת נקודה אחרת - x_2 - בה הפונקציה משיגה את המינימום הגלובלי שלה. פירוש הנחה זו הוא ש - $f(x_2) < f(x_1)$. בשל היות הפונקציה f קמורה נוכל לכתוב עבור כל $\alpha \in (0,1)$ כי:

$$f(x_1 + \alpha(x_2 - x_1)) \le (1-\alpha)f(x_1) + \alpha f(x_2) \le$$
$$\le f(x_1) + \alpha[f(x_2) - f(x_1)] < \max\{f(x_1), f(x_2)\} \le f(x_1)$$

פירוש הדבר שעבור α קרוב מאוד כרצוננו לאפס יש לפונקציה ערכים נמוכים מערכה בנקודת המינימום המקומי x_1, וזה עומד בסתירה להיותה נקודת מינימום לוקלי. לכן - ברור כי נקודה x_1 היא נקודת מינימום גלובלי. ■

מהמשפט הנ"ל נובע כי כל נקודות המינימום של פונקציה קמורה מרוכזות יחדיו, והן גם נקודת מינימום גלובליות.

משפט 3.2 : תהי $f(x) \in C^1$ פונקציה קמורה מעל תחום הגדרה קמור Ω. אם קיימת נקודה $x^* \in \Omega$ כך שלכל $y \in \Omega$ יתקיים כי $\nabla f(x^*)(y - x^*) \ge 0$, אזי נקודה זו היא נקודת מינימום גלובלית של הפונקציה בתחום ההגדרה.

הוכחה: למעשה, ע"י שימוש בלמה 3.3 נקבל מיידית כי:

$$f(y) \ge f(x^*) + \nabla f(x^*)(y - x^*) \ge f(x^*)$$

כלומר, לכל y ישנה רק עליה או שמירה על אותו ערך בפונקציה, ולכן ברור כי זו נקודת מינימום לוקלית. כיוון שלפי משפט קודם כל נקודת מינימום לוקלית היא גם נקודת מינימום גלובלית הרי שמשפט זה הוכח. ■

למעשה, ניתן לראות את נכונותו של משפט זה גם ע"י שימוש במשפט 2.4, כיוון שבחירת y מגדירה וקטור כיוון כלשהו d, ולמעשה משפט זה דורש שלכל כיוון חוקי יתקבל $\nabla f(x^*)d \ge 0$.

עד כה דיברנו על מינימיזציה של פונקציה קמורה. כעת נדון במקסימיזציה של פונקציה קמורה, ונראה כי פתרון בעיה זו מוגשם על שפת תחום ההגדרה ובנקודת קצה של שפה זו. אך עלינו להגדיר תחילה מהי נקודת קצה.

הגדרה 3.4: נקודה x בקבוצה קמורה Ω מוגדרת כנקודת קצה אם לא קיימות שתי נקודות $y_1, y_2 \in \Omega$ שונות מ- x כך שהנקודה x מצויה על הישר אשר מתוח ביניהן.

מההגדרה הנ"ל ברור כי כל נקודה המצויה בתוך קבוצה קמורה קומפקטית (כלומר שיש כדור קטן דיו סביבה, אשר מוכל כולו בקבוצה) אינה נקודת קצה. נקודות קצה יכולות להתקבל רק על שפתה של הקבוצה הקומפקטית.

משפט 3.3: תהי נתונה פונקציה קמורה $f(x)$ מעל קבוצה קמורה וקומפקטית Ω. אזי, אם לפונקציה זו יש מקסימום ממש בתחום הגדרתה, נקודת המקסימום תתקבל על נקודת קצה של Ω.

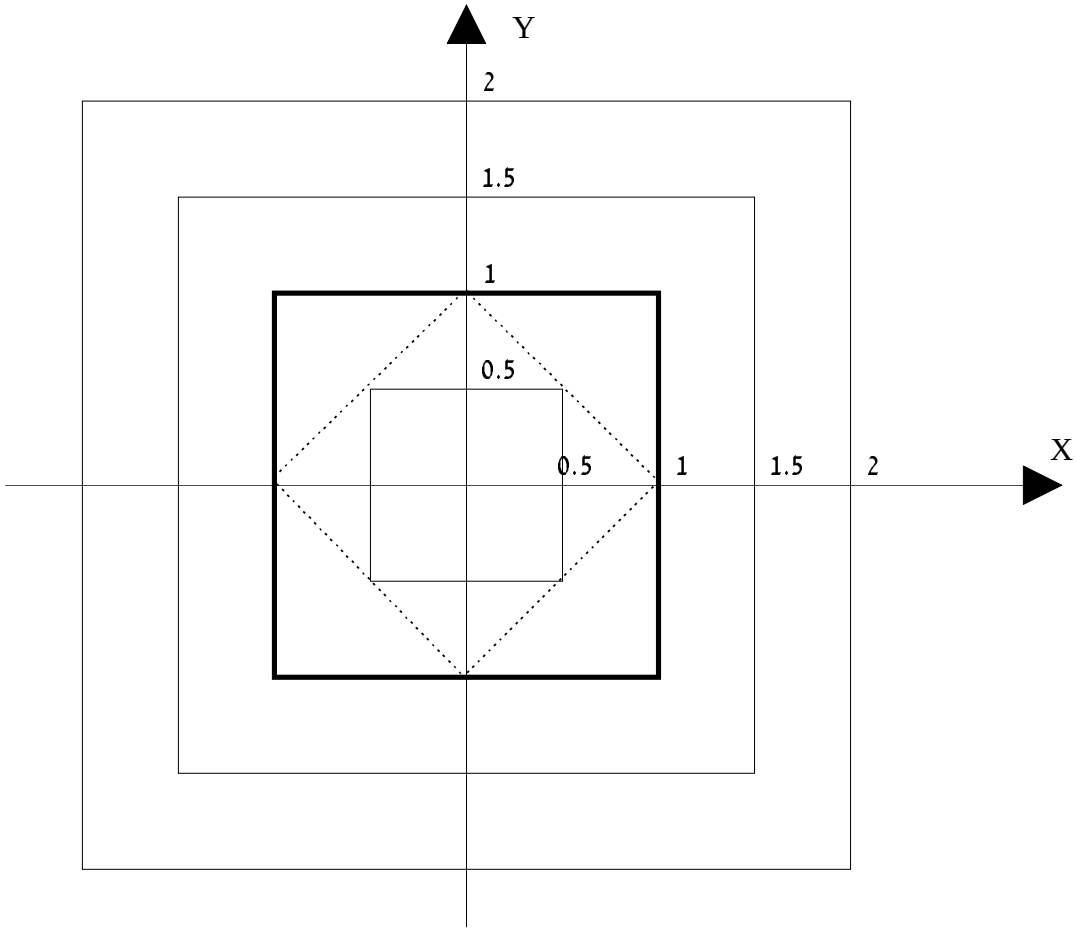
הוכחה: נניח כי הפונקציה מקבלת מקסימום בנקודה $x^* \in \Omega$. נניח כי x^* אינה על בנקודת קצה של תחום ההגדרה. אזי נוכל למתוח קו כלשהו המתחיל בנקודה y_1, חולף דרך הנקודה x^*, ומגיע לנקודת יעד y_2, כאשר $y_1, y_2 \in \Omega$. הנקודה x^* יכולה להיכתב כ- $x^* = \alpha y_1 + (1-\alpha) y_2$, ובשל היות הפונקציה קמורה נקבל:

$$f(x^*) = f(\alpha y_1 + (1-\alpha) y_2) \le \alpha f(y_1) + (1-\alpha) f(y_2) \le \max\{f(y_1), f(y_2)\}$$

וזה עומד בסתירה לכך שנקודה x^* היא נקודת מקסימום ממש. מכאן שאמנם יוגשם המקסימום בנקודת קצה של תחום ההגדרה. ■

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 דוגמה:

דוגמה 3.1: הפונקציה $f(x,y) = |x+y| + |x-y|$ נראית כמתואר בציור 3.3 ע"י קוים שוי גובה. נניח כי פונקציה זו מוגדרת מעל התחום - $\Omega = \{(x,y) \mid -1 \le x, y \le 1\}$. קל לראות כי שפתה של הקבוצה Ω הוא בדיוק קו הגובה המקסימלי של הפונרציה בתחום ההגדרה, ואמנם, המקסימום במקרה זה מתקבל על נקודות שפה ולא רק בקצוות. אם לעומת זאת נחליף את Ω להיות ריבוע שאורך צלעו 1.4142 המוטה בזוית של 45 מעלות ומסביב הראשית (אף זו קבוצה קמורה), נקבל כי נקודות המקסימום תהיינה ארבע נקודות הקצה של תחום ההגדרה.



ציור 3.3 - המחשת האפשרות שפונקציה תקבל את המקסימום שלה בנקודות שאינן נקודות קצה

3.4 סיכום

בפרק זה עסקנו בפונקציות וקבוצות קמורות, הכרנו את תכונותיהן, וראינו כיצד תכונות אלה משפיעות על מיקום נקודות מינימום ומקסימום של הפונקציה, והקשר בין נקודות מינימום מקומיות וגלובליות. כפי שנאמר כבר בתחילת פרק זה, בשל תכונותיהן הטובות של פונקציות אלו, אנו נניח בניתוח אלגוריתמים שיפותחו בהמשך התנהגות קמורה מקומית כל פונקציה בה נטפל, ובאופן זה נוכל להבטיח התכנסות לנקודת מינימום מקומי.

פרק 4 - עקרונות של אלגוריתמים למינימיזציה של פונקציה

בשלב זה אנו מוכנים להתחיל את הדיון המעשי על בנייתם של אלגוריתמים למינימיזציה של פונקציה. ישנו מגוון רחב של אלגוריתמים לפתרון בעיות אופטימיזציה, עם מגוון עקרונות, אך למרביתם (אולי אף כולם!?) בסיס משותף עליו נתעכב במסגרת פרק זה. מרבית האלגוריתמים הנומריים אשר מוצעים למינימיזציה של פונקציה פועלים באופן **איטרטיבי ויורד**. פרק זה יעסוק בעקרונותיהם של אלגוריתמים בעלי מבנה זה.

4.1 מבנה עקרוני של אלגוריתם נומרי למינימיזציה של פונקציה

המושג **איטרטיבי** מתייחס לכך שהאלגוריתם מתחיל מניחוש ראשוני כלשהו של הפתרון, ובסדרה של צעדים הוא משפר בכל פעם את הפתרון הקודם ע"י הפעלת כלי חישובי כלשהו עליו, והפקת פתרון עכשווי טוב יותר. המושג **יורד** מתייחס לעובדה שבכל צעד (איטרציה) של האלגוריתם גורם לירידה בערכה של פונקציה כלשהי המהווה אומדן לקירבה למינימום. באופן זה יורדת פונקציה זו עוד ועוד ומבטיחה התקרבות לפתרון הרצוי.

ההבדל המהותי בין האלגוריתמים השונים (והסיבה לשפע האלגוריתמים המוצעים) מרוכז בדרך בה בוחר האלגוריתם לבצע את עדכון הפתרון בכל איטרציה, ובפונקציה אותה מקטין האלגוריתמים עם התקדמותו. שינויים אלו יוצרים שינויים בקצב ההתכנסות, במספר האיטרציות הדרוש להשגת הפתרון, בכמות החישובים הדרושה וכו'.

הזכרנו קודם כי כל אלגוריתם איטרטיבי מתחיל בנקודת איתחול שרירותית כלשהי אשר מייצגת ניחוש ראשוני של הפתרון. בהתייחס לכך נגדיר את המושג הבא:

הגדרה 4.1: אם אלגוריתם כלשהו A מובטח להתכנס לנקודת פתרון (מינימום לוקלי) לכל נקודת איתחול שרירותית אשר תיבחר, אזי האלגוריתם A מוגדר כבעל תכונת התכנסות גלובלית.

נדמה כי תכונה זו טריוויאלית ואם היא מתקיימת לאיתחול אחד, היא צריכה להתקיים לאיתחולים אחרים, אך במפורש אין זה כך, כפי שנראה בהמשך. האלטרנטיבה לתכונת ההתכנסות הגלובלית היא מצב בו סידרת הנקודות כלל לא מתכנסת (למשל מגיעה למצב של מחזוריות או כאוס), או מתכנסת לנקודה שאינה פתרון.

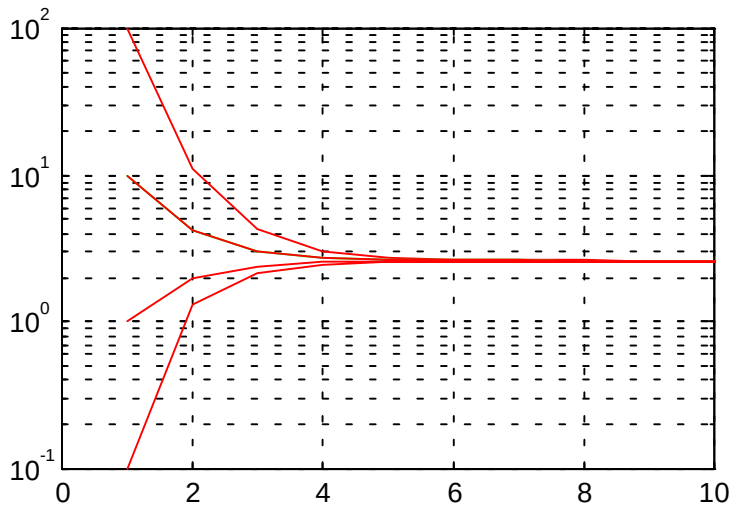
אלגוריתם A יהיה עבורנו מיפוי המקבל נקודה x ומניב נקודה y. באופן זה, כשהאלגוריתם מופעל כל פעם על מוצא השלב הקודם אנו מקבלים סידרת נקודות מהצורה $x_{k+1} = A\{x_k\}$. משוואה זו מניחה כי קביעת הנקודה הבאה נעשית על-סמך הנקודה הנוכחית בלבד - דהיינו - זכרון של נקודה אחת לאחור. ניתן באופן דומה להציע אלגוריתמים הקובעים את הנקודה הבאה על סמך

K נקודות אחרונות - זכרון מסדר K, כשברור כי הגדלת הזכרון מגדילה את יכולתו של האלגוריתם אך גם את סיבוכיותו. במהלך קורס זה נכיר אלגוריתמים בעלי זכרון מסדר אחד או שניים לכל היותר. עפ"י תיאור זה נציע את ההגדרה הבאה:

הגדרה 4.2: אלגוריתם A מוגדר כמיפוי ממרחב הנעלמים של בעיית האופטימיזציה $x \in \Omega$, ולתוכו, דהיינו - יוצר נקודה במרחב זה.

הגדרה זו שונה במקצת מזו אשר נוסחה בספרו של Luenberger - שם הוא בחר להגדיר אלגוריתם כמיפוי של נקודה לקבוצה - ובכך להקל על עצמו ניתוח של אלגוריתמים בהמשך הספר. אנו לא נידרש לכך, ולכן בחרתי לשנות מעט את ההגדרה.

דוגמה 4.1: מוצע האלגוריתם A הבא - $A\{x\} = \sqrt{x} + 1$ אשר נועד למציאת נקודת מינימום של הפונקציה $f(x) = |x - 2.618|^4$ (נניח כי הפונקציה מעל המדומים). באיתחול אלגוריתם זה בערכים $x_0 = 0.1,\ 1,\ 10,\ 100$ תחול התכנסות לערך 2.168 מהצורה המתוארת בציור הבא:



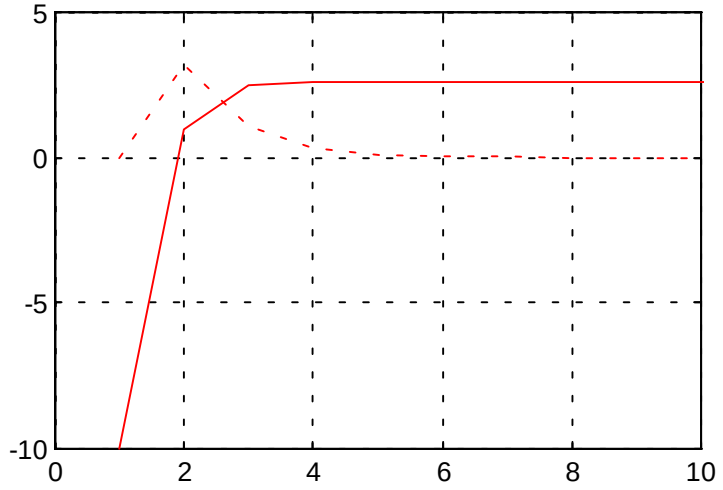
ציור 4.1 - תיאור x_k עבור איתחולים חיוביים שונים בדוגמה 4.1

נדמה כי איתחול בערך שלילי ייתן תוצאה שונה. להלן גרף של הערך הממשי והערך המדומה (שהרי התוצאה קומפלקסית במקרה זה) של ההתכנסות עבור איתחול ב- 10-. כפי שרואים, שוב חלה התכנסות ולאותו ערך. נוכל לומר (ללא הוכחה) כי אלגוריתם זה הינו בעל תכונת התכנסות גלובלית, כיוון שמכל נקודה הוא מתכנס ליעד הנדרש.

דוגמה 4.2: עבור:

$$f(x) = |x| \quad ; \quad A\{x\} = \begin{cases} 0.5(x-1)+1 & x > 1 \\ 0.5x & x \le 1 \end{cases}$$

ברור כי היעד הוא מציאת $x_\infty = 0$. עבור איתחול בערכים מעל 1 נקבל כי האלגוריתם מתכנס ל- 1. עבור איתחול קטן מ- 1 האלגוריתם יתכנס ל- 0 כנדרש. זהו אלגוריתם שאינו בעל תכונת התכנסות גלובלית.



ציור 4.2 - תיאור החלקים הממשי והמדומה של x_k עבור איתחול שלילי בדוגמה 4.1

<u>קו מלא - החלק הממשי, קו מרוסק - החלק המדומה</u>

4.2 פונקצית ירידה

הזכרנו כי אלגוריתם A אמור להביא ירידה בערכה של פונקציה כלשהי בכל פעם בה הוא מופעל. נסמן את קבוצת הפתרונות כ- $\Gamma \subseteq \Omega$. הרעיון הבסיסי של פונקצית ירידה הוא שכל נקודה מחוץ לקבוצת הפתרונות תמופה לנקודה חדשה באופן שמניב ירידה ממש בפונקציה. להלן הגדרה מדויקת:

הגדרה 4.3: תהי $\Gamma \subseteq \Omega$ קבוצת הפתרונות של בעיית אופטימיזציה כלשהי, ויהיה A אלגוריתם אשר הוצע לפתרונה. פונקציה ממשית ורציפה Z מעל Ω תוגדר כפונקצית ירידה אם היא מקיימת כי:

1. אם $x \notin \Gamma$ אזי $Z(A\{x\}) < Z(x)$
2. אם $x \in \Gamma$ אזי $Z(A\{x\}) \le Z(x)$

ניתן כמובן להגדיר באופן ישיר את פונקצית הירידה של בעיה כפונקציה $f(x)$ עצמה, ואת קבוצת הפתרונות כאוסף נקודות המינימום המקומיות, אך ניתן גם להגדיר את $\|\nabla f(x)\|$ כפונקצית הירידה, ואת אוסף הנקודות שמאפסות את הגרדיאנט כקבוצת הפתרונות.

4.3 מיפויים סגורים

הזכרנו קודם כי האלגוריתם A הוא מיפוי. נגדיר כעת את המושג מיפוי סגור, אשר בו נעשה שימוש בהמשך.

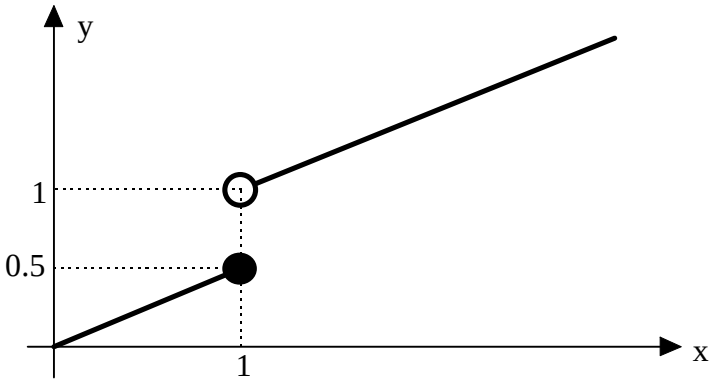
הגדרה 4.4: מיפוי A מנקודה לנקודה בקבוצה Ω קרוי סגור אם מתוך ההנחות כי $x_k \xrightarrow[k\to\infty]{} x$ ו- $y_k \equiv A\{x_k\} \xrightarrow[k\to\infty]{} y$ משתמע כי $y = A\{x\}$

הגדרה זו דומה מאוד לרציפות של אופרטור A אך שונה, ובאמת קיים קשר ישיר שני המושגים - מיפוי רציף הוא סגור (אך לא להפך), כך שרציפות היא דרישה חזקה יותר. ניתן להמיר את תכונת המיפוי הסגור במושג קל יותר להבנה - קבוצה סגורה. זאת ע"י תיאור המיפוי כגרף על מערכת הצירים x,y - אם הקבוצה $\{(x,y)|x\in\Omega,\ y = A\{x\}\}$ סגורה, אזי המיפוי סגור. כשמשרשרים מיפויים סגורים מעל תחום קומפקטי, או כאשר משרשרים מיפוי רציף A עם מיפוי סגור B (דהיינו- C=BA) מתקבל מיפוי כולל סגור (דורש הוכחה שלא תובא כאן).

דוגמה 4.3: בדוגמה 4.2 הוצע מיפוי שאינו סגור. למעשה מיפוי זה אף אינו רציף כיוון שמתקיים כי:

$$A\{x\} = \begin{cases} 0.5(x-1)+1 & x>1 \\ 0.5x & x\le 1 \end{cases} \Rightarrow \begin{array}{l} A\{1+\varepsilon\} = 1+0.5\varepsilon \\ A\{1-\varepsilon\} = 0.5-0.5\varepsilon \end{array}$$

תיאור גרפי של הקבוצה $\{(x,y)|x\in\Omega,\ y = A\{x\}\}$ נראה כך:



ציור 4.3 - המיפוי כגרף המתאר קבוצה פתוחה - דוגמה 4.3

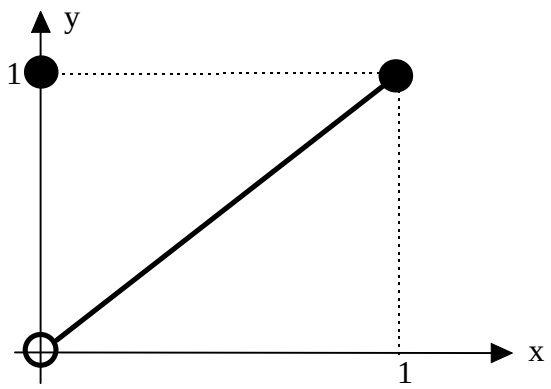
דוגמה 4.4: נתבונן במיפוי הבא: $A\{x\} = \begin{cases} x-2^{-1/x} & x\ne 0 \\ 1 & x=0 \end{cases}$ הפועל בתחום [0,1]. נראה כי מיפוי זה אינו סגור. לדוגמה, הסדרה:

$$\{x_k\}_{k=0}^{\infty} = 2^{-k} \quad ; \quad x_\infty = \lim_{k\to\infty} x_k = 0$$

מגדירה את הסדרה הבאה:

$$\{y_k\}_{k=0}^{\infty} = \{A\{x_k\}\}_{k=0}^{\infty} = 2^{-k} - 2^{-2^k} \quad ; \quad y_\infty = 0$$

אך כמו כן מתקיים כי - $A\{x_\infty\} = A\{0\} = 1 \neq y_\infty$. כמקודם, גם הפעם נכרת בעיית הסגירות בשל אי-רציפותה של A. הקבוצה $\{(x,y) | x \in \Omega,\ y = A\{x\}\}$ תיראה כך:



ציור 4.4 - המיפוי כגרף המתאר קבוצה פתוחה - דוגמה 4.4

4.4 משפט ההתכנסות הגלובלית

הגענו כעת אל עיקרו של פרק זה, וכל שנאמר עד כה נועד לשרת את הגדרתו, הבנתו והוכחתו של המשפט הבא הקרוי **משפט ההתכנסות הגלובלית**.

משפט 4.1: (משפט ההתכנסות הגלובלית). יהי A אלגוריתם מעל Ω, ונניח כי בהינתן נקודת איתחול כלשהי x_0, מיוצרת הסדרה $\{x_k\}_{k=1}^{\infty}$ ע"י $x_{k+1} = A\{x_k\}$. יהי $\Gamma \subseteq \Omega$ קבוצת פתרונות נתונה, ונניח כי:

1. כל הנקודות $\{x_k\}_{k=1}^{\infty}$ מוכלות בקבוצה קומפקטית $S \subseteq \Omega$.
2. קיימת פונקצית ירידה Z (לפי הגדרה 4.3).
3. המיפוי A סגור לכל הנקודות מחוץ לקבוצת הפתרונות.

אזי, הגבול של כל תת-סידרה מתכנסת של $\{x_k\}_{k=1}^{\infty}$ הוא פתרון.

הוכחה: נניח כי תת סידרה כלשהי של $\{x_k\}_{k=1}^{\infty}$ אשר נסמנה $\{x_k\}_{k\in\text{א}}$ מתכנסת לנקודה x שאינה פתרון, ונראה סתירה. נעשה זאת בשני שלבים - בשלב ראשון נראה כי התכנסות תת-סידרה זו ל-x מחייבת כי $\{Z(x_k)\}_{k=1}^{\infty} \xrightarrow[k\to\infty]{} Z(x)$. בשלב שני נראה כי הקביעה כי x אינה פתרון יוצאת סתירה.

כיוון ש- $\{x_k\}_{k\in\text{א}} \to x$, וכיוון שהפונקציה Z רציפה (ראה את הגדרתה), מתקבל כי $\{Z(x_k)\}_{k\in\text{א}} \to Z(x)$. בשל העובדה ש- $\{Z(x_k)\}_{k\in\text{א}} \to Z(x)$ מתקבל כי לכל $\varepsilon > 0$ קיים $k_0 \in \text{א}$, כך ש- $Z(x_{k_0}) - Z(x) < \varepsilon$. בשל היות Z פונקציה מונוטונית לא עולה אנו מקבלים כי לכל $k \ge k_0$ (בסדרה המקורית) קיים כי $Z(x_k) - Z(x) \ge 0$ מחד, ומאידך -

$$Z(x_k) - Z(x) = Z(x_k) - Z(x_{k_0}) + Z(x_{k_0}) - Z(x) \le Z(x_{k_0}) - Z(x) < \varepsilon$$

(כיוון שהגודל $Z(x_k) - Z(x_{k_0})$ קטן או שווה לאפס). מכאן ש- $\{Z(x_k)\}_{k=1}^{\infty} \xrightarrow[k\to\infty]{} Z(x)$.

כעת נחזור להנחתנו הבסיסית כי x אינה נקודת פתרון ונראה סתירה לנתונים. אם x אינה פתרון, פירוש הדבר שהפעלת A עליה תיתן ירידה ממש ב- Z, כלומר $Z(A\{x\}) < Z(x)$ (נובע מהגדרת פונקצית ירידה). מצד שני, נוכל לקחת את הסדרה $\{x_{k+1}\}_{k\in\text{א}}$ (הזזת תת-הסדרה ממקודם ב- 1), וקל לראות כי סידרה זו היא בעצם $\{A(x_k)\}_{k\in\text{א}}$. סידרה זו מתכנסת ל- x כיוון שהזזה ב- 1 (באינדקס) לא משפיעה על יעד ההתכנסות. בשל היותה של A מיפי סגור נדרוש כי:

$$\left.\begin{array}{c} \{x_k\}_{k\in\text{א}} \to x \\ \{A\{x_k\}\}_{k\in\text{א}} \to x \end{array}\right\} \quad x = A\{x\}$$

אך זה עומד בסתירה לעובדה ש: $Z(A\{x\}) < Z(x)$ על מנת ש- x לא יהיה פתרון. לכן ברור כי תת הסדרה השרירותית אשר מתכנסת, מתכנסת בהכרח לפתרון, והמשפט מוכח. ■

זהו משפט מורכב, ואינו אינטואיטיבי. בעיה אחת בהבנתו היא עניין הצורך בסגירות של A. בעיה אחרת היא ההתעסקות בתת-סדרות במקום בכל הסדרה כמקשה אחת. בדוגמאות הבאות ננסה ללבן נקודות אלה.

דוגמה 4.5: ניקח את האלגוריתם מדוגמה 4.3. נניח שאלגוריתם זה נועד למציאת מינימום לפונקציה $f(x)=|x|$. ברור כי איתחול האלגוריתם בערך גבוה מ- 1 ייתן כי האלגוריתם נתקע על 1 ולא מוצא את הפתרון הנכון.

מבחינת המשפט, התנאי היחידי שאינו מתקיים הוא סגירותו של המיפוי. כל החצי הראשון של ההוכחה לעיל נכון למקרה פרטי זה, אלא שנקודת ההתכנסות של A (היא 1), וחצי ההוכחה השני אינו תקף כאן.

דוגמה 4.6: נניח כי רצוננו למצוא מינימום לפונקציה $f(x)=(x-1)^2\cdot(x+1)^2$, וסידרת הנקודות המוצעת ע"י אלגוריתם כלשהו היא:

$$x_k=(-1)^k\cdot\left(2^{-k}+1\right)$$

לקיחת האיברים הזוגיים בסדרה זו תיתן את הסדרה $y_k=x_{2k}=\left(2^{-2k}+1\right)\xrightarrow[k\to\infty]{}1$.

לקיחת האיברים האי-זוגיים תיתן את הסדרה $z_k=x_{2k-1}=-\left(2^{-2k-1}+1\right)\xrightarrow[k\to\infty]{}-1$.

באופן זה ניכר ששתי תתי-סדרות אלו מתכנסות כל אחת לפתרון נכון אך כשהן משולבות הן נותנות סידרה המתנהגת באופן קופצני ב- x, אך עדיין מניבה ירידה מונוטונית של הפונקציה f(x). דוגמה זו באה להמחיש מדוע טענת המשפט מתמקדת בתת-סדרה מתכנסת.

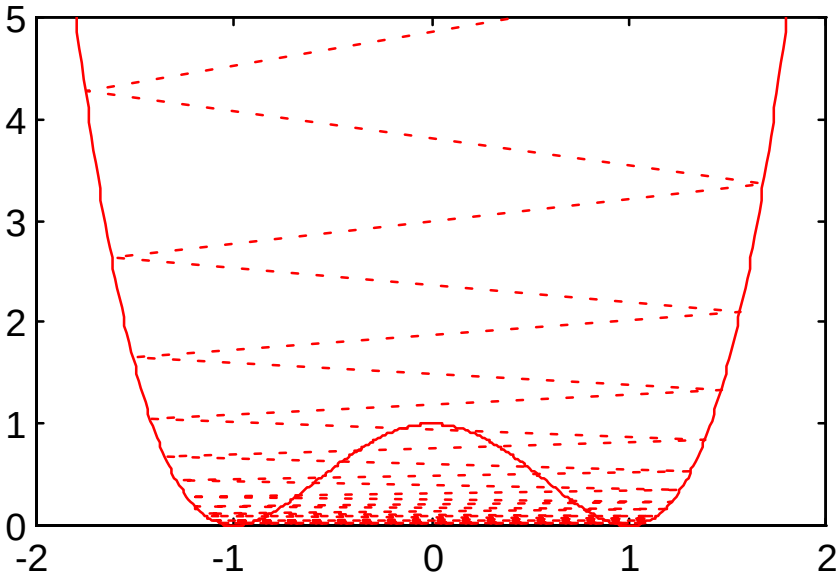
דוגמה 4.7: נמשיך את דוגמה 4.1. נראה כי כל הדרישות מתקיימות ולפיכך האלגוריתם המוצע בעל תכונת התכנסות גלובלית.

א. המיפוי $A(x)=\sqrt{x}+1$ רציף לכל x (גם מעל הקומפלקסיים).

ב. קיימת ירידה - ניתן להראות זאת ע"י הוכחת הקשר - $\forall x \quad |x-2.618|^4>\left|\sqrt{x}-1.618\right|^4$.

ג. הנקודה x בתחום סגור וסופי - נוכל פשוט להחליט שרירותית כי x הוא כל מספר קומפלקסי בתוך מעגל ברדיוס r כלשהו.

מדוע משפט זה חשוב לנו ? ובכן, במהלך ניתוח אלגוריתמים אשר מתיימרים לפתור בעיות אופטימיזציה ניתקל שוב ושוב באותו מבנה עקרוני של איטרציות היוצרות סידרה של נקודות תחום קומפקטי, פונקצית ירידה המוקטנת בכל שלב, וישנה גם כמובן קבוצת פתרונות. ע"י בדיקה של קיום תנאי משפט זה נוכל לקבוע בנוחות כי האלגוריתם מתכנס.



ציור 4.5 - הפונקציה וסידרת הנקודות עבורה - דוגמה 4.6

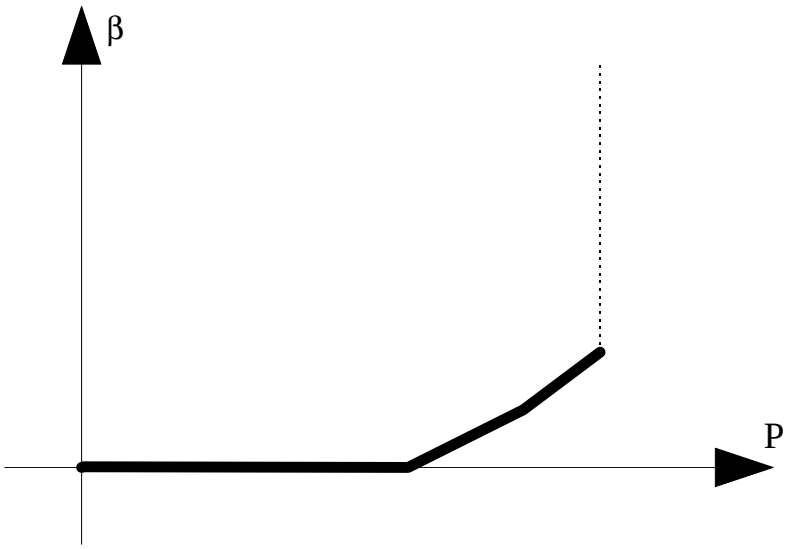
4.5 אומדנים למהירות התכנסות

כיוון שאנו עוסקים בעקרונות של אלגוריתמים איטרטיביים, ישנו עניין נוסף בו האלגוריתמים זוכים למכנה משותף והוא עניין קצב ההתכנסות. בהינתן סידרת נקודות $\{x_k\}_{k=1}^{\infty}$ המתקבלת ע"י $x_{k+1} = A\{x_k\}$, ואשר אמורה להתכנס ל- x_∞, נוכל להגדיר את שגיאת הפתרון ה- k ע"י $e_k = \|x_k - x_\infty\|$, ולדון בשאלה כיצד אומדים קצב התכנסות של אלגוריתם.

הגדרה 4.5: תהי הסדרה $\{x_k\}_{k=1}^{\infty}$ מתכנסת ל- x_∞. סדר ההתכנסות (order of convergence) לסדרה זו מוגדר ע"י הסופרמום על המספרים האי-שליליים p המקיימים כי:

$$0 \le \overline{\lim_{k\to\infty}} \frac{\|x_{k+1} - x_\infty\|}{\|x_k - x_\infty\|^p} = \overline{\lim_{k\to\infty}} \frac{e_{k+1}}{e_k{}^p} < \infty$$

בעיקרון, עלינו להציב את כל ערכי p החיוביים ובדוק את ערכו של הגבול הנ"ל. תיאור גראפי של הגבול כפונקציה של p ייראה כגרף המתחיל באפס (בדרך כלל) וגרף זה מונוטוני לא יורד. בנקודה כלשהי p הגבול יהי סופי (יתכן גם אפס) ולכל ערך מעבר לו נקבל עליה לאינסוף. נקודת שבירה זו בה עדיין הגבול סופי היא סדר ההתכנסות, וגובה הגבול שם מוגדר כיחס ההתכנסות.



ציור 4.6 - התפתחות הגבול כפונקציה של הערך p

הערה: השימוש בגבול סופריורי ($\overline{\lim}$) נועד להגן ממצבים בהם מתקבל כי יש יותר מגבול אחד, ואז לפי ההגדרה יש לקחת את הגבול הגבוה ביותר. בדרך כלל נוכל להניח כי זהו גבול lim רגיל.

מההגדרה הנ"ל ניכר כי אנו מתייחסים לזנבה של הסדרה בקובענו את סדר ההתכנסות - עובדה לא מבורכת!! משמעות הגדרה זו הוא כמה מהר דועך זנב סידרת שגיאות הפתרון - ככל ש- p גדול יותר כך גם ניתן לומר כי הדעיכה מהירה יותר. עבור אותו ערך של p ערך נמוך יותר של הגבול פירושו התכנסות מהירה יותר. יש לשים לב לכך שאנו יכולים לקבל לשתי סדרות של שגיאות פתרון אותו סדר אך ערך שונה לגבול:

$$\overline{\lim_{k\to\infty}} \frac{\|x_{k+1}-x_\infty\|}{\|x_k-x_\infty\|^p} = \overline{\lim_{k\to\infty}} \frac{e_{k+1}}{e_k{}^p} = \beta$$

במקרה כזה - הסדרה לה ערך הגבול הנמוך יותר היא הטובה יותר.

דוגמה 4.8: נתונות שתי סדרות של שגיאות פתרון - $\{f_k\}_{k=1}^{\infty} = 3^{-k}$; $\{e_k\}_{k=1}^{\infty} = 2^{-k}$. נחשב את סדר ההתכנסות של שתי סדרות אלה ונשווה ביניהן.

עבור הסדרה הראשונה ניתן לראות כי:

$$\overline{\lim_{k\to\infty}}\frac{\|x_{k+1}-x_\infty\|}{\|x_k-x_\infty\|^p}=\overline{\lim_{k\to\infty}}\frac{e_{k+1}}{e_k{}^p}=\overline{\lim_{k\to\infty}}\frac{2^{-k-1}}{2^{-pk}}=\begin{cases}0 & p=0\\ 0.5 & p=1\\ \infty & p=2\\ \infty & 0=3\end{cases}$$

לכן לכאורה ברור כי סדר ההתכנסות הוא p=1 (הסופרמום על כל הערכים הנ"ל שאינם אינסוף). לכאורה - כיוון שיתכן ש- p הוא למשל 1.5. כדי לקבוע בודאות כי אמנם p=1, עלינו לבדוק את ערך הגבול ב- $p=p+\varepsilon$ ל- ε קטן מאוד:

$$\overline{\lim_{k\to\infty}}\frac{e_{k+1}}{e_k{}^{1+\varepsilon}}=\overline{\lim_{k\to\infty}}\frac{2^{-k-1}}{2^{-k}\cdot 2^{-\varepsilon k}}=0.5\cdot\overline{\lim_{k\to\infty}}\,2^{\varepsilon k}=\infty$$

ולכן סדר ההתכנסות הוא p=1 וערכו של β הוא 0.5. באופן דומה קל לראות כי לסדרה השניה יש אותו סדר התכנסות אך עם ערך β נמוך יותר - 0.3333, וכצפוי - היא מתכנסת מהר יותר.

דוגמה 4.9: הסדרה $\{e_k\}_{k=1}^{\infty}=2^{-\left(2^k\right)}$ מתכנסת בסדר p=2, כיוון ש:

$$\overline{\lim_{k\to\infty}}\frac{\|x_{k+1}-x_\infty\|}{\|x_k-x_\infty\|^p}=\overline{\lim_{k\to\infty}}\frac{e_{k+1}}{e_k{}^p}=\overline{\lim_{k\to\infty}}\frac{2^{-\left(2^{k+1}\right)}}{2^{-p\left(2^k\right)}}=\overline{\lim_{k\to\infty}}\,2^{-(2-p)\left(2^k\right)}=\begin{cases}0 & p=0\\ 0 & p=1\\ 1 & p=2\\ \infty & 0=3\end{cases}$$

לרוב האלגוריתמים אשר יידונו במסגרת קורס זה יש התכנסות מסדר 1, ולכן נראה כי רצוי לעיין בקבוצת האלגוריתמים בעלי סדר התכנסות 1 ביתר פירוט.

הגדרה 4.6: אם הסדרה $\{x_k\}_{k=1}^{\infty}$ מתכנסת ל- x_∞ באופן שמתקיים כי:

$$\lim_{k\to\infty}\frac{\|x_{k+1}-x_\infty\|}{\|x_k-x_\infty\|}=\lim_{k\to\infty}\frac{e_{k+1}}{e_k}=\beta<1$$

אזי הסדרה מוגדרת כמתכנסת ליניארית עם יחס התכנסות β.

למעשה, כיוון שהנ"ל מתקיים, קל לראות כי סידרת שגיאות הפתרון מתנהגת כ- $e_k=c\beta^k$, ולכן קרויה התכנסות זו גם התכנסות אקספוננציאלית או גיאומטרית. ישנם אלגוריתמים אשר

מניבים התכנסות מסדר 1 עם יחס התכנסות 0 - סדרות אלו ייקראו בעלות התכנסות סופר-ליניארית.

דוגמה 4.10: הסדרה $\{e_k\}_{k=1}^{\infty} = (1/k)^k$ מתכנסת בסדר p=1, כיוון ש:

$$\lim_{k\to\infty}\frac{e_{k+1}}{e_k{}^p} = \lim_{k\to\infty}\frac{\left(\frac{1}{k+1}\right)^{k+1}}{\left(\frac{1}{k}\right)^{pk}} = \lim_{k\to\infty}\left(\frac{k^p}{k+1}\right)^k\cdot\left(\frac{1}{k+1}\right) = \begin{cases} 0 & p=0 \\ 0 & p=1 \\ \infty & p=2 \\ \infty & 0=3 \end{cases}$$

כך שלסדרה זו התכנסות סופר-ליניארית. כאן השתמשנו בתכונה $\lim_{k\to\infty}\left[\frac{k}{1+k}\right]^k = \frac{1}{e}$.

מכיוון שני, עבור סידרת שגיאות מהצורה - $e_k = 1/k^2$ ניתן להראות כי סדר ההתכנסות הוא 1 ויחס ההתכנסות אף הוא 1 - כלומר - זוהי הסידרה האיטית ביותר מבין הסדרות בעלות סדר ההתכנסות הליניארי.

ניתן להגדיר במגוון דרכים אחרות קצב התכנסות. בחרנו להיעצר כאן משום שרוב הגדרות אלה דומות בעיקרן, ולמעשה נשתמש בזו שהוצגה כאן. נקודה אחרונה שרצוננו להעלות במסגרת הדיון בקצב ההתכנסות היא בחירת מטריקת המרחק והשפעתה על תוצאת קצב ההתכנסות. בהגדרה עד כה הסתכלנו בשגיאת הפתרון במונחים של מרחק אויקלידי בין הנקודה ה- k ובין נקודת המינימום אליה אנו חותרים. נניח כי אנו עוסקים בהתכנסות ליניארית, ואנו בוחרים להשתמש כתחליף למרחק הנ"ל בערכה של הפונקציה בנקודות אלו. כלומר -

$$\lim_{k\to\infty}\frac{\|x_{k+1}-x_\infty\|}{\|x_k-x_\infty\|} = \beta < 1 \quad\Rightarrow\quad \lim_{k\to\infty}\frac{\|f(x_{k+1})-f(x_\infty)\|}{\|f(x_k)-f(x_\infty)\|} \overset{?}{=} \beta$$

והשאלה שנשאלת האם יתכן שנקבל למדדי מרחק שונים יחס התכנסות שונה? הלמה הבאה (אשר תינתן ללא הוכחה) עונה לשאלה זו.

למה 4.1: בהינתן שתי פונקציות שגיאה $f(x,x_\infty),\ g(x,x_\infty)$ בהן הנקודה x_∞ נותנת שגיאה אפס, ומתקיים כי לכל x קיימים שני קבועים a ו- b כך ש:

$$0 < a \cdot g(x, x_\infty) \le f(x, x_\infty) \le b \cdot g(x, x_\infty) < \infty$$

אזי, אם סידרה מתכנסת בקצב ליניארי עם יחס β בהתייחס לאחת מפונקציות אלה, כך גם יתקבל לגבי הפונקציה השנייה.

דוגמה 4.11 : נניח כי אנו מציעים את הפונקציות הבאות:

1. $f(x, x_\infty) = \|x - x_\infty\|^2$
2. $g(x, x_\infty) = (x - x_\infty)^T Q (x - x_\infty)$

כאשר Q מטריצה חיובית מוגדרת (אחרת אין זו מטריקה נאותה). אזי ניתן להראות כי עבור בחירת a ו- b כערכים העצמיים הקטן ביותר והגדול ביותר של Q בהתאמה יתקבל כי תנאי הלמה מתקיימים, ולכן מדידת מרחק אויקלידי במישור הפתרונות, או את ערכה של פונקציה קוודרטית - תוצאת קצב ההתכנסות זהה:

$$\lambda_{min} \|x - x_\infty\|^2 \le (x - x_\infty)^T Q (x - x_\infty) \le \lambda_{max} \|x - x_\infty\|^2$$

4.6 סיכום

בפרק זה התמקדנו במשפחה הרחבה ביותר של אלגוריתמים נומריים אשר נועדו לפתור בעיות אופטימיזציה, ונוקטים בגישה של תהליך איטרטיבי המוריד באופן עקבי בערכה של פונקציה. הראינו כי תהליך כזה מבטיח תחת תנאים מסוימים (לגמרי לא אינטואיטיביים) כי תחול התכנסות גלובלית (דהיינו - מכל נקודת איתחול שרירותית).

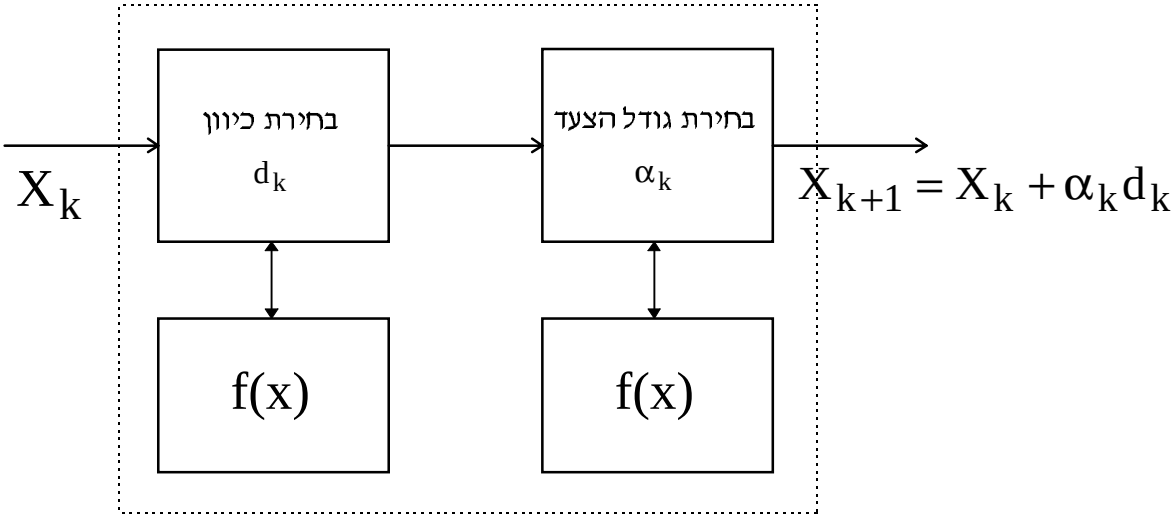
זהו הפרק האחרון בו נדון באופן כה תיאורטי בבעיות אופטימיזציה נטולות אילוצים. בפרקים הבאים ניגע באלגוריתמים נומריים ממש. לטיפול תיאורטי נשוב לקראת הטיפול בבעיות אופטימיזציה עם אילוצים.

פרק 5 - אלגוריתמי חיפוש על ישר

בפרק זה נתאר אלגוריתמים אשר נועדו למצוא נקודת מינימום של פונקציות במשתנה יחיד - דהיינו - בעיות אופטימיזציה חד מימדיות. בסעיף הראשון נדון במוטיבציה לעיסוק במקרה פרטי מעין זה של בעיות אופטימיזציה, ובסעיפי ההמשך של פרק זה נתאר את האלגוריתמים המוצעים וניתוחם.

5.1 מדוע חיפוש על ישר ?

בפרק קודם הצגנו אלגוריתם אופייני כמכונה איטרטיבית אשר בכל צעד יוצרת ע"י מיפוי עדכון לפתרון, באופן שיקטין פונקציה כלשהי. כעת נאמר יותר מכך - נניח כי האלגוריתם אותו רצוננו להפעיל נמצא בנקודה כלשהי x המוצעת כפתרון, ורצונו לעדכנה. כל אלגוריתם בו נעסוק כמעט יבצע את פעולת האיטרציה בשני שלבים - בשלב ראשון ייבחר האלגוריתם כיוון d (עם סימן) בו האלגוריתם מוצא כי כדאי לנוע, ובשלב שני ימצא האלגוריתם מהו המרחק האופטימלי (או תת-אופטימלי) בו יש ללכת כדי להשיג ירידה מירבית. הן חישוב הכיוון והן מציאת המרחק יבוצעו ע"י פניה לפונקציה ל-"התייעצות". רעיון זה מתואר בציור 5.1. כל אלגוריתם בעל מבנה כזה ישתמש שוב ושוב (בכל איטרציה פעם אחת) לרוטינה המבצעת חיפוש מינימום על ישר. פרק זה מספק עיון מפורט בדרכים לביצוע חיפוש מינימום על ישר.



ציור 5.1 - סכימת בלוקים של האלגוריתם העקרוני

אל כל האמור לעיל ניתן גם להתיחס באופן הבא - בהתקבל בעיית מינימיזציה בה יש נעלמים רבים, אנו מציעים את המרתה בפתרון בעיות רבות של מינימיזציה חד-מימדית, ופרק זה עוסק בשאלה - כיצד מבצעים מינימיזציה כזו.

סוגיה חשובה בה אנו נדרשים לטפל היא זו - בהינתן כיוון אשר נבחר - d. האם אלגוריתם אשר איתר את נקודת המינימום על הקרן אשר הוגדרה (ישר חד כיווני) הוא אופרטור סגור? אנו ראינו בפרק קודם כי תכונה זו חשובה לשם הוכחת התכנסותו הגלובלית של האלגוריתם הכולל. נגדיר את המיפוי אשר מבוצע בגישה המתוארת כאן:

משפט 5.1: המיפוי $S:E^{2n}\to E^{n}$ מוגדר באופן הבא -

$$S\{x,d\}=\left\{y:\ y=x+\alpha_0 d\ \ ,\ \alpha_0=\arg\min_{\alpha}\{f(x+\alpha d)\}\right\}$$

תהי הפונקציה f רציפה מעל E^{n}. אזי המיפוי אשר הוגדר לעיל - $S\{x,d\}$ הינו סגור עבור כל $d\neq 0$.

הוכחה: תהיינה $\{x_k\}_{k=1}^{\infty}$ ו- $\{d_k\}_{k=1}^{\infty}$ סדרות כלשהן אשר מתכנסות לצמד -(x,d) כאשר $d\neq 0$. לכל צמד (x_k,d_k) מותאם α_0^k אשר יביא למינימום את הפונקציה $f(x_k+\alpha d_k)$ ובכך תוגדר נקודה $y_k=x_k+\alpha_0^k d_k$, כך ש- $y_k=S\{x_k,y_k\}\xrightarrow[k\to\infty]{}y$.

כדי להראות כי המיפוי המוצע סגור עלינו להראות כי $y=S\{x,d\}$. לשם כך נשתמש בעובדה שסידרת הסקלרים המתקבלת מתכנסת ל:

$$\left\{\alpha_0^k\right\}_{k=1}^{\infty}=\left\{\frac{y_k-x_k}{d_k}\right\}_{k=1}^{\infty}\xrightarrow[k\to\infty]{}\alpha_0^{\infty}=\frac{y-x}{d}$$

(זהו המקום בהוכחה בו אנו נדרשים לקביעה - $d\neq 0$, ויש לישם לב לעובדה שהחלוקה כאן הינה של וקטור בוקטור המניבה סקלר). לכן - $y=x+\alpha_0^{\infty}d$. נשתמש בעובדה שמתקיים כי לכל α:

$$f(y_k)\le f(x_k+\alpha d_k)\ \ \Rightarrow\ \ f(y)\le f(x+\alpha d)$$

ולכן, מתקבל גם הקשר $f(y)\le\min_{\alpha}f(x+\alpha d)$. קשר זה פירושו ש- $y=S\{x,d\}$, ולכן המיפוי S סגור כפי שנטען. ■

משמעות משפט זה היא שכל אלגוריתם מעל תחום קומפקטי אשר מתבסס על גישה זו (קביעת כיוון חיפוש ומציאת נקודת מינימום על הישר שהוגדר), כולל בחלק של איתור המינימום על

הישר מיפוי סגור. לכן, על מנת להבטיח שאלגוריתם זה מתכנס כל שעלינו לעשות הוא לדאוג שקביעת הכיוון באלגוריתם מהווה מיפוי רציף, ושיש פונקצית ירידה לאלגוריתם, ואז מתקיימות כל דרישות משפט ההתכנסות הגלובלית.

דוגמה 5.1: נמחיש את עניין הדרישה $d \neq 0$. נניח כי רצוננו להביא למינימום את הפונקציה $f(x) = (x-1)^2$, ונראה כי $S\{x,y\}$ אינו סגור בנקודה $(x,0)$. נניח כי אנו נמצאים בנקודה כלשהי x (לדוגמה - x=0), ונבחר את סידרת הכיוונים $\left\{d_k = 2^{-k}\right\}_{k=1}^{\infty} \to 0$. אזי:

$$S\{0,d_k\} = \left\{y:\ y = x + \alpha_0 d\ ,\ \alpha_0 = \arg\min_{\alpha}\{f(x+\alpha d)\}\right\} =$$
$$= \left\{y:\ y = 0 + \alpha_0 2^{-k}\ ,\ \alpha_0 = 2^k\right\} = 1 = y_k$$

ולכן -גבול סידרת המיפוי יהיה $y = 1$. לעומת זאת, נוכל לומר כי:

$$S\{0,0\} = \left\{y:\ y = x + \alpha_0 d\ ,\ \alpha_0 = \arg\min_{\alpha}\{f(x+\alpha d)\}\right\} = \{y:\ y = 0 + \alpha_0 0\ ,\ \alpha_0 = 0\} = 0$$

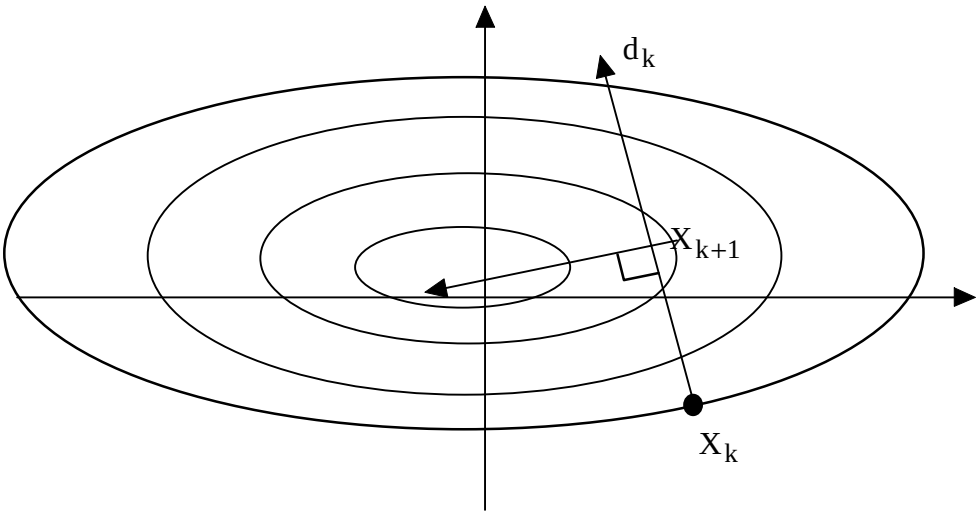
כך שאמנם, בשל היות d=0 קיבלנו כי המיפוי אינו סגור.

משפט 5.2: (משפט הניצבות). תהי הפונקציה f רציפה מעל E^n, ויהי המיפוי $S: E^{2n} \to E^n$ נתון כמקודם. אזי קיים כי $\nabla f(S\{x,d\}) \cdot d = 0$, דהיינו גרדיאנט הפונקציה בנקודת המינימום על הישר אנכית לכיוון החיפוש d.

הוכחה: למעשה, גודל הצעד α_0 מקיים כי $\frac{\partial f(x+\alpha d)}{\partial \alpha} = 0 = \nabla f(x+\alpha_0 d) \cdot d$. ■

לפני שנכנס לעניין מציאת נקודת מינימום על ישר, נגדיר את הבעיה אותה אנו רוצים לפתור: נתונה פונקציה f של משתנה סקלרי x. נתונה נקודה בה אנו עומדים כעת - x_0. רצוננו להפעיל אלגוריתם A כלשהו אשר יספק סידרת נקודות $\{x_k\}_{k=1}^{\infty}$ אשר יקיימו כי f(x) מוקטנת באופן מונוטוני עבורם, דהיינו -

$$f(x_1) > f(x_2) > f(x_3) > \cdots \geq f(x_n) \geq \cdots$$



ציור 5.2 - המחשת משפט הניצבות

והשאלה אותה אנו שואלים היא - מהו A אותו יש להפעיל? למרות שנדמה ששאלה זו מוגדרת היטב, היא עדיין אינה כזו, כיוון שלא אמרנו כמה מורכבות נתיר ל- A לפתח בפתרונה. למשל - אלגוריתם אשר נוכל להציע בקלות הוא כזה אשר יבצע חיפוש סידרתי בקפיצות גסות ויאתר את נקודת המינימום, סביבה בתחום מסוים יחזור על החיפוש באופן עדיו וחוזר חלילה - זהו אלגוריתם אשר יתכנס, אך במחיר לא נעים, והיינו רוצים להימנע מגישות כאלה אשר מקילות על החשיבה במחיר של סיבוך האלגוריתם.

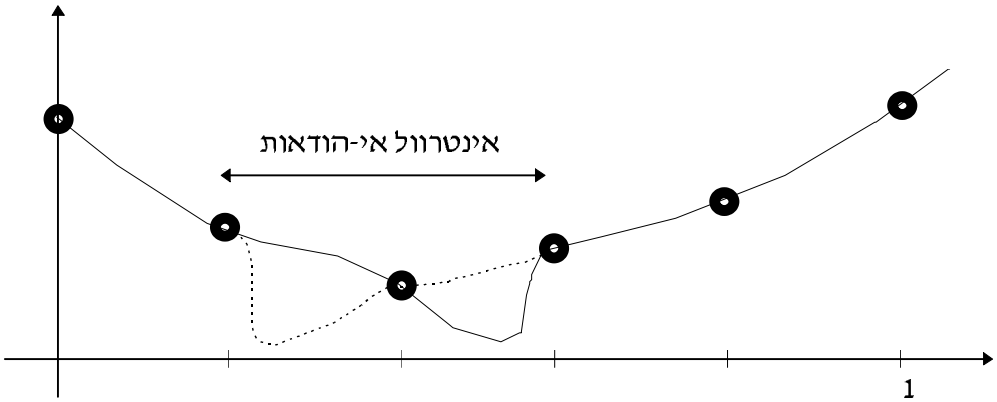
בהקשר לסיבוכיות אותה הזכרנו, יוכלו אלגוריתמים מסויימים להסתפק בקריאה לרוטינה המחזירה את ערך הפונקציה, אלגוריתמים אחרים יעשו שימוש ברוטינה אשר תחזיר את ערך הגרדיאנט (שהינו סקלר בשל היות הפונקציה חד מימדית), ואלגוריתמים מורכבים יותר יתירו לעצמם שימוש בנגזרות מסדר גבוה יותר - וברור שככל שהמידע העומד לרשות האלגוריתם גדול יותר כך גדולה יותר יכולתו לספק התכנסות מהירה יותר. במסגרת פרק זה נכיר אלגוריתמים מכל קטגוריות אלו.

5.2 סידרת פיבונצ'י וחיפוש הזהב

שיטת חיפוש הזהב אותה נכיר מיד היא בעלת יופי תיאורטי, אך חשיבותה המעשית אינה רבה. על בעית המינימיזציה שתוארה לעיל נוסיף את ההנחה כי החיפוש נעשה מעל אינטרוול סגור $[0,C]$. כן נניח כי הפונקציה אותה רצוננו להביא למינימום היא חד-אופנית (uni-modal) , דהיינו - יש לה מינימום מקומי אחד. לרוב הבעיות באופטימיזציה בהן ניתקל אין בהנחה הראשונה משום פגיעה קשה כיוון שהחיפוש הוא כיווני עם סימן, ונוכל לחסום את מרחק העניין שלנו מהנקודה בה אנו נמצאים - וכך יתקבל אינטרוול סגור. לגבי ההנחה השניה, ישנם מצבים בהם הנחה זו מגבילה ביותר - אך ללא הנחה זו מרבית אלגוריתמי החיפוש על ישר קורסים. משמעותה המעשית של הנחה זו כי כנראה שאם יש יותר ממינימה מקומית אחת - אנו נינעל מקומית ולא גלובלית.

השאלה אותה אנו באים לפתור היא הבאה: בידנו פונקציה $f(x)$ סקלרית חד-אופנית של משתנה סקלרי המצוי בתחום $[0,C]$, ורצוננו למצוא בדיוק טוב ככל האפשר את מיקום המינימום של פונקציה זו תוך שימוש ב- N חישובים של ערך הפונקציה בנקודות $\{x_k\}_{k=1}^{N}$ שנבחר. נמחיש את עניין קביעת דיוק מיקום המינימום דרך דוגמה:

דוגמה 5.2: עבור פונקציה חד-אופנית כמתואר בציור 5.3 ואשר מוגדרת מעל האינטרוול [0,1] רצוננו למצוא את נקודת המינימום ע"י מדידת גובהה ב- 6 מקומות. נציע את המיקומים - 0.0, 0.2, 0.4, 0.6, 0.8 ו- 1.0. נבחר את הנקודה בה גובה הפונקציה מינימלי כנקודת המינימום המקורבת. עפ"י הציור ניכר כי נקודת המינימום האמיתית יכולה להיות ימינה או שמאלה מהנקודה שנקבעה במרווח של 0.2 לכל היותר. אינטרוול אי-הודאות הוא לפיכך 0.4.



ציור 5.3 - הצגת עיקרון אינטרוול אי-הוודאות המשמש בשיטת הזהב

מהן הנקודות בהן כדאי לבחור לקבלת אינטרוול אי-ודאות מינימלי? מהי אסטרטגיית החיפוש האופטימלית (במובן של אי-ודאות מינימלית באיתור המינימום), ומהי סטיית הפתרון המירבית? המענה לכל שאלות אלו טמון בגישת חיפוש הזהב הבנויה על סידרת פיבונצ'י. אנו לא נוכיח כאן את היותה אופטימלית, אך נתאר אותה בקצרה. תחילה נגדיר את סידרת פיבונצ'י:

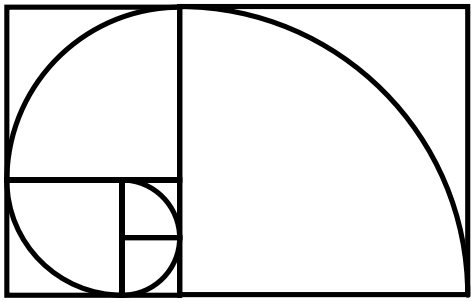
הגדרה 5.1: סידרת פיבונצ'י נתונה ע"י הקשרים הבאים:

$$F_0 = F_1 = 1 \qquad F_n = F_{n-1} + F_{n-2}$$

$$\Rightarrow \{F_0,\ F_1,\ F_2,\ F_3,\ F_4,\ \ldots\} = \{1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34,\ \ldots\}$$

לפני שנספר על שיטת חיפוש הזהב, נספר מעט על סידרה זו שזכתה במהלך השנים לעניין רב. פיבונצ'י היה מתמטיקאי בן המאה ה- 12 אשר חי בפיזה. הוא התעניין בשאלת התרבותן של ארנבות בזמן, ובהקשר זה הגדיר את הסידרה הקרויה על שמו. מאוחר יותר התגלה כי מגוון רחב של תופעות טבע הינן בעלות מבנה הקשור באופן הדוק לסידרת פיבונצ'י. ביניהן נזכיר את:

- מספר פיצולים בעץ כפונקציה של גובהו.
- ספירלות בטבע (קונכיות, גרעינים בחמניה, אבקנים בפרחים) מסתדרות באופן הקשור באופן הדוק לסידרת פיבונצ'י. נתאר תהליך בניית ספירלה: בנה ריבוע בעל צלע באורך יחידה. בנה לצידו ריבוע זהה. הצמד\י למלבן המתקבל ריבוע הצמוד לצלע הגדולה. חזור\י על צעד זה שוב ושוב במסלול המקיף את הראשית. ניתן כעת לצייר ריבעי מעגלים אשר מתחברים לספירלה. אורכי הצלעות של הריבועים המצטרפים הם סידרת פיבונצ'י.



ציור 5.4 - ספירלה הקשורה לסידרת פיבונצ'י

- היוונים הגדירו את יחס הזהב כיחס 0.618. מקור קביעה זו הוא בקו ובו 3 נקודות כמתואר בציור 5.5 - נקודות A, B ו- C. רצוננו למצוא את מיקום הנקודה C כך שיתקבל השיויון הבא באורכי הקטעים:

$$\frac{AC}{CB} = \frac{CB}{AB}$$

ציור 5.5 - יחס הזהב באורכי אינטרוולים

אם מסמן את אורך האינטרוול BC ב- a ואת אורכו של האינטרוול AC ב- b, הרי שנקבל כי המשוואה הנ"ל היא:

$$\frac{b}{a} = \frac{a}{a+b} \quad \Rightarrow \quad b^2 + ab - a^2 = 0 \quad \Rightarrow \quad \begin{Bmatrix} a = 1 \\ b = 0.618 \text{ or } -1.618 \end{Bmatrix}$$

בהתבסס על היחס הנ"ל, היוונים הגדירו כמלבן אסטתי ביותר את המלבן אשר יחס צלעותיו הוא 0.618. יחס זה הוא מספר המאפיין את סידרת פיבונצ'י, כפי שנראה בהמשך.

- יחס הזהב הנ"ל מתקבל כשורש של המשוואה הריבועית $\tau^2-\tau-1=0$. בהינתן השורש τ, מסתבר כי ניתן באמצעותו לחשב את כל חזקותיו בפשטות, כביטוי פשוט וליניארי עליו, אשר שוב חושף את סידרת פיבונצ'י:

$$\begin{aligned}\tau^2 &= \tau+1\\ \tau^3 &= 2\tau+1\\ \tau^4 &= 3\tau+2\\ &\vdots\\ \tau^n &= F_{n-1}\tau+F_{k-2}\end{aligned}$$

פרטים רבים נוספים על סידרת פיבונצ'י ודברים סמוכים ניתן למצוא באתר האינטרנט הבא, אותו מנהל מתמטיקאי בריטי: http:\\www.mcs.surrey.ac.uk\personal\R.Knott\Fibonacci.

נחזור לענייני חיפוש על ישר. האלגוריתם מתואר בציור 5.6. אלגוריתם זה מניב סידרת אינטרוולי אי-ודאות ההולכים ויורדים, ואשר אורכם נתון ע"י:

$$\varepsilon_k=\left(\frac{F_{N-k+1}}{F_N}\right)\cdot C$$

וכאמור, זהו האלגוריתם האופטימלי תחת התנאים שתוארו לעיל. יש לשים לב כי תכונה לא רצויה של אלגוריתם זה היא שיש להחליט מראש על מספר הצעדים בו ננקוט בחיפוש - במקרים רבים פירוש הדבר חוסר יעילות משווע.

דוגמה 5.3: נתאר את התפתחותו של האלגוריתם עבור N=2,3,5 על אינטרול יחידה. לכל אורך התיאור נניח כי תמיד הנקודה הקרובה לראשית היא בעלת ערך המינימום.

N=2: במקרה זה שתי הנקודות הינן זהות - $\left(\frac{F_1}{F_2}\right)=\left(1-\frac{F_1}{F_2}\right)=0.5$. מעשית יש לבחור את הנקודות $x_1=0.5+\delta$, $x_2=0.5-\delta$. הנקודה הנמוכה תעיד על אינטרוול באורך 0.5 בו מצוי המינימום. אם נבחר שתי נקודות אחרות על האינטרוול - a ו- b (נניח כי a<b) אינטרוול אי הודאות הגרוע ביותר יהיה - $\max\{b,1-a\}$. לכל בחירה של צמד זה, ברור כי גודל זה גדול מ- 0.5 (אם b קטן מ- 0.5, הרי ש- 1-a גדול מ- 0.5, ואם b גדול מ-0.5 - שוב הגודל הזה גדול מ- 0.5).

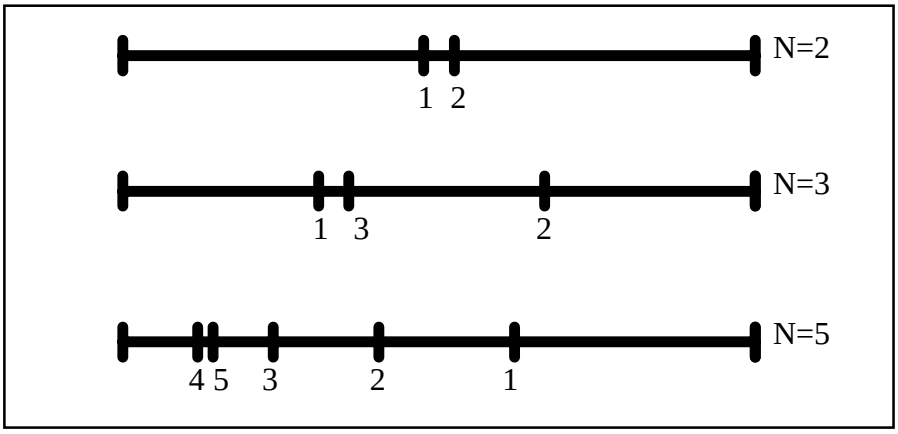
1. בשלב האיתחול אי הודאות למיקומו של המינימום היא $\varepsilon_1 = C$
2. מקם את שתי המדידות הראשונות בנקודות $\left[F_{N-1}/F_N\right]C$ ו- $\left[1-F_{N-1}/F_N\right]C$.
3. מצא מבין שתי המדידות את הנמוכה יותר. סביב נקודה זו באינטרוול באורך $\varepsilon_2 = \left[F_{N-1}/F_N\right]C$ צפויה להימצא נקודת המינימום.
4. מקם את המדידה השלישית בתוך אינטרוול זה בנקודה סימטרית לנקודה בה נמצא המינימום עד כה.
5. חזור על סעיפים 2 ו- 3 לנקודות הבאות עד לנקודה ה- N.
6. בסיום התהליך יימצא בידנו אינטרוול אי-ודאות שגודלו $\varepsilon_N = \left[F_1/F_N\right]C$

<u>ציור 5.6 - אלגוריתם חיפוש הזהב עפ"י סידרת פיבונצ'י</u>

N=3: במקרה זה שתי הנקודות הראשונות הן - $\left(1-\frac{F_2}{F_3}\right)=\frac{1}{3}$; $\left(\frac{F_2}{F_3}\right)=\frac{2}{3}$. (נניח כי מצאנו כי הנקודה 0.3333 היא בעלת המינימום) - אזי - מוגדר כעת אינטרוול באורך 0.666666 של אי-ודאות למיקום המינימום, והנקודה הנתונה ממוקמת באמצעיתו. לכן, בדומה לקודם, מיקום סימטרי נותן את אותו מיקום - נמקם בהסטה קטנה מאוד, ונקבל על-פי מיקום המינימום כי אינטרוול האי-ודאות הוא באורך 0.3333.

N=5: נקודות ההתחלה יהיו $\left(1-\frac{F_4}{F_5}\right)=\frac{3}{8}$; $\left(\frac{F_4}{F_5}\right)=\frac{5}{8}$ ואינטרוול האי-ודאות הוא באורך 5/8. באופן סימטרי לנקודה 3/8 ממקומת הנקודה 2/8, ומגדירה אינטרוול אי-ודאות באורך 3/8. באופן סימטרי לנקודה האחרונה ממוקמת הנקודה 1/8, ובאופן סימטרי אליה מתקבלת שוב אותה נקודה , ולכן נזיז אותה קלות.

ציור 5.7 מתאר את התפתחות החיפוש לשלושת מקרים אלו.



ציור 5.7 - תיאור חיפוש הזהב - דוגמה 5.3

משפט 5.3: עבור אלגוריתם חיפוש הזהב אשר תואר לעיל, ועבור N גדול דיו מתקבל כי אלגוריתם זה מתכנס בסדר ליניארי עם יחס 0.618 לנקודת המינימום באינטרוול.

הוכחה: סידרת פיבונצ'י מוגדרת כמשוואת ההפרש - $F_n = F_{n-1} + F_{n-2}$. ניתן לפתור משוואת הפרש זו באופן אנליטי. הפתרון הכללי ביותר נתון ע"י:

$$F_n = A\tau_1^n + B\tau_2^n$$

את הגדלים השונים בביטוי זה מוצאים ע"י הצבה ישירה למשוואת הרקורסיה של הטור. כאשר τ_1, τ_2 הם שורשיו של הפולינום האופייני - $\tau^2 - \tau - 1 = 0 \quad \Rightarrow \quad \tau_1, \tau_2 = -0.618,\ 1.618$. המספר 1.618 קרוי חלוקת הזהב, והוא מוכר מהקשרים אחרים במתמטיקה. עבור מספר גדול של נקודות N מתקבל כי:

$$\lim_{N\to\infty}\left(\frac{F_{N-1}}{F_N}\right) = \frac{1}{1.618} = 0.618 = \tau_1$$

ואז אינטרוולי אי הודאות הינם - $\varepsilon_k \cong [\tau_1]^{k-1} \cdot C$ -עובדה המעידה כי אלגוריתם זה נותן התכנסות ליניארית ויחס ההתכנסות הוא 0.618. ■

עבור N גדול מאוד מקובל לותר על השימוש בסידרת פיבונצ'י ולהשתמש ישירות ביחס הזהב - 0.618. במקרה זה האלגוריתם קרוי אלגוריתם יחס הזהב, ותכונותיו זהות לאלו של חיפוש פיבונצ'י.

5.3 התאמת עקומים - גישת ניוטון

בנוסף להנחות אשר קבענו בתחילת הדיון בחיפוש הזהב (חיפוש באינטרוול סופי, וחד-אופניות), נראה כי יש קבוצה גדולה למדי של פונקציות אשר מעבר לכך מקיימות סוג כלשהו של חלקות אותה נוכל לנצל לצרכי חיפוש יעיל. השיטות אשר יתוארו בסעיף זה ובבאים אחריו עושות שימוש בהנחה נוספת זו, ובאופן זה מוצעות מגוון גישות חדשות. ההבדל המהותי בין הגישות הוא באמצעים אשר אנו מעמידים לרשותן - חלקן לא יעשה שימוש בנגזרות כלל, חלקן בנגזרות מסדר ראשון וחלקן בנגזרות מסדר שני.

נתונה פונקציה f 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 :

$$q'(x_{k+1}) = 0 = f'(x_k) + f''(x_k)(x_{k+1} - x_k) \quad \Rightarrow \quad x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

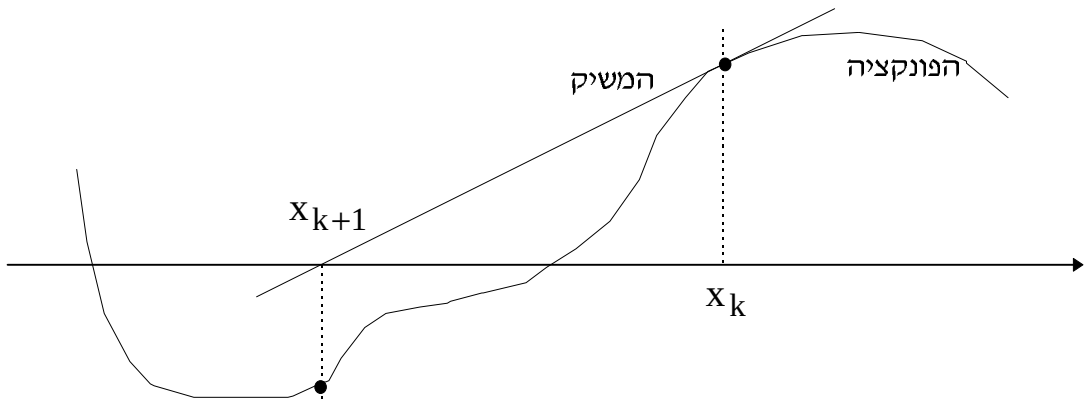
סיכום האלגוריתם מובא בציור 5.8. יש לשים לב לעובדה שלערך הפונקציה במיקום x_k אין השפעה על קביעת המיקום הבא. כן יש לשים לב לעובדה שאלגוריתם זה אשר במקורו נועד לאיתור נקודת מינימום של פונקציה $f(x)$, יכול להתפרש כאלגוריתם לחיפוש פתרון של המשוואה $q'(x) = 0$. נראה זאת באופן הבא. בהינתן פונקציה $g(x)$ רצוננו 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 המתואר בציור 5.9 קרוי אלגוריתם ניוטון-רפסון, והוא אינו אלא אלגוריתם ניוטון המוצג כאן. השוני הוא התפיסה - בעוד שאלגוריתם ניוטון חושב במונחים של מינימיזציה, הרי שאלגוריתם ניוטון-רפסון חושב במונחים של פתרון משוואה - שתי בעיות שקולות, כיוון שנקודת המינימום של הבעיה הראשונה היא נקודת ההתאפסות של נגזרת הפונקציה.

1. התחל בנקודה שרירותית x_0. חשב את ערך שתי נגזרותיה הראשונות של הפונקציה בנקודה זו - $f'(x_0), f''(x_0)$.
2. באיטרציה ה- k - בה אנו בנקודה - x_k - ועבורה ידועים שתי נגזרותיה מסדר ראשון ושני - $f'(x_k), f''(x_k)$, חשב את הנקודה הבאה ע"י - $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$.

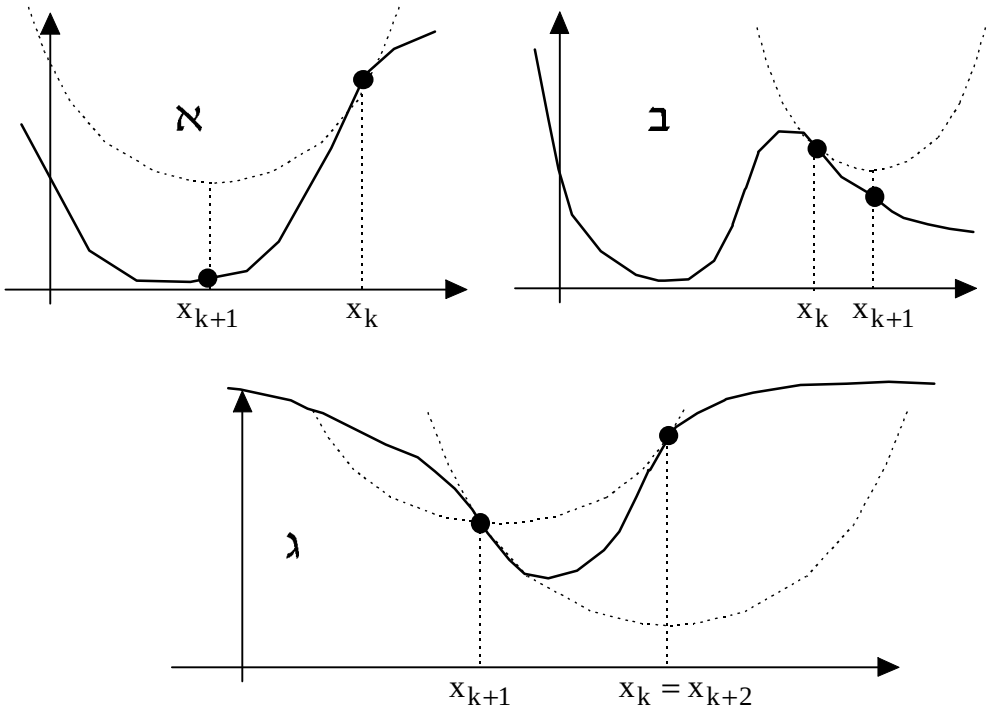
ציור 5.8 - אלגוריתם ניוטון לחיפוש על ישר



ציור 5.9 - המחשת אלגוריתם ניוטון-רפסון לפתרון משוואה

דוגמה 5.4: בציור 5.7 מתוארות שלוש פונקציות והפעלת איטרציה אחת עליהן בגישת ניוטון. כפי שרואים, בחלק א' הפונקציה חד-אופנית וניכר כי האלגוריתם יתכנס אל יעדו. בחלק ב' הפונקציה אינה חד אופנית ותחול התבדרות. בחלק ג' מתקבלת התנדנדות בין שתי נקודות ללא התכנסות למרות שהפונקציה חד-אופנית.

אין למעשה כל הבטחה שאלגוריתם ניוטון יתכנס, ומהדוגמה הנ"ל גם ברור כי הדבר צפוי לקרות לעיתים קרובות - עובדה לא כל כך מעודדת מבחינתנו. אם הנגזרת השניה שלילית בנקודה אליה הגענו - אנו בצרות! עם זאת, אם אנו מצויים בקירבת נקודת המינימום האמיתית, והפונקציה חלקה דיה, בה המשפט הבא לטעון כי תחול התכנסות בקצב גבוה - סדר 2.



ציור 5.7 - אלגוריתם ניוטון לחיפוש על ישר - דוגמאות

משפט 5.4: תהי הפונקציה f בעלת נגזרות רציפות עד סדר שלישי. תהי x^* נקודה בה הנגזרת הראשונה מתאפסת, והנגזרת השניה גדולה מאפס כלשהי - $f''(x^*) > 0$, $f'(x^*) = 0$. אזי, אם נקודה x_0 קרובה דייה לנקודה x^*, סידרת הנקודות שמתקבלת מאלגוריתם ניוטון - $\{x_k\}_{k=1}^{\infty}$ - תתכנס ל- x^* בסדר התכנסות 2 (דהיינו - כמו הסידרה) $2^{-(2^k)}$.

הוכחה: נוכל לכתוב כי -

$$x_{k+1} - x^* = x_k - x^* - \frac{f'(x_k) - f'(x^*)}{f''(x_k)} = -\frac{f'(x_k) - f'(x^*) - f''(x_k)(x_k - x^*)}{f''(x_k)}$$

וממשפט טיילור נוכל לפתח את $f'(x^*)$ לטור סביב הנקודה x_k, ולקבוע כי קיים $\theta \in [0,1]$ כך ש:

$$f'(x^*) = f'(x_k) + f''(x_k)(x^* - x_k) + \frac{1}{2} f'''(\theta x_k + (1-\theta)x^*)(x_k - x^*)^2$$

$$\Rightarrow f'(x_k) - f'(x^*) - f''(x_k)(x_k - x^*) = -\frac{1}{2} f'''(\theta x_k + (1-\theta)x^*)(x_k - x^*)^2$$

לכן מתקבל ע"י הצבת הקשר האחרון בקודמו כי:

$$x_{k+1}-x^*=-\frac{f'(x_k)-f'(x^*)-f''(x_k)(x_k-x^*)}{f''(x_k)}=\frac{f'''(\theta x_k+(1-\theta)x^*)}{2f''(x_k)}(x_k-x^*)^2$$

בשל רציפות הפונקציה ונגזרותיה עד סדר שלישי נוכל לקבוע כי בשל קירבת x_k ל- x^*:

$$\exists 0<k_1<\infty \quad \left|f'''(\theta x_k+(1-\theta)x^*)\right|<k_1$$
$$\exists 0<k_2<\infty \quad \left|f''(x_k)\right|>k_2>0$$

ולכן:

$$\left|\frac{(x_{k+1}-x^*)}{(x_k-x^*)^2}\right|=\frac{f'''(\theta x_k+(1-\theta)x^*)}{2f''(x_k)}<\frac{k_1}{2k_2}$$

אשר שקול לאמירה - קצב התכנסות מסדר שני. ■

ניתן להוסיף לאלגוריתם ניוטון מרכיב אשר יקבע בכל איטרציה את האינטרוול אשר בו מובטח כי תוכל נקודת המינימום, בהנחת יוני-מודליות. מכל הנקודות אשר נאספו עד כה, נבחר את הנקודה הקטנה ביותר אשר הניבה כיוון ירידה (לפי נגזרת ראשונה), ואת הנקודה הגדולה ביותר אשר הניבה עליה. לא צריך כמובן לזכור את כל נקודות העבר. אם האינטרוול באיטרציה ה- k (לפני חישוב הנקודה x_{k+1}) הוא $\left[x_{min}^k, x_{max}^k\right]$ אזי עדכונו ייעשה לפי:

$$\left[x_{min}^k, x_{max}^k\right]=\begin{cases}\left[\max\left\{x_{min}^k, x_k\right\}, x_{max}^k\right] & x_{k+1}>x_k\\ \left[x_{min}^k, \min\left\{x_{max}^k, x_k\right\}\right] & x_{k+1}<x_k\end{cases}$$

ולמעשה המשפט הנ"ל קובע כי אינטרוול זה מתכווץ בקצב מסדר שני, אם אנו בקרבת נקודת המינימום.

5.4 התאמת עקומים - גישת האינטרפולציה

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:

$$q(x)=f(x_k)+f'(x_k)(x-x_k)+\frac{1}{2}\cdot\frac{f'(x_k)-f'(x_{k-1})}{x_k-x_{k-1}}(x-x_k)^2$$

ומכיוון שהביטוי $\frac{f'(x_k)-f'(x_{k-1})}{x_k-x_{k-1}}$ אינו אלא קירוב נומרי לנגזרת מסדר שני, ניכר הדמיון בין הגישה הנוכחית, ואלגוריתם ניוטון. 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. סיכום האלגוריתם נתון להלן:

1. התחל בשתי נקודות שרירותיות x_0,x_1. חשב את ערך נגזרותיה מסדר ראשון של הפונקציה בנקודות אלו - $f'(x_0),f'(x_1)$.
2. באיטרציה ה- k - בה אנו בנקודה - x_k - ועבורה ידועים נגזרתה ונגזרת קודמתה מסדר ראשון - $f'(x_k),f'(x_{k-1})$, חשב את הנקודה הבאה ע"י - $x_{k+1}=x_k-\frac{(x_k-x_{k-1})\cdot f'(x_k)}{f'(x_k)-f'(x_{k-1})}$.

<u>ציור 5.10 - אלגוריתם האינטרפולציה לחיפוש על ישר</u>

ניתן להוכיח כי קצב ההתכנסות של אלגוריתם זה הוא מסדר 1.618 (כן! יש קשר לסידרת פיבונצ'י). בדומה לאלגוריתם ניוטון, אלגוריתם זה עלול להתבדר, לשוטט ללא התכנסות או להתנדנד - תלוי בבחירת נקודת ההתחלה, והפונקציה עליה אנו עובדים. לא ניתן כמובן בשל כך לטעון להתכנסות גלובלית. בדומה לאלגוריתם ניוטון, גם כאן ניתן לקדם אינטרוול בו צריך להימצא המינימום. למעשה, דרך קידומו של אינטרוול זה זהה לנאמר קודם.

נראה עקרונית כיצד ניתן לקבל סדר התכנסות 1.618 (זו אינה הוכחה!). ובכן, מסתבר לאחר פיתוח אלגברי מייגע במקצת שקיימת המשוואה הבאה לשגיאת האלגוריתם בזמן:

$$\varepsilon_k = x_k - x^* \quad \Rightarrow \quad \varepsilon_k = \varepsilon_{k-1} \cdot \varepsilon_{k-2}$$

ע"י פעולת Log על שני האגפים נקבל - $L_\varepsilon(k) = L_\varepsilon(k-1) + L_\varepsilon(k-2)$. זאת משוואת הפרש רקורסיבית זהה לסידרת פיבונצי, ולכן ברור שהגודל $L_\varepsilon(k)$ מתכנס ליניארית ביחס 0.618, דהיינו כמו - 0.618^k. לכן, השגיאה עצמה ε_k מתכנסת כמו $\exp\left\{0.618^k\right\}$.

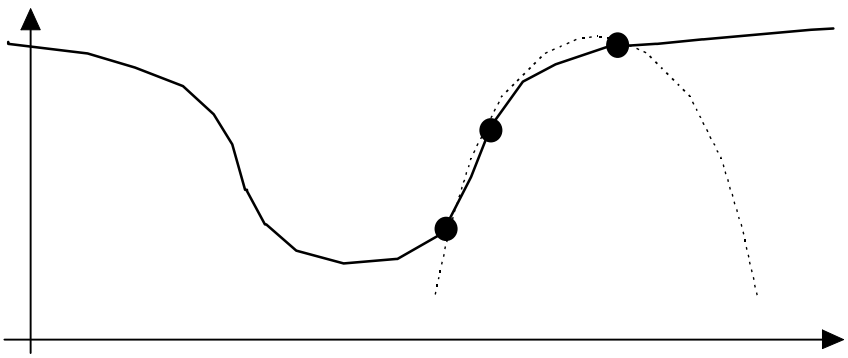
5.5 התאמת עקומים - גישת ההתאמה הריבועית

על פי אותו קו, נוכל להציע ויתור על נגזרות ושימוש בערכי הפונקציה בלבד - בדומה לגישה שננקטה בחיפוש הזהב. עבור שלושת הנקודות האחרונות x_{k-2}, x_{k-1}, x_k, בידיעת הערכים של הפונקציה בנקודות אלה נקבל כי הפונקציה הריבועית אשר עוברת דרכם היא:

$$\begin{aligned} q(x) = f(x_k)\frac{(x-x_{k-1})(x-x_{k-2})}{(x_k-x_{k-1})(x_k-x_{k-2})} + \\ + f(x_{k-1})\frac{(x-x_k)(x-x_{k-2})}{(x_{k-1}-x_k)(x_{k-1}-x_{k-2})} + \\ + f(x_{k-2})\frac{(x-x_k)(x-x_{k-1})}{(x_{k-2}-x_k)(x_{k-2}-x_{k-1})} \end{aligned}$$

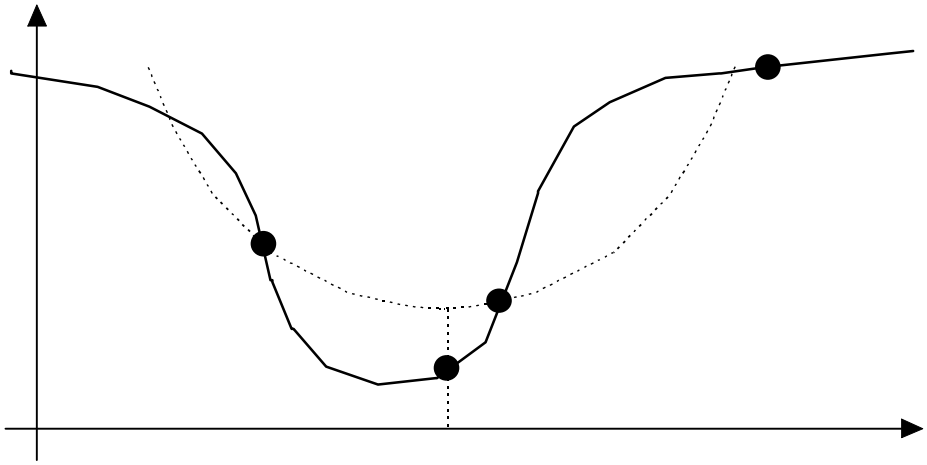
וניתן (ואנו לא נטרח לעשות זאת) לכתוב את הביטוי אשר מייצג את נקודות המינימום של פונקציה זו. אם האיתחול של גישה זו נבחר להיות קרוב לנקודת המינימום, ניתן להראות כי אלגוריתם זה מתכנס בקצב מסדר 1.3.

הבעיה המרכזית של אלגוריתם זה היא שבחירה כלשהי של נקודות עלולה לתת כי הפרבולה המקרבת אינה בעלת מינימום. לשם המחשת אפשרות זו ראה ציור 5.11.



ציור 5.11 - חיפוש על ישר בגישת ההתאמה הריבועית - בעיה אופיינית

על מנת לפתור בעיה זו נציע את האלגוריתם הבא שהינו וריאציה של התפיסה שהוצגה לעיל. בוריאציה זו אנו מקדמים בכל איטרציה שלוש נקודות בעלות מבנה המחייב קיומו של מינימום. יתרה מכך - תמיד יתקבל כי נקודת המינימום האמיתית תהיה בתוך האינטרוול המוגדר מקצוות השלישיה הנתונה, וכיוון שגודל זה דועך מאיטרציה לאיטרציה, תתקבל התכנסות לכל איתחול (דהיינו - התכנסות גלובלית).



ציור 5.12 - ההתאמה הריבועית בגישה המשופרת - דוגמה

1. התחל בשלוש נקודות שרירותיות $x_0 < x_1 < x_2$, אשר יקיימו כי $f(x_0) \geq f(x_1) \leq f(x_2)$ (יש דרכים למציאת שלישיה כזו). חשב את ערך הפונקציה בנקודות אלו.
2. נניח כי אנו באיטרציה ה- k עבורה יש בידנו שלישיית נקודות $x_0 < x_1 < x_2$ אשר בדומה לקודם מקיימות כי $f(x_0) \geq f(x_1) \leq f(x_2)$.
3. חשב את הנקודה הבאה ע"י הבאת $q(x)$ כפי שהוגדר קודם למינימום, וקבל את x^*.
4. חשב את $f(x^*)$, ומתוך ארבעת הנקודות הנתונות $x_0 < x_1 < x_2$ ו- x^* בחר שלישיה $x'_0 < x'_1 < x'_2$ אשר תכלול את x^* ואשר תהיה במבנה של פונקציה ריבועית עם מינימום.

ציור 5.13 - אלגוריתם ההתאמה הריבועית המתוקן לחיפוש על ישר

5.6 קריטריוני עצירה לאלגוריתמי החיפוש על ישר, וחיפוש מקורב

בדרך כלל לא נרצה לבצע את האלגוריתמים שתוארו לעיל באינסוף איטרציות על מנת להבטיח התכנסות לנקודה המדוייקת. ברור כי נרצה לעצור את התהליך בנקודה כלשהי ולהסתפק בה, בידיעה שתוצאת המינימיזציה החד-מימדית הינה קירוב בלבד. השאלה הנשאלת היא כיצד נקבע את נקודת העצירה?

גישה מומלצת אחת היא קביעת סף הסטיה מנקודת היעד באופן יחסי. נניח כי הפרמטר α אמור להתכנס לערך מדויק α_∞. אזי, קביעת ערך סף $c<<1$ וקבלת כל פתרון המקיים כי $\left|\alpha-\alpha_\infty\right|/\alpha_\infty \le c$ היא דרך יעילה המבטיחה כי אנו קרובים מאוד לפתרון. בכל הגישות שהזכרנו קודם קיבלנו כפועל יוצא אינטרוול המקודם ומוקטן בזמן ובו צריכה להימצא נקודת המינימום. נוכל לכן לומר כי באיטרציה ה- k:

$$\alpha_\infty \in \left[\alpha_k^{min},\alpha_k^{max}\right] \;\Rightarrow\; \frac{\left|\alpha-\alpha_\infty\right|}{\alpha_\infty} \le \frac{\alpha_k^{max}-\alpha_k^{min}}{\alpha_k^{min}} \le c$$

כך שעל-פי אינטרוול אי-האמינות נקבל חסם עליון לשגיאת השערוך היחסית.

ברור כי לגישה הזו חולשה במקרים בהם נקודת המינימום של אינטרוול אי-האמינות קרובה ל- 0 - עובדה שאינה מעידה בהכרח על בעיית התכנסות. לכן, כתחליף ניתן להציע כקריטריון עצירה את גודל אינטרוול זה (ללא יחסיות).

ישנן דרכים נוספות לקביעת העצירה של אלגוריתמים אלו - לא נסקור אותן במסגרת זו. חשוב לציין עם זאת את העובדה החשובה הבאה: ביצוע חיפוש מקורב על ישר - הן בשל התיחסות לאינטרוול סגור חלקי לתחום בו נדרשת מונימיזציה, והן בשל מציאת נקודה שאינה מינימום לוקלי - הוא עדיין מיפוי סגור, ולכן שימוש בו לא מסכן התכנסותם של אלגוריתמים (דורש הוכחה).

5.7 סיכום

בפרק זה הצגנו משפחה של אלגוריתמים אשר מבצעים מינימיזציה רבת מימדים ע"י פתרון איטרטיבי בו יש מינימיזציה על ישר. הצגנו מספר דרכים שונות ומגוונות לחיפוש על ישר - חלקן בעלות יכולת התכנסות גלובלית, וחלקן האחר עם בעיות בהיבט זה. ניתן להציע דרכים מגוונות נוספות, ושילובים שונים של אלגוריתמים לקבלת כלי חיפוש חזקים על ישר. אנו נניח לצורך ניתוח אלגוריתמים כוללים כי מופעל חיפוש מדויק על ישר - בידיעה ברורה שמותר לחרוג ולהציע חיפוש מקורב. בציור 5.12 מובאת טבלה בה השוואה בין השיטות שהועלו במסגרת פרק זה במגוון היבטים.

לא ניתן לקבוע מי מהשיטות טובה יותר - שיפוט מעיין זה מחייב ידיעה נוספת על הפונקציה עליה אנו מבצעים מינימיזציה, שכן חישוב נגזרות יכול להיות פשוט או מורכב ביותר.

	חיפוש הזהב	גישת התאמת עקום - ניוטון	שיטת האינטרפולציה	שיטת ההתאמה הריבועית
סדר התכנסות האלגוריתם	ליניארי עם יחס 0.618	2	1.618	1.31
חישוב ערכי הפונקציה	כן	לא	לא	כן
חישוב נגזרת ראשונה	לא	כן	כן	לא
חישוב נגזרת שניה	לא	כן	לא	לא
התכנסות גלובלית	כן	לא	לא	כן
הערות	נדרש לדעת את מספר הצעדים מראש	ניתן להציע גירסה בעלת התכנסות גלובלית	ניתן להציע גירסה בעלת התכנסות גלובלית	התכונות הנ"ל מתיחסות לאלגוריתם המשופר

ציור 5.14 - טבלת השוואה בין שיטות החיפוש המדויקות על ישר

פרק 6 - אלגוריתמי מינימיזציה ללא חישוב נגזרות

בדומה לגישה שננקטה בטיפול בבעיות חד-מימדיות, גם בבעיות רב מימדיות ניתן להרכיב אלגוריתמים אשר ללא חישוב גרדיאנט רב-מימדי מבצעים מינימיזציה. אנו נציג מספר גישות ואלגוריתמים כאלה במסגרת פרק זה. עיקר תרומתם תקפה במצבים בהם לא ניתן או קשה לחשב נגזרות באופן ישיר.

6.1 אלגוריתם דעיכה על קואורדינטות

תהי נתונה פונקציה בת n נעלמים - $[x_1, x_2, \ldots, x_n]$ - אותה רצוננו להביא למינימום. בהינתן נקודה בה אנו נמצאים $[x_1^0, x_2^0, \ldots, x_n^0]$ מציעה גישת הדעיכה על הקואורדינטות למצוא מינימום לפי אחת הקואורדינטות (נניח x_1) וזאת ע"י הפעלת אלגוריתם חיפוש על ישר. וכך בכל איטרציה תיבחר קואורדינטה אחרת עליה יבוצע מינימום. יש לשים לב לכך שבגישה זו החיפוש על הישר אינו על קרן אלא על ישר דו-כיווני כי לא ידוע כיוון הירידה (שהרי אין חישובי גרדיאנט). למעשה ניתן לראות אלגוריתם זה כאלגוריתם ירידה בו בחירת הכיוונים היא הבסיס הטריוויאלי e_i.

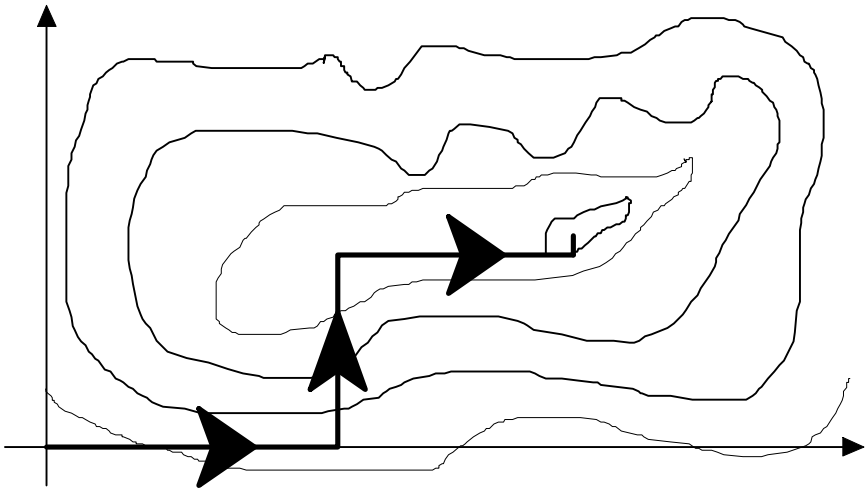
תפיסה אחת אשר הוצעה למימוש אלגוריתם זה הציעה סריקת קואורדינטות ציקלית - מ-1 ועד n וחוזר חלילה. גישה אחרת (Aitken Double Sweep Method) הציעה גישה בה מתחילים מ-1 ועולים עד n ובחזרה מ- n-1 ועד 1, וחוזר חלילה.

אם ניתן לחשב את הגרדיאנט של הפונקציה ניתן לקבוע את סדר הקואורדינטות לפי גודל הנגזרות החלקיות, ולבחור את זו בעלת הגודל הגדול ביותר, ואף לקבוע את סימן הכיוון לפיה. גישה זו קרויה Gauss-Southwell Method.

ככלל ביצועי של גישת דעיכה על קואורדינטות ירודים. עם זאת היא פשוטה למימוש ולהבנה. ניתן להראות כי אלגוריתם זה הוא בעל התכנסות גלובלית, כיוון שכל מרכיביו רציפים, וקיימת פונקצית דעיכה אותה הוא מממש.

6.2 אלגוריתם ה- Simplex הכללי

אלגוריתם אחר אשר אותו נכיר באופן יותר מפורט כשנדון בבעיות תכנות ליניארי (שהינן בעיות עם אילוצים) הוא אלגוריתם ה- Simplex. יש למעשה מגוון רחב של וריאציות לאלגוריתם זה - אנו מציג אחת מהן. משותף לכל וריאציות אלו ביסוס חזק על אינטואיציה, ואי-יכולת להוכיח התכנסות למעט עבור מקרים פרטיים פשוטים (כגון בעיית התכנות הליניארי בה נתעמק בפרקים מתקדמים). למרות זאת, גישת ה- Simplex מקובלת לשימוש וקלה ליישום.



ציור 6.1 - המחשת שיטת הדעיכה על הקואורדינטות

1. אתחל את האלגוריתם בנקודה שרירותית $x_0 = \left\{x_0^1, x_0^2, \cdots, x_0^n\right\}$.
2. נניח כי אנו באיטרציה ה- k (עבור האיטרציות k=1 עד k=n):
   מצא את נקודת המינימום γ_{opt} של הישר $f\left(x_{k-1}^1, x_{k-1}^2, \cdots, x_{k-1}^{k-1}, \gamma, x_{k-1}^{k+1}, \cdots, x_{k-1}^n\right)$ וקבע את הנקודה הבאה להיות -
   $$x_k = \left\{x_{k-1}^1, x_{k-1}^2, \cdots, x_{k-1}^{k-1}, \gamma_{opt}, x_{k-1}^{k+1}, \cdots, x_{k-1}^n\right\}$$
3. עבור האיטרציות k=n+1 ועד k=2n-1, בצע באופן דומה כשקואורדינטות המינימיזציה מתחילים מהסוף.
4. עצור את האלגוריתם לאחר מכפלה שלמה של 2n-1 איטרציות, כאשר ביחס לתוצאת הסבב הקודם חל שינוי קטן מסף מסויים.

ציור 6.2 - אלגוריתם Aitken Double Sweep Method

הגדרה 6.1: במרחב E^n, Simplex הוא פוליגון בעל n+1 נקודות- $\{x_k\}_{k=0}^{n}$. Simplex מוגדר כחוקי אם נפחו אינו אפס. מרכזו של Simplex מוגדר כ- $x_{center} = (n+1)^{-1}\sum_{k=0}^{n} x_k$.

הרעיון הבסיסי של שיטת ה- Simplex הוא לקבוע פוליגון כזה התחלי (אשר אם הוא כולל את נקודת המינימום הרב-מימדי זהו יתרון), ולכווץ עוד ועוד את הפוליגון, כששומרים על חוקיותו, וכשהוא כולל בכל שלב את נקודת המינימום. באופן זה מתכנסים לנקודת המינימום. מקובל להתיחס לנקודת מרכז ה- Simplex כפתרון הביניים בסיום כל איטרציה בה עודכן ה- Simplex. אחד מאלגוריתמי ה- Simplex מתואר בציור 6.3.

הגדרה 6.2: בהינתן Simplex במרחב E^n בעל הנקודות - $\{x_k\}_{k=0}^{n}$, והפונקציה f(x), נגדיר את שני הגדלים הבאים:

1. אתחל את האלגוריתם ע"י קביעת Simplex (אם יכיל את נקודת המינימום זה יתרון).
2. באיטרציה החדשה - חשב את f_{max}, x_{max}, f_{min}, x_{min} ואת $\hat{x}$.
3. חשב את נקודת השיקוף - $x_{ref} = 2\hat{x} - x_{max}$
4. חשב את הנקודה x_{new} אשר תחליף את x_{max} לפי:

$$x_{new} = \begin{cases} \begin{cases} 2x_{ref} - \hat{x} & f(2x_{ref} - \hat{x}) < f(x_{ref}) \\ x_{ref} & \text{otherwise} \end{cases} & f(x_{min}) > f(x_{ref}) \\ x_{ref} & \text{otherwise} \\ \begin{cases} [x_{max} + \hat{x}]/2 & f(x_{max}) \le f(x_{ref}) \\ [x_{ref} + \hat{x}]/2 & \text{otherwise} \end{cases} & \max_j \{ f(x_j) \mid x_j \ne x_{max} \} \le f(x_{ref}) \end{cases}$$

5. לאחר ההחלפה התקבל Simplex חדש.
6. עצור את האלגוריתם כאשר $1 - f_{min}/f_{max} \le \varepsilon$

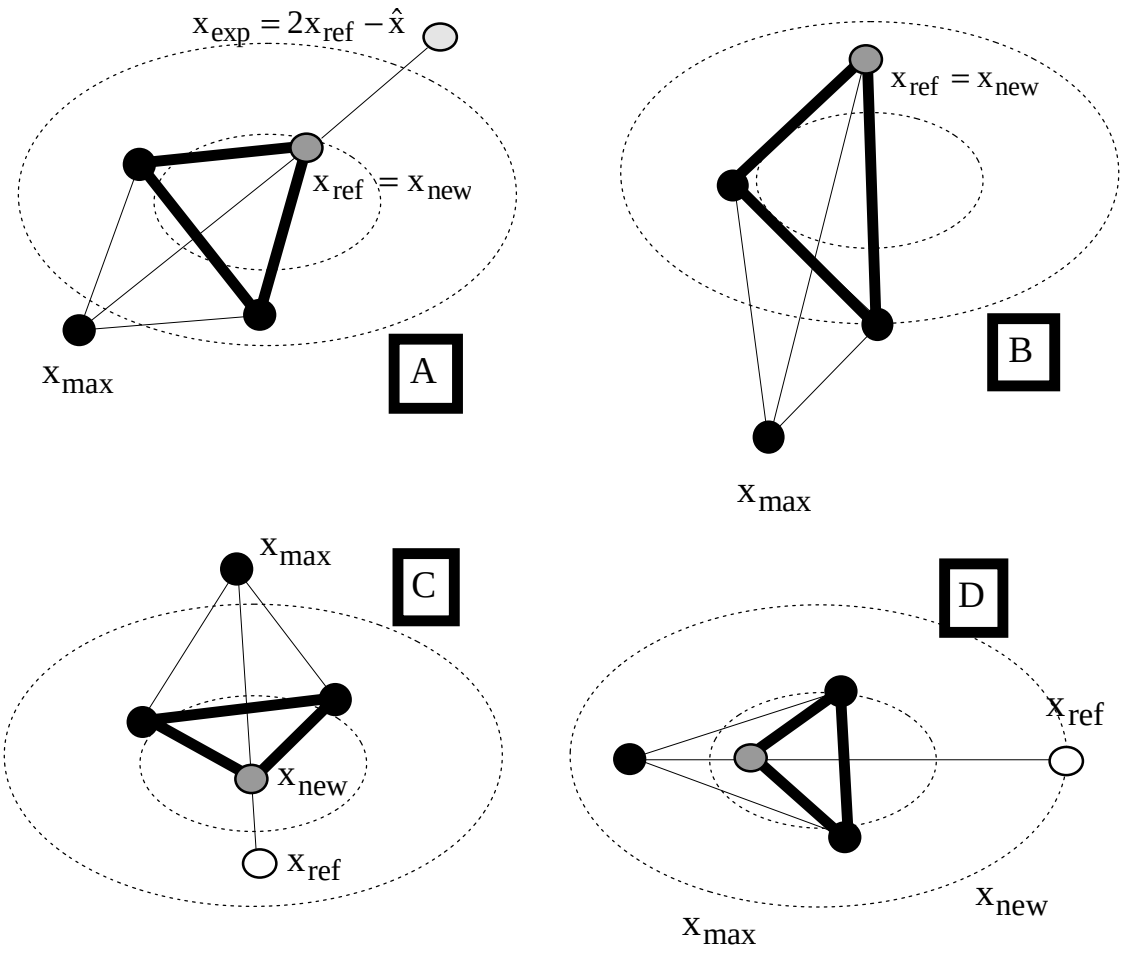
<u>ציור 6.3 - אלגוריתם ה- Simplex הכללי</u>

$$x_{min} = \arg\min_{x_j}\left\{f\left(x_j\right)\right\} \quad ; \quad f_{min} = \min_{x_j}\left\{f\left(x_j\right)\right\}$$

$$x_{max} = \arg\max_{x_j}\left\{f\left(x_j\right)\right\} \quad ; \quad f_{max} = \max_{x_j}\left\{f\left(x_j\right)\right\}$$

$$\hat{x} = \frac{1}{n}\left[\sum_{j=0}^{n} x_j - x_{max}\right]$$

בציור 6.4 מתוארים מצבים שונים של עדכון ה- Simplex. מתוך ציורים אלו גם ברור ההגיון בבחירת הנקודה המחליפה. בכל המצבים המתוארים אנו רואים כי המשולש (simplex) החדש קרוב יותר לנקודת המינימום מבחינת נקודת מרכזו.



ציור 6.4 - מגוון צורות ההתקדמות של ה- Simplex - A, B, C ו- D לפי סדר הופעתם באלגוריתם (המשולש העבה הוא ה- Simplex החדש)

6.3 קירובי נגזרות

מלכתחילה, הסיבות האפשריות להימנעות מחישוב אנליטי של הנגזרות יכולות להיות ש- 1) f(x) היא "קופסא שחורה" ולכן אי-אפשר לחשב נגזרת, 2) f(x) מסובכת לגזירה בשל המבנה שלה, או 3) כמות החישובים בנגזרת האנליטית גבוהה מאוד. בדומה לדרך בה הדבר נעשה באלגוריתמי החיפוש על ישר, אנו יכולים לקרב את חישוב הנגזרת ע"י חישוב ערכי הפונקציה ושימוש בהם. פירוש הדבר שנוכל בפועל לעשות שימוש בכל האלגוריתמים שיוצגו בהמשך, ואשר דורשים שימוש בגרדיאנט ו\או בהסיין של פונקציה.

לצורך שערוך וקטור הגרדיאנט של פונקציה ניתן להציע את אחת משתי הדרכים הבאות - הגישה החד-צדדית או הגישה הדו-צדדית:

$$\frac{\partial f(x)}{\partial x_j} \approx \frac{1}{\delta}\left[f\left(x+\delta e_j\right)-f(x)\right] \qquad \frac{\partial f(x)}{\partial x_j} \approx \frac{1}{2\delta}\left[f\left(x+\delta e_j\right)-f\left(x-\delta e_j\right)\right]$$

כאשר x_j היא הקואורדינטה ה- j של הוקטור x, ו- e_j הוא הכיוון בקואורדינטה זו. הגישה הדו-צדדית דורשת יותר חישובים כיון שיש לחשב את ערך הפונקציה ב- 2n מקומות (בהשוואה ל- n+1 מקומות בגישה הראשונה), אך גישה זו מדוייקת יותר. נראה זאת ע"י פיתוח לטור טיילור:

$$\frac{1}{\delta}\left[f\left(x+\delta e_j\right)-f(x)\right]=\frac{\partial f(x)}{\partial x}e_j+\frac{1}{\delta}o\{\delta\}$$
$$\frac{1}{2\delta}\left[f\left(x+\delta e_j\right)-f\left(x-\delta e_j\right)\right]=\frac{\partial f(x)}{\partial x}e_j+\frac{1}{2\delta}o\left\{\delta^2\right\}$$

כיוון ש:

$$f\left(x+\delta e_j\right)=f(x)+\frac{\partial f(x)}{\partial x}e_j\delta+\frac{1}{2}\delta^2 e_j^T\frac{\partial^2 f(x)}{\partial x^2}e_j+o\left\{\delta^2\right\}$$
$$f\left(x-\delta e_j\right)=f(x)-\frac{\partial f(x)}{\partial x}e_j\delta+\frac{1}{2}\delta^2 e_j^T\frac{\partial^2 f(x)}{\partial x^2}e_j+o\left\{\delta^2\right\}$$

נזכור כי לפי הגדרה - $\lim_{\alpha\to 0}\frac{o\{\alpha\}}{\alpha}=0$. הנגזרת השניה בדרך השניה נופלת בשל הופעתה פעמיים ובסימן הפוך. לכן מתקבל דיוק טוב יותר בגישה הדו-צדדית. יתרון נוסף של הגישה הדו-צדדית היא יכולתה לספק בקלות גם את ערכי האלכסון הראשי של מטריצת ההסיין כיוון שערכי הפונקציה שחושבו משמשים ישירות לפי:

$$\frac{\partial^2 f(x)}{\partial x_j^2} \approx \frac{1}{\delta^2}\left[f\left(x+\delta e_j\right)+f\left(x-\delta e_j\right)-2f(x)\right]$$

וכך, רק ע"י הוספת חישוב ערך הפונקציה בנקודת המרכז x, אנו מקבלים מין המוכן את כל האלכסון הראשי של ההסיין. ובמקרה הכללי ניתן לקרב את הנגזרות השניות לפי הביטויים הבאים (שוב שתי הגישות - החד-צדדית והדו-צדדית):

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \approx \frac{1}{\delta}\left[\frac{\partial f\left(x+\delta e_j\right)}{\partial x_i}-\frac{\partial f(x)}{\partial x_i}\right]=\frac{1}{\delta^2}\left[f\left(x+\delta e_i+\delta e_j\right)-f\left(x+\delta e_j\right)-f\left(x+\delta e_i\right)+f(x)\right]$$

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \approx \frac{1}{2\delta}\left[\frac{\partial f\left(x+\delta e_j\right)}{\partial x_i}-\frac{\partial f\left(x-\delta e_j\right)}{\partial x_i}\right]=\frac{1}{4\delta^2}\left[f\left(x+\delta e_j+\delta e_i\right)-f\left(x+\delta e_j-\delta e_i\right)-\right.$$
$$\left.-f\left(x-\delta e_j+\delta e_i\right)+f\left(x-\delta e_j-\delta e_i\right)\right]$$

המכשלה המרכזית בגישה זו של קירוב הנגזרות היא עניין דיוק השערוך, והשפעת השגיאות בהיבט זה על האלגוריתמים המיושמים. ברור כי רצויה בחירה של δ קטנה ככל האפשר, אך מאידך יש להזהר לא לבחור סטיה כה קטנה אשר קרובה לשגיאות הנומריות בחישוב באורך מילה סופי במחשב, משום שאז יתקבלו תוצאות שערוך שגויות לחלוטין. בהרבה מקרים אף לא ניתן לקבוע ערך אחד של δ אשר יתאים לשערוך כל הנגזרות, וזאת בשל ה- Scale השונה של הפונקציה המטופלת בכל ציר. באשר לטיפול בבעיות אלו - הגישה המומלצת היא ניסוי וטעיה, וישנם כללי אצבע מגוונים אשר נועדו לקבוע את הפרמטר δ באופן שימנע שגיאות מתבדרות.

הערה: הסוגיה אשר העלנו כאן - עניין בחירת גודל הסטיה δ נדון בהרחבה באנליזה נומרית במסגרת הדיון בפתרון נומרי של משוואות דיפרנציאליות חלקיות.

6.4 סיכום

בפרק זה הצגנו דרכים לעקוף את הצורך לחשב נגזרות של פונקציה על מנת לבצע עליה מינימיזציה. ישנן דרכים ישירות - כלומר אלגוריתמים אשר פועלים ללא חישוב גרדיאנטים, אך אלגוריתמים אלו מאופיינים ע"י ביצועים ירודים בד"כ. דרכים עקיפות למניעת הצורך בחישוב נגזרות היא שערוכם מתוך ערכי הפונקציה - גישה נומרית ידועה בהקשרים אחרים וקרויה שיטת הפרשים סופיים.

פרק 7 - אלגוריתמי מינימיזציה מבוססי גרדיאנט

בפרק זה נציג את האלגוריתמים המוכרים ביותר באופטימיזציה - אלגוריתם השיפוע המירבי ואלגוריתם ניוטון. שני אלגוריתמים אלו מבוססים על חישובו של הגרדיאנט ובשל כך ביצועיהם טובים משל האלגוריתמים אשר הוצגו בפרק הקודם. מכיוון ששני אלגוריתמים אלו הם בעלי בסיס תיאורתי משותף, נציג קו מרכזי בו שני אלגוריתמים אלו מהוים מקרה פרטי.

7.1 אלגוריתמי מינימיזציה מבוססי גרדיאנט

נניח כי האלגוריתם למינימיזציה יבחר את כיוון התנועה d להיות מכפלה של מטריצה חיובית מוגדרת וסימטרית כלשהי במינוס וקטור הגרדיאנט, דהיינו:

$$d_k = -\mathbf{M}_k \nabla f^T(x_x) \quad \mathbf{M}_k > 0, \ \mathbf{M}_k = \mathbf{M}_k^T$$
$$x_{k+1} = x_k - \alpha_k \mathbf{M}_k \nabla f(x_x)$$

מוטיבציה לבחירה זו באה מכך שוקטור הגרדיאנט מייצג מבחינה גאומטרית את היטלו של הנורמל למשטח אותו רצוננו להביא למינימום. אינטואיטיבית ניכר כי זהו גם כיוון העליה המירבית בפונקציה. לכן הליכה בכיוון הפוך לו פירושה ירידה בפונקציה - זהו היעד המבוקש. ההכפלה במטריצה חיובית מוגדרת שקולה במקרה הסקלרי למכפלה במספר חיובי, ובמקרה המטריצי פירושה מישקול שונה מעט של מרכיבי הכיוון, ללא פגיעה בכיוון הכללי. יש לזכור כי אם נכתוב את **M** בתלות בוקטורים ובערכים העצמיים שלה נקבל כי הכיוון שייקבע ע"י הכפלה ב- **M** הוא:

$$\mathbf{M} = \sum_{k=1}^{n} \lambda_k u_k u_k^T \Rightarrow \ \mathbf{M}v = \sum_{k=1}^{n} \lambda_k \langle u_k, v \rangle u_k$$

בעוד שהכיוון ללא הכפלה ב- **M** הוא:

$$v = \sum_{k=1}^{n} \langle u_k, v \rangle u_k$$

כשעשינו שימוש בעובדה שהוקטורים העצמיים u_k הם אורתונורמליים. כך שכפי שנאמר, משמעות הכפלה ב- **M** היא מישקול שונה של מרכיבי הכיוון ללא פגיעה בכיוון הכללי.

משפט 7.1: בחירת הכיוון $\mathbf{M}_k > 0$ $d_k = -\mathbf{M}_k \nabla f^T(x_k)$ מבטיחה עבור $\alpha_k > 0$ קטן דיו ירידה בפונקציה f, ועבור בחירת $\mathbf{M}=\mathbf{I}$ ווקטור עדכון בנורמה קבועה קטנה דייה ε^2 מתקבלת הירידה הגדולה ביותר.

הוכחה: לפי נוסחת טיילור נוכל לכתוב כי:

$$f(x) = f(x_k) + \nabla f(x_k)(x - x_k) + o\{\|x - x_k\|\}$$

הצבת הקשר - $x_{k+1} = x_k - \alpha_k \mathbf{M}_k \nabla f(x_x)$ נותנת:

$$f(x_{k+1}) = f(x_k) - \alpha_k \nabla f(x_k) \mathbf{M} \nabla f^T(x_k) + o\{\|\alpha_k \mathbf{M} \nabla f^T(x_k)\|\}$$

ועבור $\alpha_k \to 0$ ברור כי חלה כאן ירידה כיוון ש - $\nabla f(x_k) \mathbf{M} \nabla f^T(x_k) > 0$. בכך הוכחנו את הטענה הראשונה של המשפט.

נורמת וקטור העדכון הכללית היא:

$$\alpha_k^2 \nabla f(x_k) M^T M \nabla f^T(x_k) = \varepsilon^2$$
$$v = \nabla f^T(x_k) \quad , \quad \mathbf{M} = \sum_{j=1}^{j} \lambda_j u_j u_j^T$$
$$\Rightarrow \quad \varepsilon^2 = \alpha_k^2 \cdot \sum_{j=1}^{n} \lambda_j^2 \langle u_j, v \rangle^2$$

עבור נורמה השואפת לאפס הראינו כי מידת הירידה נתונה ע"י:

$$\Delta_k(M, x_k) = \alpha_k \nabla f(x_k) M \nabla f^T(x_k) = \alpha_k \cdot \sum_{j=1}^{n} \lambda_j \langle u_j, v \rangle^2 .$$

לכן, אם נגדיר את הגדלים - $\hat{\lambda}_{k,j} = \alpha_k \lambda_j$, אזי ברצוננו לפתור כאן בעיית מקסימיזציה עם אילוץ מהצורה:

$$\underset{\hat{\lambda}_{k,j}}{\text{Max}}\ \underset{j=1\ldots n}{}\left\{\sum_{j=1}^{n}\hat{\lambda}_{k,j}\left\langle u_j,v\right\rangle^2\right\}\quad \text{s.t.}\quad \varepsilon^2=\sum_{j=1}^{n}\hat{\lambda}_{k,j}^2\left\langle u_j,v\right\rangle^2$$

ובגישת כופלי לגרנז' נקבל כי עלינו להביא למקסימום את הביטוי:

$$\underset{\hat{\lambda}_{k,j}}{\text{Max}}\ \underset{j=1\ldots n}{}\left\{\sum_{j=1}^{n}\hat{\lambda}_{k,j}\left\langle u_j,v\right\rangle^2+\mu\left[\varepsilon^2-\sum_{j=1}^{n}\hat{\lambda}_{k,j}^2\left\langle u_j,v\right\rangle^2\right]\right\}$$

גזירה והשוואה לאפס נותנת את הקשר:

$$\left\langle u_j,v\right\rangle^2\left[1-2\mu\hat{\lambda}_{k,j}\right]=0\quad \text{for}\quad j=1,2,3,\ldots,n\ \Rightarrow\ \hat{\lambda}_{k,j}=\frac{1}{2\mu}$$

כלומר כל הערכים העצמיים צריכים להיות שוים. הצבת פתרון זה באילוץ תיתן כי:

$$\varepsilon^2=\sum_{j=1}^{n}\hat{\lambda}_{k,j}^2\left\langle u_j,v\right\rangle^2=\frac{1}{4\mu^2}\sum_{j=1}^{n}\left\langle u_j,v\right\rangle^2=\frac{1}{4\mu^2}\quad\Rightarrow\quad \mu=\frac{1}{2\varepsilon}\quad\Rightarrow\quad \hat{\lambda}_{k,j}=\varepsilon$$

וכך קיבלנו כי M צריכה להיות מטריצה בעלת ערכים עצמיים זהים - שאינה אלא מטריצת היחידה.

לכן, הכיוון אשר מניב את הירידה המירבית הוא $d_k=-\nabla f^T(x_k)$. ■

עבור הבחירה $\mathbf{M}_k=I$ ושימוש בגודל צעד $\alpha_k=\alpha=\text{const}$ מתקבל אלגוריתם הקרוי אלגוריתם השיפוע המירבי Steepest Descent. עבור חיפוש מדויק על הישר וקביעת α_k האופטימלי בכל איטרציה מתקבל אלגוריתם השיפוע המירבי המנורמל - Normalized SD. עבור בחירת $\mathbf{M}_k=\mathbf{F}^{-1}(x_k)$ (היפוך מטריצת ההסיין בנקודה בה אנו נמצאים) וקביעה $\alpha_k=1$ מתקבל אלגוריתם ניוטון. עבור בחירת $\mathbf{M}_k$ כמטריצה אלכסונית בה כל איבר על האלכסון מכוונן להיות אופטימלי לצעד ה- k קרויה שיטת ה- DNSD (Diagonal NSD).

7.2 הפונקציה הריבועית ושימושיה באופטימיזציה

בכל האלגוריתמים אשר נציג במסגרת קורס זה נתעניין במיוחד בביצועיהם עבור הבעיה הריבועית. נגדיר תחילה בעיה זו, ונסביר לאחר-מכן את המוטיבציה לשימוש בה.

הגדרה 7.1: הבעיה הריבועית מוגדרת כמינימיזציה של הפונקציה הבאה:

$$f(x)=\frac{1}{2}x^T\mathbf{Q}x-x^Tb+c$$

כאשר **Q** היא מטריצה חיובית מוגדרת.

ע"י גזירה והשוואה לאפס נקל לראות כי נקודת המינימום של פונקציה זו נתונה ע"י פתרון מערכת המשוואות $\mathbf{Q}x^*=b$. ישנו תחום רחב של מחקר בו עוסקים בפתרון מערכות משוואות ליניאריות עם מטריצה חיובית מוגדרת - ישום העולה במגוון אפליקציות. פירוש הדבר שכל השיטות אשר תקפות לפתרון מערכות משוואות כאלה (עם מטריצה חיובית מוגדרת **Q**), רלוונטיות גם לנו למינימיזציה של בעיות ריבועיות.

ניתן גם לכתוב את הפונקציה הריבועית בדרך אלטרנטיבית:

$$f(x)=\frac{1}{2}x^T\mathbf{Q}x-x^Tb+c=\frac{1}{2}(x-x^*)^T\mathbf{Q}(x-x^*)-\frac{1}{2}x^{*T}\mathbf{Q}x^*+c$$

ומכיתוב זה ניכר כי משטחים שווי גובה של פונקציה זו הינם אליפסואידים רב-מימדיים. הפונקציה הריבועית היא פונקציה קמורה (רק אם **Q** חיובית מוגדרת! - ראה\י למה 3.4).

דוגמה 7.1: נניח כי רצוננו למדוד וקטור Z של אות מוחי EEG המורכב מ- N ערכים סקלריים. אנו מבצעים את פעולת המדידה L פעמים במיקומים שונים ומקבלים בכל פעם וקטור מדידה $\{Z_j\}_{j=1}^{L}$.
ידוע כי צריך להתקיים הקשר $Z_j=C_jZ+n_j$, כאשר C_j מטריצה ריבועית כלשהי (ידועה) המייצגת את דיכוי התווך, ו- n_j רעש מדידה. רצוננו לשערך את Z מתוך המדידות, וזאת ע"י מציאת מינימום לפונקציה:

$$f(Z)=\sum_{j=1}^{L}\left(Z_j-C_jZ\right)^T\left(Z_j-C_jZ\right)$$

ע"י ריכוז איברים נוכל לכתוב פונקציה זו כפונקציה ריבועית קלאסית:

$$f(Z)=\sum_{j=1}^{L}\left(Z_j-C_jZ\right)^T\left(Z_j-C_jZ\right)=\frac{1}{2}Z^T\mathbf{Q}Z-Z^Tb+c$$

$$\mathbf{Q}=2\sum_{j=1}^{L}C_j^TC_j \quad ; \quad b=2\sum_{j=1}^{L}C_j^TZ_j \quad ; \quad c=\sum_{j=1}^{L}Z_j^TZ_j$$

הבעיה המתוארת לעיל חוזרת שוב ושוב במדעים המדויקים וקרויה שערוך ע"י ריבועים פחותים (Least Squares). אנו רואים בעיית ה- LS הופכת להיות בעיה ריבועית קלאסית.

בקירוב העקומים אשר הוצג בפרק 5 לטובת עניין חיפוש על ישר, קירבנו פונקציה ע"י פרבולה, ונעזרנו בקירוב זה לשם מציאת נקודת המינימום. בדומה לכך, אנו יכולים לבצע קירוב של משטח $f(x)$ לריבועי ע"י קירוב לפונקציה ריבועית ובכך להכליל את שנעשה בפרק 5. נוכל לטעון כי כל פונקציה חלקה מספיק ניתנת לקירוב מקומי כפונקציה ריבועית. דרך הקירוב עושה שימוש בקירוב טיילור באופן הבא:

$$f(x)=f(x_0)+\nabla f(x_0)(x-x_0)+\frac{1}{2}(x-x_0)^T\mathbf{F}(\xi)(x-x_0)$$

$$\text{where:} \quad \xi=\theta x+(1-\theta)x_0 \qquad \theta\in[0,1]$$

כך שהגרדיאנט וההסיין ממלאים את תפקידם של b ו- $\mathbf{Q}$ בהתאמה. אם אנו נמצאים מספיק קרוב לנקודה סביבה נעשה הפיתוח לטור, ברור כי טור זה מתאר קירוב טוב.

לסיכומו של דבר, הפונקציה הריבועית הינה בעלת מינימה גלובלית ולוקלית יחידה, זוהי פונקציה קמורה, והיא מהווה קירוב טיילור מסדר שני לכל פונקציה שהיא אשר הינה חלקה דייה. כמו כן, במקרים רבים בעיית האופטימיזציה הינה בעיה ריבועית ממש - פועל יוצא של שיטת הריבועים הפחותים (Least-Squares) הפופולרית כל-כך (ראה פרק 11). לכן, מרבית ניתוחי האלגוריתמים אשר ייעשו במסגרת קורס זה יעשו שימוש בבעיה הריבועית כבעיית יחוס בעלת עניין רב.

7.3 ההתכנסות של אלגוריתם ה- SD עבור הבעיה הריבועית

מכיוון שגרדיאנט הפונקציה בנקודה x הוא $\nabla f(x)=\mathbf{Q}x-b$, משוואת הרקורסיה של עדכון פתרונו נתונה ע"י הקשר $x_{k+1}=x_k+\alpha\left[b-\mathbf{Q}x_k\right]$.

משפט 7.2: (אלגוריתם ה- SD). עבור הבעיה הריבועית מובטח כי אלגוריתם ה- SD מתכנס עבור כל $\alpha \in \left(0,2/\lambda_{\max}\{\mathbf{Q}\}\right)$, ההתכנסות הינה ליניארית עם יחס התכנסות הנתון ע"י הגודל $\max_k\left\{\left|1-\alpha\lambda_k\{\mathbf{Q}\}\right|\right\}$.

הוכחה: הפעלת המשוואה הריקורסיבית של ה- SD נותנת את הקשר:

$$x_{k+1} = [I-\alpha\mathbf{Q}]x_k + \alpha b = \cdots = [I-\alpha\mathbf{Q}]^{k+1}x_0 + \alpha\cdot\sum_{j=0}^{k}[I-\alpha\mathbf{Q}]^j b$$

שימוש בקשר - $x^* = \mathbf{Q}^{-1}b$ נותן כי:

$$\begin{aligned} x_{k+1} &= [I-\alpha\mathbf{Q}]^{k+1}x_0 + \alpha\cdot\sum_{j=0}^{k}[I-\alpha\mathbf{Q}]^j\mathbf{Q}x^* = \\ &\underset{\substack{\uparrow \\ \sum_{j=1}^{k}A^j=\left[I-A^{k+1}\right][I-A]^{-1}}}{=} [I-\alpha\mathbf{Q}]^{k+1}x_0 + \left[I-[I-\alpha\mathbf{Q}]^{k+1}\right]Q^{-1}Qx^* = \\ &= x^* + [I-\alpha\mathbf{Q}]^{k+1}\left[x_0 - x^*\right] \end{aligned}$$

עשינו שימוש בנוסחאת טור גאומטרי של מטריצות - נוסחה זו נכונה אם $\det[1-A]\neq 0$. אנו רואים שאם המטריצה $[I-\alpha\mathbf{Q}]^{k+1}$ תדעך לאפס נקבל התכנסות לפתרון המדויק. על מנת שמטריצה כלשהי בחזקת k תתכנס כפונקציה של k נדרוש שהערכים העצמיים שלה יהיו בתוך מעגל היחידה. כיוון ש- $\mathbf{Q}$ היא מטריצה חיובית מוגדרת, ניתן לייצגה כ- $\mathbf{Q} = \mathbf{U}^*\Delta\mathbf{U}$ כאשר Δ מטריצה אלכסונית שאיברי אלכסונה הם ערכיה העצמיים של $\mathbf{Q}$, ו- $\mathbf{U}$ מטריצה יוניטרית. לכן:

$$U^*[I-\alpha\Delta]U = [I-\alpha\mathbf{Q}]$$

ופירוש הדבר הוא שהערכים העצמיים של $[I-\alpha\mathbf{Q}]$ הם $1-\alpha\lambda_k\{\mathbf{Q}\} = \lambda_k\{I-\alpha\mathbf{Q}\}$. אנו דורשים:

$$-1 < 1-\alpha\lambda_k\{\mathbf{Q}\} < 1 \quad\Rightarrow\quad 0 < \alpha < \frac{2}{\lambda_{\max}\{\mathbf{Q}\}}$$

כטענת המשפט. אם נניח כי הפרמטר α נבחר בתוך תחום זה נקבל:

$$x_{k+1} = U^*[I-\alpha\Delta]^{k+1}Ux_0 + U^*\left[I-[I-\alpha\Delta]^{k+1}\right]Ux^*$$

אנו רואים כי השפעת האיתחול דועכת באופן אקספוננציאלי לאפס, וכך גם ההתכנסות לפתרון x^*. כזכור, התכנסות אקספוננציאלית הינה בעצם התכנסות ליניארית לפי הגדרתנו. קל לראות כי קיים הקשר:

$$U(x_{k+1}-x^*) = [I-\alpha\Delta]^{k+1}Ux_0 - [I-\alpha\Delta]^{k+1}Ux^*$$
$$U(x_k-x^*) = [I-\alpha\Delta]^{k}Ux_0 - [I-\alpha\Delta]^{k}Ux^*$$

$$\Rightarrow \quad U(x_{k+1}-x^*) = [I-\alpha\Delta]U(x_k-x^*)$$

הפעלת נורמה אויקלידית על שני האגפים נותנת:

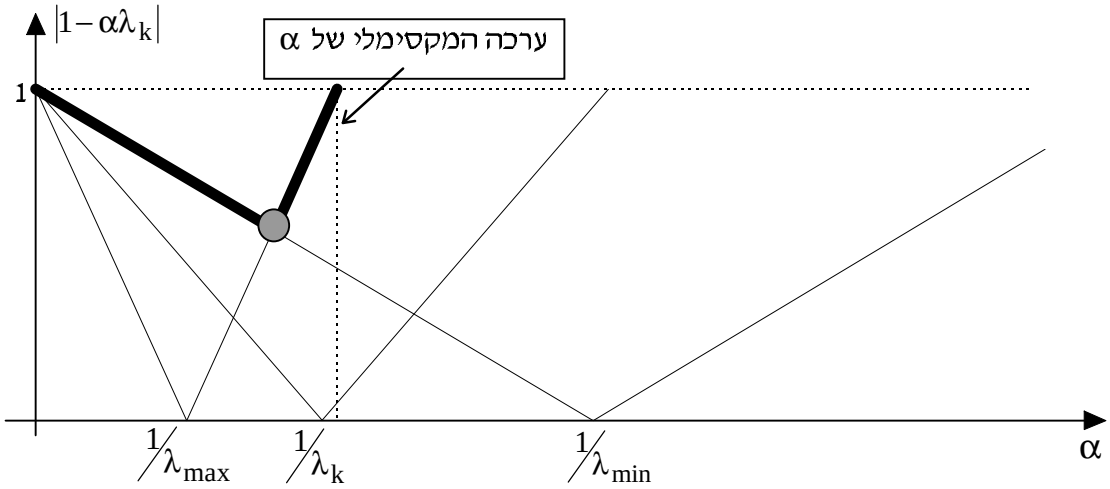
$$\|x_{k+1}-x^*\| \le \|I-\alpha\Delta\|\|x_k-x^*\| \Rightarrow \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} \le \|I-\alpha\Delta\| = \max_k\left\{\left|1-\alpha\lambda_k\{\mathbf{Q}\}\right|\right\}$$

ובכך הוכח המשפט. ■

לאלו שהזדעזעו מהשורה האחרונה, נעיר כי ישנה דרך פשוטה להבינה ללא כל מעורבותן של נורמות על מטריצות. נשתמש בסימון הבא: $e_k = U[x_k - x^*]$. המכפלה $e_{k+1} \le [I-\alpha\Delta]e_k$ הינה מכפלה במטריצה אלכסונית $[I-\alpha\Delta]$ אשר כל איבר אלכסון בה הוא $\left[1-\alpha\lambda_j\{\mathbf{Q}\}\right]$. ברור כי נרצה כי כל קואודינטה של הוקטור e_k תדעך, ולכן נדרוש שכל ערכי האלכסון יהיו בערכם המוחלט קטנים ממש מ- 1. יתרה מזו - האיבר הקרוב ביותר ל- 1 (בערכו המוחלט) יהיה החוליה החלשה בשרשרת, ויכתיב את קצב ההתכנסות - זוהי בדיוק משמעות התוצאה הסופית, בה ביצעהו מקסימיזציה על פני כל $\left[1-\alpha\lambda_j\{\mathbf{Q}\}\right]$, למציאת הכופל הגרוע ביותר.

מהמשפט הקודם ניתן גם לחלץ את הפרמטר α בעל קצב ההתכנסות הגבוה ביותר. קל לראות זאת ע"י תיאור גרפי של הגורמים $\left|1-\alpha\lambda_k\right|$ (ציור 7.1). אנו רוצים בחירה של α אשר תיתן את הערך המכסימלי הנמוך ביותר על כל הפונקציות הנ"ל, כלומר:

$$\alpha_{opt} = \min_\alpha \max_k\left\{\left|1-\alpha\lambda_k\{\mathbf{Q}\}\right|\right\}$$



ציור 7.1 - מציאת α האופטימלי לאלגוריתם ה- SD

הפונקציה max{...} מתוארת בקו עבה בציור לעיל. רואים כי α האופטימלי צריך לקיים:

$$1-\alpha_{opt}\lambda_{min} = \alpha_{opt}\lambda_{max} - 1 \quad \Rightarrow \quad \alpha_{opt} = \frac{2}{\lambda_{max}+\lambda_{min}}$$

ואז יחס ההתכנסות הוא:

$$1-\frac{2\lambda_{min}}{\lambda_{min}+\lambda_{max}} = \frac{\lambda_{max}-\lambda_{min}}{\lambda_{min}+\lambda_{max}} \underset{r=\frac{\lambda_{max}}{\lambda_{min}}}{\overset{}{\underset{\uparrow}{=}}} \frac{r-1}{r+1}$$

ביישום מעשי של אלגוריתם ה- SD יש בעיה בקביעת α. שימוש בשיטת ה- SD מקובל עבור בעיות רבות מימדים (מאות ועד מיליוני נעלמים) בהן קשה לחשב את הערכים העצמיים של **Q**. עם זאת, יש מגוון דרכים לחסום מלמעלה ומלמטה את הערכים העצמיים של מטריצה ללא חישוב ישיר שלהם, ובקלות יחסית של חישובים. לדוגמה, שימוש בנורמת אינסוף הקלה לשימוש:

$$\lambda_{max} \le \|Q\|_\infty \equiv \max_i \sum_j |Q(i,j)|$$

אלגוריתם ה- SD מסוכם בציור 7.2. האלגוריתם המתואר מתיחס למקרה בו הבעייה היא ריבועית.

1. בחר נקודת איתחול שרירותית x_0.
2. קבע על פי חסמים על הערכים העצמיים את ערכו של α.
3. באיטרציה ה- k בצע את העדכון לפי $x_{k+1} = x_k + \alpha[b - \mathbf{Q}x_k]$.
4. עצור את האלגוריתם כאשר $\|b - \mathbf{Q}x_k\| = \|\nabla f(x_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

ציור 7.2 - סיכום אלגוריתם ה- SD לבעיה הריבועית

עבור בעיה שאינה ריבועית יש לחשב בכל נקודה את הגרדיאנט של הפונקציה, והאלגוריתם המוצע מתואר בציור 7.3. תכונות ההתכנסות במקרה הכללי דומות מאוד אם האיתחול קרוב דיו לנקודת המינימום הדרושה. ברור כי אם ההסיין משתנה מהותית תוך כדי האלגוריתם, עלול α לא להתאים, ואז צפויה התבדרות האלגוריתם.

1. בחר נקודת איתחול שרירותית x_0.
2. קבע או שערך את ההסיין בנקודת האיתחול, ועל פי חסמים על הערכים העצמיים של מטריצה זו קבע את ערכו של α.
3. באיטרציה ה- k בצע את העדכון לפי $x_{k+1} = x_k - \alpha\nabla f(x_k)$.
4. עצור את האלגוריתם כאשר $\|\nabla f(x_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

ציור 7.3 - סיכום אלגוריתם ה- SD לבעיה שאינה ריבועית

7.4 ההתכנסות של אלגוריתם ה- NSD עבור הבעיה הריבועית

נזכור כי אלגוריתם ה- NSD צריך לחשב את α_k בכל איטרציה לשם מציאת מינימום על הישר שנבחר. כל חתך של פונקציה ריבועית הינו בצורת פרבולה, ולכן ניתן לקבל ביטוי סגור עבור ערכו של α_k.

משפט 7.3: באלגוריתם ה- NSD ($x_{k+1} = x_k + \alpha_k e_k$), ערכו האופטימלי של α_k נתון ע"י:

$$\alpha_k = \frac{e_k^T e_k}{e_k^T \mathbf{Q} e_k} \quad \text{where} \quad e_k = b - \mathbf{Q} x_k$$

הוכחה: ערכו של α_k צריך להיקבע כך שייתן מינימום לפונקציה

$$f(x_k + \alpha_k e_k) = \frac{1}{2}(x_k + \alpha_k e_k)^T \mathbf{Q}(x_k + \alpha_k e_k) - b^T(x_k + \alpha_k e_k)$$

גזירת ביטוי זה לפי α_k והבאתו לאפס נותנת:

$$\frac{\partial}{\partial \alpha_k} f(x_k + \alpha_k e_k) = e_k{}^T \mathbf{Q}(x_k + \alpha_k e_k) - e_k{}^T b = 0$$

$$\Rightarrow \alpha_k = \frac{e_k{}^T e_k}{e_k{}^T \mathbf{Q} e_k}$$

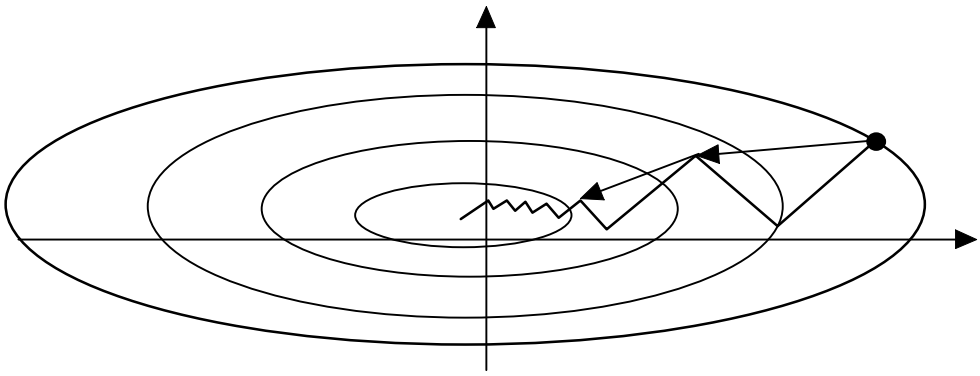
וטענת המשפט הוכחה. ■

למעשה, יכולנו לקבל את אותה התוצאה מכיוון אחר. עפ"י משפט הניצבות מפרק קודם אנו מקבלים כי צריך להתקיים - $e_k{}^T e_{k+1} = 0$. משפט הניצבות תקף כאן כי אנו מבצעים חיפוש מדויק על ישר. לכן צריך להתקיים:

$$e_k{}^T e_{k+1} = e_k{}^T[b - \mathbf{Q} x_{k+1}] = e_k{}^T[b - \mathbf{Q}(x_k + \alpha_k e_k)] = e_k{}^T[e_k - \alpha_k \mathbf{Q} e_k] = 0$$

וגם מכאן ניכר כי מגיעים לאותה תוצאה. יש לשים לב לעובדה הבאה - אנו פוסעים באלגוריתם ה- NSD באופן אופטימלי בכל איטרציה, ובשיעור אופטימלי, אך פירוש הדבר תופעת זיג-זוג שהרי הגרדיאנטים העוקבים (שאינם אלא הכיוונים בזמנים עוקבים) אנכיים זה לזה. תיאור גרפי ממחיש את הבזבוז שבדבר.

תופעת הזיג-זג מראה כי "אנרגיית חיפוש" מבוזבזת בהליכה לא ממוקדת לקראת נקודת הפתרון. התיאור הגרפי מעלה גם את הרעיון הבא - צמדים של איטרציות יכולים עם זאת להיות מאוחדים לכלל כיוון אחד, ועליו נבצע חיפוש על ישר. אנו נפגוש רעיון זה מכיוון אחר בפרק הבא.



ציור 7.4 - תיאור ה- NSD ותופעת הזיג-זג

משפט 7.4: עבור הבעיה הריבועית, אלגוריתם ה- NSD מתכנס לנקודת המינימום באופן ליניארי (אקספוננציאלי) עם יחס התכנסות **<u>החסום</u>** ע"י $\frac{r-1}{r+1}$, כאשר r הוא ה- # Condition של המטריצה **Q**.

הוכחה: נשתמש בפונקציה $E(x)=\frac{1}{2}(x-x^*)^T\mathbf{Q}(x-x^*)$, אשר כפי שראינו קודם היא שקולה לפונקציה הריבועית עד כדי קבוע. אזי מתקיים:

$$\frac{E(x_k)-E(x_{k+1})}{E(x_k)}=\frac{(x_k-x^*)^T\mathbf{Q}(x_k-x^*)-(x_k+\alpha_k e_k-x^*)^T\mathbf{Q}(x_k+\alpha_k e_k-x^*)}{(x_k-x^*)^T\mathbf{Q}(x_k-x^*)}=$$

$$=-\frac{(\alpha_k e_k)^T\mathbf{Q}(\alpha_k e_k)+2(\alpha_k e_k)^T\mathbf{Q}(x_k-x^*)}{(x_k-x^*)^T\mathbf{Q}(x_k-x^*)}$$

שימוש בעובדה ש- $e_k=b-\mathbf{Q}x_k=\mathbf{Q}[x^*-x_k]$ והצבת הנוסחה עבור α_k נותנים:

$$\frac{E(x_k)-E(x_{k+1})}{E(x_k)}=-\frac{\frac{(e_k{}^Te_k)^2}{e_k{}^T\mathbf{Q}e_k}-2\frac{(e_k{}^Te_k)^2}{e_k{}^T\mathbf{Q}e_k}}{e_k{}^T\mathbf{Q}^{-1}e_k}=\frac{[e_k{}^Te_k]^2}{[e_k{}^T\mathbf{Q}^{-1}e_k][e_k{}^T\mathbf{Q}e_k]}$$

עבור הביטוי אשר התקבל קיים חסם הקרוי חסם Kantrovitch לפיו:

$$\frac{E(x_k)-E(x_{k+1})}{E(x_k)}=\frac{\left[e_k{}^T e_k\right]^2}{\left[e_k{}^T \mathbf{Q}^{-1} e_k\right]\left[e_k{}^T \mathbf{Q} e_k\right]}\geq\frac{4aA}{(a+A)^2}$$

כאשר A ו- a הם הערכים העצמיים הגדול ביותר והקטן ביותר של **Q**. לכן קיבלנו בפועל כי:

$$E(x_{k+1})\leq\left[1-\frac{4aA}{(A+a)^2}\right]E(x_k)=\left[\frac{A-a}{A+a}\right]^2 E(x_k)=\left[\frac{r-1}{r+1}\right]^2 E(x_k)$$

כאשר r הוא יחס הערכים העצמיים A ו- a. לפי למה 4.1, נוכל לקבוע כי:

$$\frac{E(x_{k+1})}{E(x_k)}\leq\left[\frac{r-1}{r+1}\right]^2 \quad\Rightarrow\quad \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|}\leq\left[\frac{r-1}{r+1}\right]$$

ומכאן שההתכנסות הינה ליניארית, ועל-פי יחס התכנסות כפי שנאמר. ■

משמעות תוצאה זו היא שעבור מטריצות **Q** בעלות r גדול מאוד תחול התכנסות איטית ביותר. יש לשים לב לעובדה שיחס ההתכנסות עבור ה- SD עם α אופטימלי זהה לזה שהתקבל כאן. אך יש לזכור ששם קיבלנו כי זהו בדיוק יחס ההתכנסות בעוד שכאן זהו חסם עליון עליו. לכן, ברור כי ה- NSD יותר טוב, או לפחות מהיר כמו ה- SD.

דוגמה 7.2: נניח כי בידנו בעיית אופטימיזציה ריבועית קלאסית. דרך אחת לפותרה היא ע"י ה- SD (או ה- NSD) באופן ישיר. דרך אחרת היא לפתור את מערכת המשוואות $\mathbf{Q}x^*=b$. אך כיוון שפתרון מערכת משוואות כזו עשוי להיות קשה (נניח כי מספר המשוואות עצום) נביא למינימום את הפונקציה - $g(x)=\|\mathbf{Q}x-b\|^2$. השאלה אותה אנו שואלים היא - מה עדיף מבחינת קצב התכנסות?

נשים לב כי גם הפונקציה g היא בעלת מבנה ריבועי, אך במקום המטריצה **Q** במבנה הרגיל יופיע $\mathbf{Q}^2$. פירוש הדבר הוא שיחס ההתכנסות יהיה:

$$f(x) = 0.5(x - x^*)^T \mathbf{Q}(x - x^*) \Rightarrow \quad \left[\frac{r-1}{r+1}\right]^2 \approx \left(1 - r^{-1}\right)^4$$

$$g(x) = (\mathbf{Q}x - b)^T (\mathbf{Q}x - b) \Rightarrow \quad \left[\frac{r^2-1}{r^2+1}\right]^2 \approx \left(1 - r^{-2}\right)^4$$

כך שהתקבל כי הגישה השניה איטית הרבה יותר. לדוגמה עבור r=100 יתקבל כי היחס במקורי הוא 0.99 בעוד שהיחס החדש הוא 0.9999. זה גם אומר שבקירוב נידרש ל- r איטרציות של הגישה השניה לקבלת תוצאת איטרציה אחת של הגישה הראשונה.

לסיכום הדיון באלגוריתם ה- NSD, בציור 7.5 מובא ריכוז של אלגוריתם לבעיה הריבועית.

1. בחר נקודת איתחול שרירותית x_0.
2. באיטרציה ה- k קבע את גודל הצעד לפי -

$$\alpha_k = \frac{e_k^T e_k}{e_k^T \mathbf{Q} e_k} \quad \text{where} \quad e_k = b - \mathbf{Q}x_k$$

3. בצע את העדכון לפי $x_{k+1} = x_k + \alpha_k e_k$.
4. עצור את האלגוריתם כאשר $\|e_k\| = \|\nabla f(x_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

ציור 7.5 - סיכום אלגוריתם ה- NSD לבעיה הריבועית

עבור בעיה שאינה ריבועית יש לחשב בכל נקודה את הגרדיאנט של הפונקציה, והאלגוריתם המוצע מתואר בציור 7.6. תכונות ההתכנסות במקרה הכללי דומות מאוד אם האיתחול קרוב דיו לנקודת המינימום הדרושה.

7.5 ההתכנסות של אלגוריתם ניוטון

עד כה ראינו כי בחירה איטואיטיבית הובילה להרכבת אלגוריתם ה- SD או ה- NSD, ולא מצאנו מוטיבציה לגישת ניוטון הכופלת את הכיוון המוצלח ביותר במטריצה חיובית מוגדרת - בהיפוכו של ההסיין. למעשה, עבור הבעיה הריבועית קל לראות כי אלגוריתם ניוטון פותר את הבעיה באיטרציה אחת לכל איתחול שהוא:

1. בחר נקודת איתחול שרירותית x_0.
2. באיטרציה ה- k קבע את הגרדיאנט המקומי $\nabla f(x_k)$.
3. קבע את גודל הצעד לעדכון α_k לפי חיפוש על הישר - $f(x_k - \alpha \cdot \nabla f(x_k))$ באחת השיטות שהוצגו.
3. בצע את העדכון לפי $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.
4. עצור את האלגוריתם כאשר $\|e_k\| = \|\nabla f(x_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 7.6 - סיכום אלגוריתם ה- NSD לבעיה שאינה ריבועית</u>

$$f(x) = \frac{1}{2}(x - x^*)^T \mathbf{Q}(x - x^*) \quad \Rightarrow \quad \begin{array}{c} \nabla f(x) = \mathbf{Q}(x - x^*) \\ \mathbf{F}(x) = \mathbf{Q} \end{array}$$

$$x_1 = x_0 - \mathbf{F}^{-1}(x)\nabla f(x) = x_0 - \mathbf{Q}^{-1}\left[\mathbf{Q}(x_0 - x^*)\right] = x^*$$

כך שתוצאה זו בהחלט מלמדת על כוחו של אלגוריתם ניוטון ועדיפותו על-פני ה- SD.

משפט 7.5: תהי $f(x) \in C^3$, וניח כי בנקודת המינימום של f אשר מסומנת x^* ההסיין חיובי מוגדר. אזי, עבור איתחול קרוב דיו לנקודת הפתרון, סידרת הנקודות המיוצרת ע"י אלגוריתם ניוטון מתכנסת לפתרון בסדר 2 לפחות.

הוכחה: ההוכחה דומה מאוד להוכחה שניתנה לאלגוריתם ניוטון החד-מימדי (משפט 5.4). נוכל לכתוב כי -

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - \mathbf{F}^{-1}(x_k)\nabla f(x_k) = \\ &= -\mathbf{F}^{-1}(x_k)\left[\nabla f(x_k) - \nabla f(x^*) + \mathbf{F}(x_k)(x^* - x_k)\right] \end{aligned}$$

ומכיוון שזהו פיתוח לטור טיילור עד סדר ראשון של נגזרת הפונקציה נוכל לקבוע כי:

$$\nabla f(x_k)-\nabla f(x^*)+\mathbf{F}(x_k)(x^*-x_k)=O\left\{\|x^*-x_k\|^2\right\}$$

לכן מתקבל ע"י הצבת הקשר האחרון בקודמו כי:

$$\|x_{k+1}-x^*\|\le\left\|\mathbf{F}^{-1}(x_k)\right\|\cdot k_1\cdot\|x_k-x^*\|^2$$
$$\Rightarrow\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|^2}\le k_1k_2$$

כאשר k_2 הוא חסם עליון על נורמת היפוך ההסיין. תוצאה זו שקולה לאמירה - קצב התכנסות מסדר שני. ■

לריכוז הדיון בנושא אלגוריתם ניוטון, בציור 7.7 תיאורו עבור בעיה שאינה ריבועית (שהרי עבור בעיה ריבועית הוא מתכנס בצעד אחד, ויישומו שקול לפתרון מערכת משוואות ליניארית).

1. בחר נקודת איתחול שרירותית x_0.
2. באיטרציה ה- k קבע את הגרדיאנט וההסיין המקומיים $\nabla f(x_k),\mathbf{F}(x_k)$.
3. בצע את העדכון לפי $x_{k+1}=x_k-\mathbf{F}^{-1}(x_k)\nabla f(x_k)$.
4. עצור את האלגוריתם כאשר $\|\nabla f(x_k)\|\le\varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 7.7 - סיכום אלגוריתם ניוטון לבעיה שאינה ריבועית</u>

7.6 וריאציות של אלגוריתם ניוטון

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:

1. מטריצת ההסיין לבעיה הכללית עלולה לא להיות חיובית מוגדרת. במקרה כזה כבר ראינו בתחילת פרק זה כי הכיוון המבוצע אינו מחייב כלל ירידה.

2. מטריצת ההסיין עלולה להיות סינגולרית -עובדה שפירושה שלא קיימת מטריצה הפוכה לה.
3. קצב התכנסות 2 הושג 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 שינוי מטריצת ההסיין ע"י תוספת של מטריצת יחידה מוכפלת בקבוע. שינוי זה נועד להבטיח שתתקבל מטריצה חיובית מוגדרת (ולכן לא סינגורלית) M. מעבר לכך מבוצע כאן חיפוש על ישר ולא נקבע כבאלגוריתם ניוטון כי $\alpha_k = 1$.

1. בחר נקודת איתחול שרירותית x_0, ופרמטר δ חיובי.
2. באיטרציה ה- k קבע את הגרדיאנט וההסיין המקומיים $\nabla f(x_k), \mathbf{F}(x_k)$.
3. מצא $\varepsilon_k > 0$ כך שיתקבל כי $\varepsilon_k I + \mathbf{F}(x_k) > \delta I$.
4. קבע את כיוון החיפוש ל- $d_k = \left[\varepsilon_k I + \mathbf{F}(x_k)\right]^{-1} \nabla f(x_k)$.
5. חשב את α_k ע"י חיפוש על הישר
$$g(\alpha) = f(x_k - \alpha_k d_k)$$
5. בצע את העדכון לפתרון לפי הביטוי
$$x_{k+1} = x_k - \alpha_k \left[\varepsilon_k I + \mathbf{F}(x_k)\right]^{-1} \nabla f(x_k).$$
6. עצור את האלגוריתם כאשר $\|\nabla f(x_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 7.8 - אלגוריתם ניוטון ההיברידי לבעיה שאינה ריבועית</u>

האלגוריתם השני קרוי אלגוריתם Levenberg-Marquardt. ההבדל המהותי ביחס לאלגוריתם הקודם הוא בשימוש בפירוק Cholesky. פירוק זה הינו מעין הוצאת שורש למטריצה בצורת מטריצה משולשת תחתונה. כפועל יוצא קובע פירוק זה אם המטריצה חיובית מוגדרת.

אלגוריתם ניוטון המקורי וגרסאותיו המשופרות הינם אלגוריתמים מורכבים אשר ביצועיהם התיאורתיים לא עולים תמיד בקנה אחד עם ביצועיהם המעשיים. מעבר לכך, יש לשים לב למורכבותם החישובית הגבוהה - רק עבור בעיות עם מספר נעלמים מועט (עד כ- 10) ועם פונקציית מינימיזציה אשר נוחה לגזירה קביל להפעיל אלגוריתמים אלו.

1. בחר נקודת איתחול שרירותית x_0, ופרמטר δ חיובי.
2. באיטרציה ה- k קבע את הגרדיאנט וההסיין המקומיים $\nabla f(x_k), \mathbf{F}(x_k)$.
3. בצע פירוק Cholesky למטריצה - $\varepsilon_k I + \mathbf{F}(x_k) = GG^T$ כאשר G מטריצה משולשת תחתונה (תהליך מורכב חישובית אך יציב). עפ"י הפירוק מצא $\varepsilon_k > 0$ כך שיתקבל כי $\varepsilon_k I + \mathbf{F}(x_k) > \delta I$.
4. קבע את כיוון החיפוש ל-פי פתרון מערכת המשוואות $GG^T d_k = \nabla f(x_k)$ (יחסית קל לביצוע כי G משולשת תחתונה).
5. חשב את α_k ע"י חיפוש על הישר
$$g(\alpha) = f(x_k - \alpha_k d_k)$$
6. בצע את העדכון לפתרון לפי הביטוי
$$x_{k+1} = x_k - \alpha_k \left[\varepsilon_k I + \mathbf{F}(x_k)\right]^{-1} \nabla f(x_k).$$
7. עצור את האלגוריתם כאשר $\|\nabla f(x_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 7.9 - אלגוריתם לוונברג-מרקרט לבעיה שאינה ריבועית</u>

בסעיף 3 של האלגוריתם הנ"ל, מבוצע פירוק Choleski. אם הוא אינו מצליח, מגדילים את ε_k וחוזרים על נסיון הפירוק.

7.7 ההתכנסות של אלגוריתם ה- DNSD עבור הבעיה הריבועית

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 שוים, אנו מקבלים כמקרה פרטי את ה- NSD. נכתוב את משוואת האלגוריתם הכללית עבור הבעיה הריבועית:

$$x_{k+1} = x_k - \mathbf{D}_k \nabla f(x_k) = x_k + \mathbf{D}_k e_k = x_k + \begin{bmatrix} \alpha_k^1 & & 0 \\ & \ddots & \\ 0 & & \alpha_k^n \end{bmatrix} e_k$$

אנו לא ננתח את קצב ההתכנסות של אלגוריתם זה - אם כי ברור כי ביצועיו טובים מאלו של ה- NSD, אך ירודים מאלו של שיטת ניוטון.

משפט 7.6: עבור אלגוריתם ה- DNSD, ערכי האלכסון האופטימליים למטריצה $\mathbf{D}_k$ נתונים ע"י הפתרון למשוואה - $[I - \mathbf{QD}_k]e_k = 0$ אם נניח כי כל איברי הוקטור e_k שונים מאפס. עבור מקרה זה - האלגוריתם אינו אלא אלגוריתם ניוטון, ולפיכך מתכנס בצעד אחד בבעיה הריבועית.

הוכחה: הפרמטרים $\left\{\alpha_k^j\right\}_{j=1}^n$ נקבעים לקבלת מינימום לערך הפונקציה הבאה:

$$f(x_{k+1}) = [x_{k+1} - x^*]^T \mathbf{Q} [x_{k+1} - x^*]$$

לכן, גזירה והשוואה לאפס לפי וקטור פרמטרים אלו - $\alpha_k^* = \left[\alpha_k^1 \quad \cdots \quad \alpha_k^n\right]$ - תיתן:

$$\frac{\partial f(x)}{\partial \alpha_k^*} = \left[\mathrm{diag}\{e_k\}\right]^T \mathbf{Q}\left[x_k - x^* + \mathbf{D}_k e_k\right] = 0$$

עשינו כאן שימוש בעובדה שניתן לגזור לפי כל אחד מהפרמטרים בנפרד ולאחד את המשוואות לאחת כללית, שנראית כפי שנכתב לעיל. גזירה של f לפי פרמטר יחיד מניבה ערך סקלרי - גזירה של f לפי וקטור של נעלמים לפיכך תיתן וקטור של ערכים אותם עלינו להשוות לאפס.

שימוש בכך שמטריצה אלכסונית סימטרית לעצמה מאפשרת השמטת פעולת השחלוף, והעובדה ש - $\mathbf{Q}[x_k - x^*] = \mathbf{Q}x_k - b = -e_k$ נותנים:

$$\mathrm{diag}\{e_k\}[I-\mathbf{QD}_k]e_k=0$$

ותחת ההנחה שכל איברי הוקטור e_k שונים מאפס נוכל להשמיט את המכפלה ב- $\mathrm{diag}\{e_k\}$ ולקבל כי נדרש $[I-\mathbf{QD}_k]e_k=0$, כטענת המשפט.

לא לחינם ניתנה הטענה של המשפט לעיל באופן סתום. פתרון של מערכת משוואות זו תהיה:

$$\left[\begin{bmatrix}1 & & & 0\\ & 1 & & \\ & & \ddots & \\ 0 & & & 1\end{bmatrix}-\begin{bmatrix}q_{11} & q_{12} & \cdots & q_{1n}\\ q_{21} & q_{22} & \cdots & q_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ q_{n1} & q_{n2} & \cdots & q_{nn}\end{bmatrix}\begin{bmatrix}\alpha_k^1 & & & 0\\ & \alpha_k^2 & & \\ & & \ddots & \\ 0 & & & \alpha_k^1\end{bmatrix}\right]\begin{bmatrix}e_k^1\\ e_k^2\\ \vdots\\ e_k^n\end{bmatrix}=$$

$$=\begin{bmatrix}1-\alpha_k^1q_{11} & -\alpha_k^1q_{12} & \cdots & -\alpha_k^nq_{1n}\\ -\alpha_k^1q_{21} & 1-\alpha_k^1q_{22} & \cdots & -\alpha_k^nq_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ -\alpha_k^1q_{n1} & -\alpha_k^1q_{n2} & \cdots & -\alpha_k^nq_{nn}\end{bmatrix}\begin{bmatrix}e_k^1\\ e_k^2\\ \vdots\\ e_k^n\end{bmatrix}=0$$

ככלל, פתרונה של מערכת משוואות ליניארית זו שקולה לצורך בהיפוכה של $\mathbf{Q}$ - ניתן לראות זאת ע"י הצגת הפתרון למשוואה זו בפשטות ע"י המשוואה:

$$[I-\mathbf{QD}_k]e_k=0 \;\Rightarrow\; \mathbf{Q}^{-1}e_k=\mathbf{D}_ke_k \;\Rightarrow\; \alpha_k^*=\mathbf{Q}^{-1}e_k./e_k$$

כאשר הסימן $./$ פירוש חלוקת וקטורים זה בזה איבר איבר (כמו ב- MATLAB). פירוש הדבר שחישוב הערכים האופטימליים אינו מעשי (מכיוון שאם כבר אנו מוכנים להפוך את $\mathbf{Q}$, נהפוך אותה בעת חישוב הפתרון למערכת המשוואות $\mathbf{Q}x^*=b$ ונסיים את מלאכת האופטימיזציה). תוצאה זו פירושה שבחירת הקבועים α_k^* תיתן שוקטור העדכון לפתרון יהיה (ראה המשוואה הנ"ל) $\mathbf{D}_ke_k=\mathbf{Q}^{-1}e_k$ וזהו אלגוריתם ניוטון במדויק. ■

המשפט הנ"ל אינו מניב תוצאה מעשית - לצוך יישום החישוב של המקדמים האופטימליים עלינו להפוך את ההסיין., ואם כבר עשינו זאת, מוטב כבר להשתמש באלגוריתם ניוטון.

גישה מקובלת לשימוש אשר מקרבת את ה- DNSD היא גישה בה עושים שימוש באלכסון הראשי בלבד של המטריצה **Q**. כלומר שימוש במישקול האלכסוני הבא:

$$\mathbf{D}_k = \begin{bmatrix} \left[\dfrac{\partial^2 f(x_k)}{\left(\partial x_k^1\right)^2}\right]^{-1} & & 0 \\ & \ddots & \\ 0 & & \left[\dfrac{\partial^2 f(x_k)}{\left(\partial x_k^1\right)^2}\right]^{-1} \end{bmatrix} = \begin{bmatrix} q_{11}^{-1} & & 0 \\ & \ddots & \\ 0 & & q_{nn}^{-1} \end{bmatrix}$$

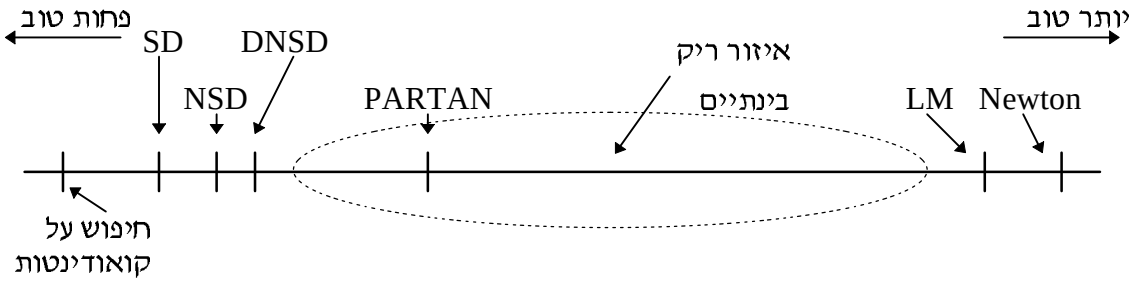
ברור שעבור מטריצה Q קרובה לאלכסונית אנו מקבלים כי הביצועים מתקרבים לביצועי ה- DNSD האופטימלי (ובעצם - לביצועי ניוטון). בכל מקרה, רצוי להשתמש בחיפוש על ישר, כיוון שאם Q אינה קרובה לאלכסונית, וקטור העדכון המתקבל יכול להביא להתבדרות בשל גודלו.

7.8 סיכום

בפרק זה הכרנו לראשונה אלגוריתמים מעשיים למינימיזציה כללית ללא אילוצים של פונקציה. הגישה שננקטה היא תהליך איטרטיבי הבוחר כיוון לחיפוש נקודת הפתרון הבאה, ומחפש נקודה נמוכה יותר בפונקציה. בפרק זה הצגנו גישות מבוססות גרדיאנט - דהיינו - גרדיאנט קובע את הכיוון הכללי בו אנו מתקדמים.

פרק 8 - שיטות האצה של אלגוריתם ה- SD

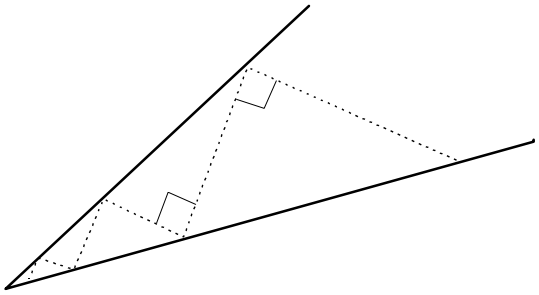
בפרק הקודם ראינו כי עבור בעיות ריבועיות (או בעיות הניתנות לקירוב מקומי ככאלה), בהן המטריצה **Q** היא בעלת Condition # גבוה, יעילותם של אלגוריתמי ה- SD וה- NSD יורדת פלאים. מאידך, אלגוריתם ניוטון אשר תכונות ההתכנסות שלו מעולות, אינו מעשי במימדים בינוניים ומעלה (מעל 10 נעלמים) בשל מורכבותו החישובית. בדרך זו הכרנו את שתי הקצוות האפשריים של האלגוריתמים האיטרטיביים. בפרק זה נכיר מגוון דרכים אשר נועדו לזרז קצבי התכנסות של אלגוריתמים איטרטיביים. אנו נפעיל גישות אלו על ה- SD ונראה כי כל דרכים אלו מניבות ממקורות שונים את האלגוריתם הקרוי שיטת הגרדיאנטים הצמודים - The Conjugate Gradients Method. אלגוריתם זה מהווה גישת ביניים מכל הבחינות בין שתי הקצוות שתוארו. בפרק הבא נייחד דיון אלטרנטיבי אשר מציג את אלגוריתם ה- CG מכיוון אחר.



ציור 8.1 - ציר איכות לאלגוריטמי אופטימיזציה - קצב התכנסות

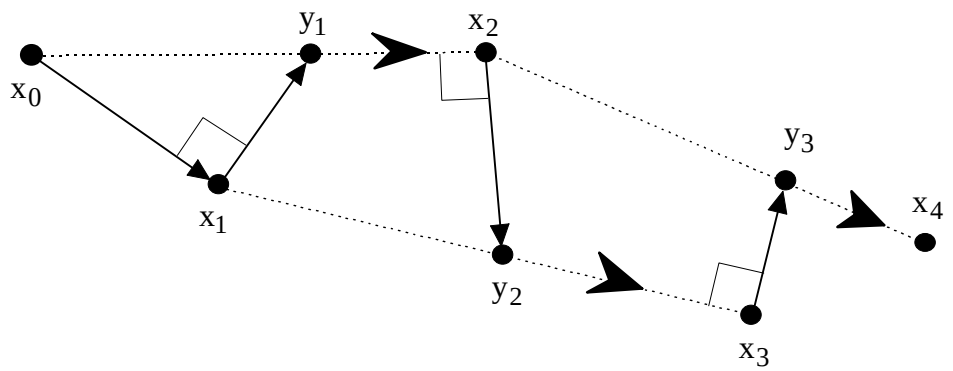
8.1 גישת PARTAN לזירוז אלגוריתם ה- NSD

ראינו בפרק 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. ראה ציור 8.2 להבנת התנהגות האלגוריתם.



ציור 8.2 - תופעת הזיג-זג ב- NSD

הצעתנו לשיפור אשר נקראית שיטת ה- PARTAN (Parallel Tangent) (הוצע לראשונה בשנת 1951 אך נותח ונוסח במלואו בשנת 1964) מציעה לקחת תוצאות של איטרציות עוקבות של ה- NSD ולבצע חיפוש על ישר אשר מאחד אותן. האלגוריתם מתחיל בנקודת איתחול שרירותית x_0. ממנה באיטרציית NSD מקבלים את x_1. באופן דומה נבצע צעד NSD נוסף ונקבל את y_1. כעת בא ההבדל - במקום לקבוע $x_2 = y_1$ כפי שעושה זאת ה- NSD, נקבע את x_2 כנקודה המשיגה את המינימום של הפונקציה הריבועית על הישר אשר נמתח בין x_0 ל- y_1. באופן דומה, אם נניח כי אנו באיטרציה ה- k - ובידנו הנקודה x_k, צעד NSD יוצר את הנקודה y_k, והנקודה x_{k+1} תתקבל על הישר הקושר את x_{k-1} ל-y_k. תיאור גרפי של הרעיון מתואר בציור 8.3.



ציור 8.3 - שיטת ה- PARTAN - תיאור עקרוני - קוים מלאים - צעדי NSD

נכתוב את משוואות האלגוריתם אשר תואר קודם מילולית:

1. $e_k = b - \mathbf{Q}x_k$

2. $y_k = x_k + \dfrac{e_k{}^T e_k}{e_k{}^T \mathbf{Q} e_k} e_k$

3. $\theta_k = \arg\min_{\theta} f\left(\theta\left[y_k - x_{k-1}\right] + x_{k-1}\right)$

4. $x_{k+1} = \theta_k\left[y_k - x_{k-1}\right] + x_{k-1}$

הפרמטר θ_k ניתן לקביעה לפי:

$$\theta_k = \arg\min_{\theta}\left[\theta\left[y_k - x_{k-1}\right] + x_{k-1} - x^*\right]^T \mathbf{Q}\left[\theta\left[y_k - x_{k-1}\right] + x_{k-1} - x^*\right]$$

$$\Rightarrow \left[y_k - x_{k-1}\right]^T \mathbf{Q}\left[\theta\left[y_k - x_{k-1}\right] + x_{k-1} - x^*\right] = 0$$

$$\Rightarrow \theta_k = \frac{\left[y_k - x_{k-1}\right]^T \mathbf{Q}\left[x^* - x_{k-1}\right]}{\left[y_k - x_{k-1}\right]^T \mathbf{Q}\left[y_k - x_{k-1}\right]} = \frac{\left[y_k - x_{k-1}\right]^T e_{k-1}}{\left[y_k - x_{k-1}\right]^T \mathbf{Q}\left[y_k - x_{k-1}\right]}$$

לדוגמה, עבור מצב בו $x^*-x_{k-1}=y_k-x_{k-1}\Rightarrow y_k=x^*$ יתקבל כי $\theta_k=1$ - תוצאה רצויה כיוון שפירושה ש- $x_{k+1}=y_k=x^*$. האלגוריתם אשר נגזר מתוצאה זו מתואר בציור 8.4. אחד מיתרונותיו הבולטים של אלגוריתם זה הוא שהוא אינו יכול להיות גרוע יותר מה- NSD - כיוון שמבוצע חיפוש על ישר הכולל את נקודת ה- NSD, ואם היא הכי נמוכה, נקבל אותה ונתלכד עם ה- NSD. פירוש הדבר הוא שאלגוריתם זה הוא בעל תכונת התכנסות גלובלית. בנוסף, אלגוריתם זה פשוט ליישום, ומורכבותו החישובית גבוהה רק במקצת בהשוואה ל- NSD. בהמשך נכיר את אלגוריתם הגרדיאנטים הצמודים , שאיננו אלא אלגוריתם זה המפותח מכיוון שונה.

1. בחר נקודת איתחול שרירותית x_0.
2. חשב את הנקודה x_1 ע"י איטרציית NSD.
3. באיטרציה ה- k אנו בנקודה x_k. חשב צעד NSD אחד וקבל את y_k.
3. על-פי - y_k, x_k, x_{k-1} חשב את ערכו של θ_k. על-פיו חשב את $x_{k+1}=\theta_k[y_k-x_{k-1}]+x_{k-1}$.
4. עצור את האלגוריתם כאשר $\|e_k\|=\|\nabla f(x_k)\|\le\varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 8.4 - סיכום אלגוריתם ה- PARTAN לבעיה הריבועית</u>

נשים לב כי האלגוריתם שהוצע הינו אלגוריתם בעל משוואת עדכון מסדר שני, כיוון שבעיקרה היא מציעה את דרך העדכון הבאה:

$$\begin{aligned}x_{k+1}&=\theta_k\left[x_k+\alpha_k\left(b-\mathbf{Q}x_k\right)\right]+\left(1-\theta_k\right)x_{k-1}=\\&=\theta_k\left[I-\alpha_k\mathbf{Q}\right]x_k+\left(1-\theta_k\right)x_{k-1}+\alpha_k\theta_k b\end{aligned}$$

השימוש בשתי נקודות קודמות בתהליך לקביעת הנקודה החדשה הוא רעיון המכליל את תפיסת האלגוריתמים מפרק קודם, ויכול לפיכך רק להועיל, אם מיושם נכון.

8.2 האצה פולינומיאלית של אלגוריתמים

גישה אחרת להאצת ה- SD היא גישה הקרויה האצה פולינומיאלית (גישה זו מאפשרת האצה של מגוון אלגוריתמים, ולא מוגבלת לטיפול ב- SD בלבד). אנו נמחיש את עניין ההאצה על אלגוריתם ה- SD - עובדה שתקל מעט בביטויים אותם יש לכתוב. הרעיון הבסיסי הוא הבא - נזכור כי עבור אלגוריתם ה- SD מתקיימים הקשרים הבאים (ראה פרק קודם):

$$x_{k+1} = x_k + \alpha e_k \quad \text{where} \quad e_k = b - \mathbf{Q}x_k$$
$$\Rightarrow x_{k+1} - x^* = x_k - x^* + \alpha\left[\mathbf{Q}x^* - \mathbf{Q}x_k\right] = \left[I - \alpha\mathbf{Q}\right]\left[x_k - x^*\right]$$

אם נסמן $\tilde{\varepsilon}_k = x_k - x^*$, קיבלנו כי מתקיים - $\tilde{\varepsilon}_{k+1} = \left[I - \alpha\mathbf{Q}\right]\tilde{\varepsilon}_k = \left[I - \alpha\mathbf{Q}\right]^{k+1}\tilde{\varepsilon}_0$.

נגדיר את הפולינום הבא מסדר k:

$$P_k(z) = \sum_{j=0}^{k}\alpha_{k,j}\cdot z^j \;\; ; \;\; P_k(1) = \sum_{j=0}^{k}\alpha_{k,j} = 1$$

נציע כעת תהליך אשר יפעל על תוצאות הנקודות המתקבלות מה- NSD ואשר יבצע את הפעולה הבאה תוך שימוש בפולנום שהוגדר:

$$y_k = \sum_{j=0}^{k}\alpha_{k,j}x_j$$

(מעריך בפולינום מומר להיות האינדקס של הוקטור). משמעות פעולה זו - חישוב ממוצע משוקלל של כל הנקודות אשר התקבלו עד כה. אנו נקרא לסידרת המוצא y_k - התהליך המואץ. כבר ראינו גישה אשר הציעה שיקלול של שתי הנקודות האחרונות - גישת ה- PARTAN. לפיכך גישה זו כללית יותר. שימוש בביטוי האחרון נותן:

$$y_k - x^* \overset{\Delta}{=} \varepsilon_k = \sum_{j=0}^{k}\alpha_{k,j}x_j - x^* = \sum_{j=0}^{k}\alpha_{k,j}\left(x_j - x^*\right) = \sum_{j=0}^{k}\alpha_{k,j}\tilde{\varepsilon}_j$$
$$\Rightarrow \varepsilon_k = \sum_{j=0}^{k}\alpha_{k,j}\tilde{\varepsilon}_j = \sum_{j=0}^{k}\alpha_{k,j}\left[I - \alpha\mathbf{Q}\right]^j\varepsilon_0 = \left\{P_k\left(\left[I - \alpha\mathbf{Q}\right]\right)\right\}\varepsilon_0$$

כאשר עשינו שימוש בעובדה שסכום מקדמי הפולינום הוא 1. מכאן גם ברור הקשר לפולינומים - התקבל כי השגיאה של התהליך המואץ היא צירוף ליניארי של שגיאות ה- SD עם משקלות שהינן

מטריצות בחזקות עולות, ובעצם - פולינום של המטריצה האחראית על ההתכנסות ב- SD - $I-\alpha\mathbf{Q}$. לדוגמה, אם אנו בסדר פולינום גבוה דיו, ונמצא את מקדמי הפולינום האופייני של $[I-\alpha\mathbf{Q}]$, נקבל כי השגיאה מאופסת זהותית. כלומר - מצוי בידנו כלי בעל פוטנציאל להקטנת השגיאה של התהליך המואץ. המחיר, עם זאת, אינו קל - יש לזכור את כל הפתרונות במצטבר עד כה במהלך שיטוטו של ה- SD, ויש כמות הולכת וגדלה של פרמטרים לביצוע הממוצע המשוקלל. לכן נציע כאלטרנטיבה את הפולינום הרקורסיבי הכללי הבא:

$$
\begin{aligned}
&P_0(z)=1\\
&P_1(z)=\gamma_1 z-\gamma_1+1\\
&P_k(z)=\rho_k(\gamma_k z-\gamma_k+1)P_{k-1}(z)+(1-\rho_k)P_{k-2}(z) \quad k\ge 2
\end{aligned}
$$

קל לראות כי פולינומים אלו מקיימים את התכונה שסכום מקדמיהם שווה 1 לכל סדר (ע"י הצבת z=1 בפולינום). זוהי משפחה רחבה מאוד של פולינומים אשר מהווה תת-קבוצה חשובה בתוך אוסף כל הפולינומים האפשריים. איבדנו כאן בהחלט כלליות אך הרווחנו תפיסה רקוסיבית כפי שטוענת הלמה הבאה:

למה 8.1: שימוש בפולינום הרקורסיבי אשר הוצע על אלגוריתם ה- SD יוצר את האלגוריתם הרקורסיבי הבא:

$$
\begin{aligned}
&y_1=[I-\gamma_1\alpha\mathbf{Q}]y_0+\gamma_1\alpha b\\
&\quad\vdots\\
&y_k=\rho_k\left[(I-\alpha\gamma_k\mathbf{Q})y_{k-1}+\alpha\gamma_k b\right]+(1-\rho_k)y_{k-2} \qquad k\ge 2
\end{aligned}
$$

הערה: הדמיון לשיטת ה- PARTAN ניכר מהיות אלגוריתם זה בעל משוואת עדכון כללית מסדר שני.

הוכחה: עבור k=1 ההוכחה מיידית ע"י כיתוב מפורש של $P_1(z)$:

$$
\begin{aligned}
&P_1(z)=\gamma_1 z-\gamma_1+1\\
&\Rightarrow\quad y_1=(1-\gamma_1)x_0+\gamma_1 x_1 \underset{\substack{x_0=y_0,\\ x_1=x_0+\alpha(b-\mathbf{Q}x_0)}}{\overset{\uparrow}{=}} [1-\gamma_1\alpha\mathbf{Q}]y_0+\gamma_1 b
\end{aligned}
$$

עבור k כללי נוכל לומר כי:

$$\begin{aligned}\varepsilon_k &= P_k([I-\alpha\mathbf{Q}])\varepsilon_0 = \\ &= \{\rho_k(\gamma_k[I-\alpha\mathbf{Q}]-\gamma_k+1)P_{k-1}([I-\alpha\mathbf{Q}])+(1-\rho_k)P_{k-2}([I-\alpha\mathbf{Q}])\}\varepsilon_0 = \\ &= \rho_k(\gamma_k[I-\alpha\mathbf{Q}]-\gamma_k+1)\varepsilon_{k-1}+(1-\rho_k)\varepsilon_{k-2}\end{aligned}$$

נוסיף לשני האגפים את הגודל x^* ומקבל:

$$\begin{aligned}y_k &= \rho_k(-\alpha\mathbf{Q}\gamma_k+1)\varepsilon_{k-1}+(1-\rho_k)\varepsilon_{k-2}+(1-\rho_k)x^*+\rho_k x^* = \\ &= \rho_k(-\alpha\mathbf{Q}\gamma_k+1)y_{k-1}+(1-\rho_k)y_{k-2}+\rho_k\gamma_k\alpha\mathbf{Q}x^* = \\ &= \rho_k(\gamma_k\alpha b+[I-\alpha\mathbf{Q}\gamma_k]y_{k-1})+(1-\rho_k)y_{k-2}\end{aligned}$$

כשהצבנו - $\mathbf{Q}x^*=b$, וטענת הלמה הוכחה. ■

נשים לב כי עבור הבחירה $\rho_k=1$ וכן:

$$\alpha\gamma_k=\frac{e_{k-1}{}^T e_{k-1}}{e_{k-1}{}^T\mathbf{Q}e_{k-1}} \quad \text{where} \quad e_k=b-\mathbf{Q}y_k$$

אנו מקבלים את ה- NSD כמקרה פרטי. עבור הבחירה הקודמת של $\alpha\gamma_k$ וחיפוש ערכו של ρ_k האופטימלי אנו מקבלים כמקרה פרטי את שיטת ה- PARTAN.

במקרה הכללי קיבלנו אלגוריתם חדש המקדם את y_k באופן רקורסיבי - השאלה היא האם אלגוריתם זה מתכנס ליעד x^*, ואם כן - באיזה קצב? בכל איטרציה יש במשוואת הרקורסיה שני פרמטרים אותם יש לקבוע ושאלה נוספת המתעוררת היא - כיצד יש לקובעם לקבלת ביצועים מיטביים. יש לשים לב לעובדה שהפרמטר החופשי ב- SD - הפרמטר α - כבר אינו חשוב לכשעצמו, אלא ערך המכפלה - $\alpha\cdot\gamma_k$, כך שללא אובדן כלליות נוכל לכתוב מעתה כי $\alpha=1$ -ובעצם להשמיטו. זה גם מסביר מדוע בחרנו לטפל ב- SD ולא ב- NSD - ההאצה הפולינומיאלית המוצעת כוללת בתוכה נירמול של גודל הצעד כפרמטר.

משפט 8.1: אם ידוע עבור האלגוריתם המואץ כי $\forall k \quad \rho_k,\gamma_k\neq 0$ והמטריצה $\mathbf{Q}$ חיובית מוגדרת, אזי אם האלגוריתם מתכנס, נקודת התכנסותו היא x^* בלבד.

הוכחה: קיבלנו קודם את הקשר:

$$\varepsilon_k = \rho_k (I - \gamma_k \mathbf{Q})\varepsilon_{k-1} + (1-\rho_k)\varepsilon_{k-2}$$

במצב יציב, צריך להתקיים:

$$\varepsilon_\infty = \rho_k (I - \gamma_k \mathbf{Q})\varepsilon_\infty + (1-\rho_k)\varepsilon_\infty$$
$$\Rightarrow \left[\rho_k (I - \gamma_k \mathbf{Q}) - \rho_k\right]\varepsilon_\infty = 0 \quad \Rightarrow \quad \rho_k \gamma_k \mathbf{Q}\varepsilon_\infty = 0$$

ומכיוון שמנתוני הלמה ידוע כי $\rho_k, \gamma_k \neq 0 \quad \forall k$ והמטריצה $\mathbf{Q}$ חיובית מוגדרת, הפתרון היחיד למערכת המשוואות שהתקבלה היא $\varepsilon_\infty = 0$ שפירוש התכנסות לנקודה x^*. ■

נציע לנקוט בדרך הבאה - בידינו אלגוריתם בעל שני פרמטרים לא ידועים - $\{\rho_k, \gamma_k\}_{k=1}^{\infty}$. בהנחה כי אנו כעת באיטרציה ה- k-1 (נמצאים בנקודה y_{k-1}) נוכל לשאול מהו צירוף הפרמטרים ρ_k, γ_k אשר יביא אותנו למינימום על המשטח הריבועי, כלומר לפתור את הבעיה הבאה:

$$\gamma_k, \rho_k = \underset{\gamma,\rho}{\operatorname{argmin}} [y_k - x^*]^T \mathbf{Q} [y_k - x^*]$$
$$\text{where} \quad y_k = \rho\left[(I - \gamma \mathbf{Q})y_{k-1} + \gamma b\right] + (1-\rho)y_{k-2}$$

גזירה והשוואה לאפס עפ"י הנעלמים - ρ_k, γ_k נותנת:

$$\frac{\partial f}{\partial \gamma} = 0 \quad \Rightarrow e_{k-1}^T \mathbf{Q}\left[\rho(y_{k-1} - y_{k-2}) + y_{k-2} + \rho\gamma e_{k-1}\right] = e_{k-1}^T b$$
$$\frac{\partial f}{\partial \rho} = 0 \quad \Rightarrow \left[(y_{k-1} - y_{k-2}) + \gamma e_{k-1}\right]^T \mathbf{Q}\left[\rho(y_{k-1} - y_{k-2}) + y_{k-2} + \rho\gamma e_{k-1} - x^*\right] =$$
$$= \left[(y_{k-1} - y_{k-2}) + \gamma e_{k-1}\right]^T b$$

ופתרון מערכת משוואות בשני נעלמים זאת תיתן את בחירת הפרמטרים ρ_k, γ_k האופטימליים. בהשוואה ל- PARTAN, אלגוריתם זה אמור להיות טוב יותר כיוון שב- PARTAN נבחר γ_k על-פי ה- NSD, והפרמטר השני חשוב כך שיתקבל מינימום על הישר. כאן אנו מלכתחילה מחפשים צמד ערכים שיתנו מינימום ולכן זה צפוי להיות מינימום נמוך יותר. אלגוריתם זה מובטח להתכנס כי הוא יכול להיות טוב יותר או זהה ל- NSD, ומכאן שיתקבל כי הוא בעל תכונת התכנסות גלובלית.

לסיכום סעיף זה, ראינו כי האצה פולינומיאלית של ה- SD מניבה כמקרים פרטיים את ה- NSD, וה- PARTAN. ישנן עוד שיטות התבויות מאחורי ייצוג זה - אחת מהן היא שיטה בה נעשה שימוש בפולינומי צ'ביצף. אנו לא נדון באלגוריתם המתקבל מהצבה זו במסגרת זו. הנקודה המרכזית אותה רצינו להעביר היא שאלגוריתם ה- PARTAN (שהינו גם אלגוריתם הגרדיאנטים הצמודים עבור הבעיה הריבועית) מתקבל גם מגישה של האצה פולינומיאלית, ובחירת פרמטרים אופטימליים במובן כלשהו.

8.3 האצת ה- SD בגישת המומנט

גישה אחרת לזירוזו של אלגוריתם ה- SD היא שימוש במגמה של ההתקדמות על מנת לזרז את עדכון הפתרון. משוואת האיטרציה המוצעת היא הבאה:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k [x_k - x_{k-1}] = x_k + \alpha_k e_k + \beta_k [x_k - x_{k-1}]$$

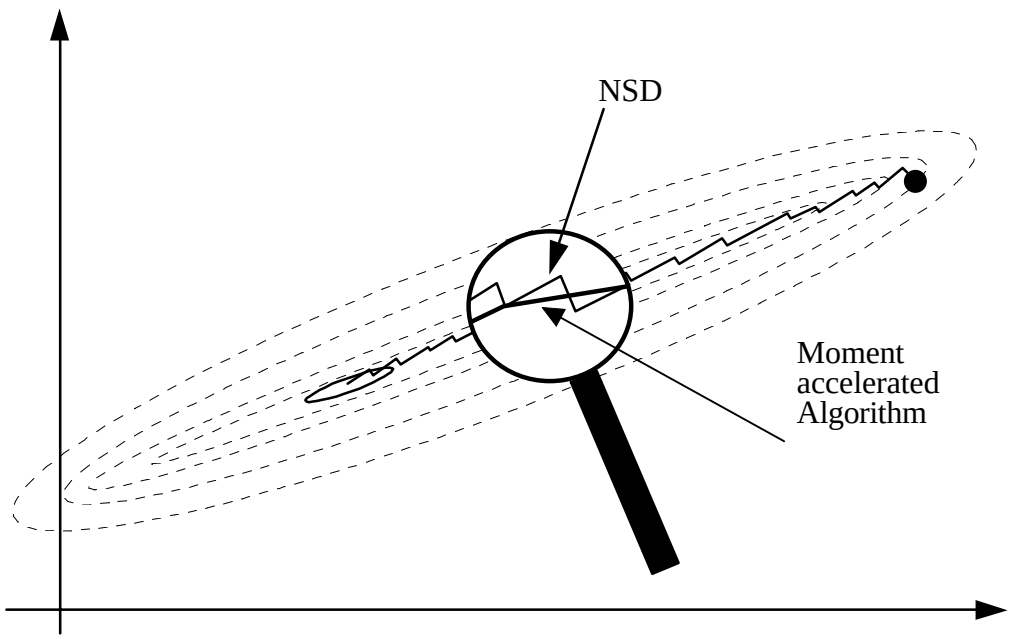
כשהגורם $\beta_k [x_k - x_{k-1}]$ בא להוסיף לעדכון הפתרון מרכיב אשר מגדיל את העדכון בכיוון שהותווה באיטרציה הקודמת. מהקשר הנ"ל ניכרים כמה פרטים חשובים:

1. עבור בחירת $\beta_k = 0$ מתקבלים כמקרים פרטיים האלגוריתמים ה- SD וה- NSD (בתלות בבחירת הפרמטר α_k).
2. בשיטת ההאצה הפולינומיאלית - למה 8.1 - קיבלנו את המשוואה האיטרטיבית הבאה:

$$\begin{aligned} y_k &= \rho_k y_{k-1} + \rho_k \gamma_k [b - \mathbf{Q} y_{k-1}] + (1-\rho_k) y_{k-2} = \\ &= y_{k-1} + \rho_k \gamma_k e_k + (\rho_k - 1)[y_{k-1} - y_{k-2}] \end{aligned}$$

לכן - בחירת $\beta_k = \rho_k - 1$, $\alpha_k = \rho_k \gamma_k$ יוצרת התלכדות של שתי גישות אלו, והן למעשה זהות. כיוון ששתי גישות אלו זהות, כל התוצאות אשר הופקו בסעיף הקודם רלוונטיות גם כאן. בפרט, נקודת ההתכנסות, ודרך חישוב הפרמטרים האופטימליים אשר יתנו ירידה מירבית.

3. יעילותה של גישה זו מתבטא בצורה מובהקת עבור מצבים בהם ה- Condition # של ההסיין הוא גרוע (כלומר גבוה). במצבים אלו נראה המשטח עליו מבוצעת המינימיזציה כעמק צר וארוך היורד לאט. שימוש באלגוריתם ה- SD נותן התנדנדות בין שתי השלוחות החובקות אותו, בעוד שאלגוריתם זה מתקדם במורד העמק במהירות רבה יותר. ציור 8.5 בא להמחיש תיאור זה.



ציור 8.5 - המחשת שיטת ההאצה בגישת המומנט

ננסה להבין את תכונות האלגוריתם המוצע בהשוואה ל- SD. בסעיף הקודם הראינו כיצד צריך לקבוע את פרמטרי השיטה לקבלת ירידה מירבית בפונקציה. כעת נראה גישה אחרת בה הפרמטרים אינם תלוי איטרציה, ונמחיש עבור מקרה זה את תנאי ההתכנסות וקצב התכנסות. ברור כי כל אשר נאמר במסגרת זו תקף גם לתפיסת ההאצה הפולינומיאלית.

משפט 8.2: יהיו $\{\mu_j\}_{j=1}^{n}$ ערכים עצמיים של המטריצה $\mathbf{Q}$. נתון האלגוריתם המואץ עם הפרמטרים α,β. אזי האלגוריתם המואץ מתכנס עבור כל צירוף פרמטרים בתחום:

$$0<\alpha<\frac{2(1+\beta)}{A} \quad ; \quad -1<\beta<1$$

כאשר A הוא הערך העצמי המקסימלי של $\mathbf{Q}$.

הוכחה: על פי המשוואה הרקורסיבית של האלגוריתם אשר ראינו קודם:

$$x_{k+1}=x_k+\alpha e_k+\beta[x_k-x_{k-1}]$$
$$\Rightarrow x_{k+1}-x^*=\varepsilon_{k+1}=x_k-x^*+\alpha e_k+\beta[x_k-x^*-x_{k-1}+x^*]=$$
$$=\varepsilon_k+\alpha\mathbf{Q}[x^*-x_k]+\beta[\varepsilon_k-\varepsilon_{k-1}]$$

$$\Rightarrow \varepsilon_{k+1} = \left[I + \beta - \alpha \mathbf{Q}\right]\varepsilon_k - \beta\varepsilon_{k-1}$$

ולכן, ניתן להציג קשר זה באופן מטריצי ע"י:

$$v_k = \begin{bmatrix} \varepsilon_k \\ \varepsilon_{k-1} \end{bmatrix} = \begin{bmatrix} (I + \beta I - \alpha \mathbf{Q}) & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{k-1} \\ \varepsilon_{k-2} \end{bmatrix} = \mathbf{T} v_{k-1}$$

ובצורת רישום זו ניכר מיד כי אם **T** מטריצה בה כל הערכים העצמיים בתוך מעגל היחידה, תחול התכנסות של השגיאה. כלומר נדרוש (לשם איתור הערכים העצמיים):

$$\det\{\mathbf{T}\} = \left\{ \begin{bmatrix} (I + \beta I - \alpha \mathbf{Q}) & -\beta I \\ I & 0 \end{bmatrix} - \lambda \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \right\} = 0$$

בשל היות מטריצה **Q** חיובית מוגדרת ניתן לייצגה כ- $\mathbf{Q} = U^* \Delta U$ ומתקבל:

$$\begin{aligned} \det\{\mathbf{T}\} &= \det\left\{-\lambda(I + \beta I - \alpha \mathbf{Q}) + \lambda^2 I + \beta I\right\} = \\ &= \det\{U^*\}\det\{U\} \cdot \det\left\{-\lambda(I + \beta I - \alpha \Delta) + \lambda^2 I + \beta I\right\} = \\ &= \det\left\{-\lambda(I + \beta I - \alpha \Delta) + \lambda^2 I + \beta I\right\} = 0 \end{aligned}$$

היתרון ברישום המשוואה האחרונה הוא בכך שהיא אלכסונית, ולכן - איפוס הדטרמיננט שלה מוגשם ע"י איפוס אחד מאיברי אלכסונה. לכל μ_j יתקבלו שני ערכים עצמיים של **T** ע"י פתרון המשוואה הריבועית:

$$\lambda^2 - \lambda\left(1 + \beta - \alpha\mu_j\right) + \beta = 0$$

ובסך הכל יתקבלו 2n ערכים עצמיים של **T**. התכנסות תתקבל אם ערכם המוחלט של כל ערכים עצמיים אלו קטן מ- 1. נכתוב באופן מפורש את הדרישה על שורשי המשוואה הריבועית:

$$\left|\frac{\left(1+\beta-\alpha\mu_j\right)\pm\sqrt{\left(1+\beta-\alpha\mu_j\right)^2-4\beta}}{2}\right|<1$$

$$\Rightarrow\quad -2<\left(1+\beta-\alpha\mu_j\right)\pm\sqrt{\left(1+\beta-\alpha\mu_j\right)^2-4\beta}<2$$

$$-3-\beta+\alpha\mu_j<\pm\sqrt{\left(1+\beta-\alpha\mu_j\right)^2-4\beta}<1-\beta+\alpha\mu_j$$

$$-3-\beta+\alpha\mu_j<\pm\sqrt{\left(1+\beta-\alpha\mu_j\right)^2-4\beta}<1-\beta+\alpha\mu_j$$

עבור סימן ה- (+) לפני השורש נקבל:

$$\left(1+\beta-\alpha\mu_j\right)^2-4\beta<\left(1-\beta+\alpha\mu_j\right)^2$$

$$\Rightarrow\quad -2\beta-2(1+\beta)\alpha\mu_j<-2\beta+2(1-\beta)\alpha\mu_j$$

$$\Rightarrow\quad 0<2\alpha\mu_j\quad\Rightarrow\quad \alpha>0$$

ועבור סימן ה - (-) נקבל:

$$\left(3+\beta-\alpha\mu_j\right)^2>\left(1+\beta-\alpha\mu_j\right)^2-4\beta$$

$$\Rightarrow\quad (3+\beta)^2-2(3+\beta)\alpha\mu_j>(1+\beta)^2-2(1+\beta)\alpha\mu_j-4\beta$$

$$\Rightarrow\quad 8+8\beta-4\alpha\mu_j>0$$

$$\Rightarrow\quad \alpha<\frac{2(1+\beta)}{\mu_j}$$

ומכיוון שדרישות אלו צריכות להתקיים לכל j מתקבל:

$$0<\alpha<\frac{2(1+\beta)}{A}$$

כאשר A הוא הערך העצמי הגודל ביותר של **Q**. עבור $A>>1$ שורשי המשוואה הם:

$$\frac{\left(1+\beta-\alpha\mu_j\right)\pm\sqrt{\left(1+\beta-\alpha\mu_j\right)^2-4\beta}}{2}\cong\frac{\left(1+\beta-\alpha\mu_j\right)\pm(1-\beta)}{2}=\left(1-\frac{1}{2}\alpha\mu_j\right);\left(\beta-\frac{1}{2}\alpha\mu_j\right)$$

ולכן נדרוש עבור שני שורשים אלו:

$$-1<\left(1-\frac{1}{2}\alpha\mu_j\right)<1 \quad\Rightarrow\quad 0<\frac{\mu_j}{A}(1+\beta)<2$$
$$\Rightarrow\quad -1<\beta<\frac{2A}{\mu_j}-1 \quad\Rightarrow\quad -1<\beta<1$$

$$-1<\left(\beta-\frac{1}{2}\alpha\mu_j\right)<1 \quad\Rightarrow\quad \frac{1}{2}\alpha\mu_j-1<\beta<\frac{1}{2}\alpha\mu_j+1$$
$$\Rightarrow\quad \frac{1+\beta}{A}\mu_j-1<\beta<1 \quad\Rightarrow\quad -1<\beta<1$$

ובכך הוכחנו את המשפט. ■

ניתן להראות (ואנו לא נעשה זאת במסגרת זו) כי קצב ההתכנסות של אלגוריתם ההאצה מבוססת המומנט היא ליניארית עם יחס התכנסות הנתון ע"י:

$$\left[\frac{\sqrt{r}-1}{\sqrt{r}+1}\right]^2 \approx \left(1-1/\sqrt{r}\right)^4$$

וזהו שיפור משמעותי ביחס להתכנסות ה- NSD הנתונה ע"י:

$$\left[\frac{r-1}{r+1}\right]^2 \approx (1-1/r)^4$$

8.4 סיכום

בפרק זה ראינו מספר גישות שונות אשר מנסות ייעול של ה- SD. למרות מקורותיהן השונים של שיטות אלו, כולן מגיעות לאחר מעשה לאלגוריתם דומה בו יש זכרון מסדר שני (כלומר תוצאת האיטרציה הנוכחית תלויה בתוצאות שתי האיטרציות הקודמות). טענו כי התוצאה הזו הינה אלגוריתם הגרדיאנטים הצמודים, אך לא הזכרנו מהו אלגוריתם זה ולא הוכחנו את הקשר אליו. פרק הבא מגדיר באופן מפורט את מושג הכיוונים הצמודים וממנו מצמיח את אלגוריתם ה- CG.

פרק 9 - אלגוריתם הגרדיאנטים הצמודים - CG

בפרק הקודם ראינו מגוון דרכים אשר נועדו לזרז קצבי התכנסות של אלגוריתמים איטרטיביים אשר הופעלו על ה- SD. התקבלו מגוון אלגוריתמים אשר לגבי כולם טענו כי הם מובילים לאלגוריתם הגרדיאנטים הצמודים בפרק זה נכיר מכיוון שונה בתכלית את אלגוריתם ה- CG (Conjugate Gradient). הדיון יתחיל בהצגת ה- CG עבור בעיות ריבועיות, ולאחר מכן תורחב היריעה לדיון בבעיות כלליות.

9.1 שיטת הכיוונים הצמודים

הגדרה 9.1: בהינתן מטריצה חיובית מוגדרת $\mathbf{Q}$ ושני וקטורים כלשהם v ו- u, וקטורים אלו קרויים **Q-**אורתוגונליים או צמודים זה לזה אם מתקיים כי $v^T\mathbf{Q}u = 0$.

למה 9.1: תהי $\mathbf{Q}$ מטריצה חיובית מוגדרת כלשהי בגודל $[n \times n]$, ויהיו $\{v_k\}_{k=0}^{n-1}$ סידרת וקטורים **Q-** אורתוגונליים. אזי וקטורים אלו בלתי תלוים ליניארית, ולכן פורסים את המרחב.

הוכחה: נניח כי וקטורים אלו תלויים ליניארית פירוש הדבר שקיימים n מקדמים סקלריים שונים מ- 0 (לפחות אחד מהם חייב להיות שונה מאפס) כך ש:

$$\sum_{k=0}^{n-1} a_k v_k = 0$$

ע"י הכפלת צירוף זה משמאל בגורם $v_j^T\mathbf{Q}$ נקבל:

$$v_j^T\mathbf{Q}\sum_{k=0}^{n-1} a_k v_k = \sum_{k=0}^{n-1} a_k v_j^T\mathbf{Q}v_k = a_j = 0$$

ובאופן דומה נוכל להראות כי כל המקדמים שווים לאפס. זה כמובן סותר את הנחתנו הראשונית שהם תלוים ליניארית, ולכן הלמה הוכחה. ■

כעת, נניח כי בידנו בעיית מינימיזציה ריבועית אותה אנו רוצים לפתור:

$$f(x) = \frac{1}{2}(x - x^*)^T\mathbf{Q}(x - x^*) = \frac{1}{2}x^T\mathbf{Q}x - b^Tx + c$$

כיוון ש- $\{v_k\}_{k=0}^{n-1}$ הינם כיוונים צמודים אשר מהוים בסיס, נוכל לייצג את הפתרון ע"י הצירוף:

$$x^* = \sum_{k=0}^{n-1} a_k v_k = \mathbf{Q}^{-1} b$$

ברור כי ידיעת מקדמי הצירוף הליניארי היא ידיעת הפתרון. ניתן לקבל מקדמים אלו בקלות ע"י הכפלה ב- $v_j^T \mathbf{Q}$:

$$v_j^T \mathbf{Q} x^* = v_j^T b = \sum_{k=0}^{n-1} a_k v_j^T \mathbf{Q} v_k = a_j v_j^T \mathbf{Q} v_j \quad \Rightarrow \quad a_j = \frac{v_j^T b}{v_j^T \mathbf{Q} v_j}$$

כלומר נוכל להציע את האלגוריתם הבא - אתחל ע"י וקטור אפסים - $x_0 = 0$. עדכן פתרון זה באופן איטרטיבי ע"י:

$$x_{k+1} = x_k + \frac{v_k^T b}{v_k^T \mathbf{Q} v_k} v_k \qquad k = 0,1,2,\ldots,n-1$$

וברור כי יתקבל כי $x_n = x^*$. אלגוריתם זה אינו מוצלח כיוון שהוא מחייב איתחול אפס. האלגוריתם המתואר בציור 9.1 ואשר נקרא אלגוריתם הכיוונים הצמודים (Conjugate Directions) פועל באופן דומה אך עם איתחול כלשהו.

1. בנה סידרת וקטורים $\{v_k\}_{k=0}^{n-1}$ אשר תהיה $\mathbf{Q}$ אורתוגונלית.
2. אתחל את האלגוריתם עם x_0 כלשהו.
3. באיטרציה ה- k חשב את המקדם $\alpha_k = \frac{-v_k^T(\mathbf{Q}x_k - b)}{v_k^T \mathbf{Q} v_k}$
4. עדכן את הפתרון לפי - $x_{k+1} = x_k + \alpha_k v_k$.
5. עצור את האלגוריתם אם $\|\mathbf{Q}x_k - b\| \le \varepsilon$ או באיטרציה ה- n.

<u>ציור 9.1 - אלגוריתם הכיוונים הצמודים - Conjugate Directions לבעיה הריבועית</u>

משפט 9.1: האלגוריתם המתואר בציור 8.1 מתכנס תוך n איטרציות לוקטור x^* ללא שגיאה.

הוכחה: האיטרציה ה- k באלגוריתם היא - $x_{k+1} = x_k - \frac{v_k^T(\mathbf{Q}x_k - b)}{v_k^T\mathbf{Q}v_k}v_k$. החסרת x^* משני האגפים והכפלה משמאל ב- $v_k^T\mathbf{Q}$ נותנת:

$$v_k^T\mathbf{Q}[x_{k+1} - x^*] = v_k^T\mathbf{Q}[x_k - x^*] - v_k^T[\mathbf{Q}x_k - b] \underset{\substack{\uparrow \\ x^* = \mathbf{Q}^{-1}b}}{=} 0$$

פירוש הדבר ששגיאת האיטרציה ה- k - $[x_{k+1} - x^*]$ - היא $\mathbf{Q}$ אורתוגונלית לוקטור v_k. למעשה, שגיאה זו היא $\mathbf{Q}$ אורתוגונלית לכל הוקטורים $\{v_j\}_{j=0}^{k}$ כיוון ש:

$$\begin{aligned}
&x_{k+1} - x^* = x_k - x^* - \frac{v_k^T(\mathbf{Q}x_k - b)}{v_k^T\mathbf{Q}v_k}v_k \\
&\Rightarrow v_{k-1}^T\mathbf{Q}[x_{k+1} - x^*] = v_{k-1}^T\mathbf{Q}[x_k - x^*] = 0 \\
&\Rightarrow v_{k-2}^T\mathbf{Q}[x_{k+1} - x^*] = v_{k-2}^T\mathbf{Q}[x_k - x^*] = v_{k-2}^T\mathbf{Q}[x_{k-1} - x^*] = 0
\end{aligned}$$

ובאינדוקציה ניתן להראות זאת עד הוקטור הראשון בקבוצה. לכן, כאשר אנו באיטרציה ה- n, נקבל כי:

$$\forall 1 \le j \le n-1 \quad v_j^T\mathbf{Q}[x_n - x^*] = 0$$

והוקטור היחיד שיכול להיות צמוד לכל הסידרה $\{v_k\}_{k=0}^{n-1}$ הוא וקטור האפס בלבד. ■

תוצאת לוואי של ההוכחה הנ"ל היא שבכל איטרציה k של האלגוריתם מתקבל כי הגרדיאנט המקומי החדש ניצב לכל הכיוונים הצמודים שכבר נעשה בהם שימוש:

$$\forall 1 \le j \le k \quad v_j^T\mathbf{Q}[x_{k+1} - x^*] = v_j^T[\mathbf{Q}x_{k+1} - b] = -v_j^T e_{k+1} = \nabla f^T(x_{k+1})v_j = 0$$

ופירוש הדבר הוא שניתן להשתמש בגרדיאנט החדש e_{k+1} שהינו אורתוגונלי לכל הוקטורים $\{v_j\}_{j=0}^{k}$, כוקטור אשר ע"י תהליך גרהם-שמידט יומר להיות v_{k+1}.

למעשה, כל שהוכחנו עד כה הוא שאלגוריתם זה מתכנס תוך n איטרציות לפתרון המדויק. בינתיים, אנו לא יכולים אף להבטיח ששגיאת הפתרון דועכת - בהחלט יתכן שהשגיאה תעלה ותרד במהלך האיטרציות - השגיאה מתאפסת באיטרציה האחרונה. המשפט הבא בא לחזק את התוצאה הנ"ל של אלגוריתם ה- CD ולקבוע כי חלה ירידה ממש במהלך האיטרציות

משפט 9.2: אלגוריתם אשר מבצע עדכון איטרטיבי של הפתרון לפי המשוואה $x_{k+1} = x_k + \alpha_k v_k$, כאשר $\{v_k\}_{k=0}^{n-1}$ היא סידרה $\mathbf{Q}$ אורתוגונלית, והפרמטר α_k נקבע לקבלת מינימום על הישר $x = x_k + \alpha_k v_k$ מניב את אלגוריתם ה- CD. כפועל יוצא של גישה זו - בכל איטרציה אנו מגשימים בפועל את מציאת המינימום של הבעיה בתוך תת המרחב הבא:

$$M_k = \left\{ x \,\middle|\, x = x_0 + v,\ \ v \in \text{subspace spanned by } \{v_j\}_{j=0}^{k} \right\}$$

הוכחה: נראה תחילה כי α_k שמגשים את המינימום על הישר יוצא אותו הביטוי אשר הוצג ב- CD:

$$\begin{aligned}\frac{\partial f(x_k + \alpha v_k)}{\partial \alpha} &= \frac{1}{2} \cdot \frac{\partial}{\partial \alpha} [x_k + \alpha v_k - x^*]^T \mathbf{Q} [x_k + \alpha v_k - x^*] = \\ &= \nabla f^T (x_k + \alpha_k v_k) v_k = -e_{k+1}^T v_k = 0\end{aligned}$$

התקבל כי α_k נובע מהמשוואה בה נדרשת ניצבות אשר כבר קיבלנו קודם. ביטוי מפורש עבור α_k הוא:

$$e_{k+1}^T v_k = [\mathbf{Q}(x_k + \alpha_k v_k) - b] v_k = 0 \quad \Rightarrow \quad \alpha_k = \frac{b v_k - x_k^T \mathbf{Q} v_k}{v_k^T \mathbf{Q} v_k} = \frac{-v_k^T (\mathbf{Q} x_k - b)}{v_k^T \mathbf{Q} v_k}$$

כך שאנו רואים כי האלגוריתם המוצע מתלכד עם ה- CD. נראה כעת את קיומה של הטענה השניה. נניח כי רצוננו למצוא את נקודת המינימום של הפונקציה הריבועית במרחב M_k. לכן, נחפש את k+1 הפרמטרים אשר יתנו מינימום לפי:

$$x_{k+1} = x_0 + \beta_0 v_0 + \beta_1 v_1 + \cdots + \beta_k v_k \quad \text{where}$$
$$\frac{\partial f\left(x_0 + \beta_0 v_0 + \beta_1 v_1 + \cdots + \beta_k v_k\right)}{\partial \beta_j} = \nabla f^T\left(x_{k+1}\right) v_j = 0$$

אבל בהוכחה של המשפט הקודם ראינו כי הגרדיאנט בנקודה ה- k+1 ניצב לכל הוקטורים $\left\{v_j\right\}_{j=0}^{k}$, כך שאמנם המקדמים מושג מינימום על פני התת-מרחב M_k. באופן זה יוצא כי כשמגיעים ל- k=n-1 תת-מרחב זה הוא כל המרחב, ולכן מאותרת נקודת המינימום המדויקת של הבעיה הריבועית. ■

למעשה, האלגוריתם המצוי בידינו אינו מוצלח כפי שהוא עלול להיחשב, מכיוון שמרכיב חיוני בו הוא הרכבת או מציאת סט של וקטורים **Q** אורתוגונליים, וזוהי עלולה להיות משימה יקרה מבחינה חישובית. נראה כיצד פעולה זו נעשית על קבוצת וקטורים כללית בגישת גרהם-שמידט, כך שהמרחב הנפרס מכל תת-קבוצה של וקטורים אלו מתלכד.

בהינתן וקטור ראשון z_0, נקבע את $v_0 = z_0$. נניח כי אנו בשלב ה- k, דהיינו הצלחנו להמיר את הקבוצה $\left\{z_j\right\}_{j=0}^{k}$ בקבוצה **Q** אורתוגונלית $\left\{v_j\right\}_{j=0}^{k}$. כעת אנו מקבלים את z_{k+1}. נגדיר:

$$v_{k+1} = z_{k+1} + \sum_{j=0}^{k} c_{k+1}^{j} v_j$$

אזי צריך להתקיים:

$$\forall 0 \le m \le k \quad v_m^T \mathbf{Q} v_{k+1} = 0 = v_m^T \mathbf{Q} z_{k+1} + \sum_{j=0}^{k} c_{k+1}^{j} v_m^T \mathbf{Q} v_j = v_m^T \mathbf{Q} z_{k+1} + c_{k+1}^{m} v_m^T \mathbf{Q} v_m$$
$$\Rightarrow c_{k+1}^{m} = \frac{-v_m^T \mathbf{Q} z_{k+1}}{v_m^T \mathbf{Q} v_m} \quad \Rightarrow \quad v_{k+1} = z_{k+1} - \sum_{j=0}^{k} \frac{v_j^T \mathbf{Q} z_{k+1}}{v_j^T \mathbf{Q} v_j} v_j$$

כשאנו באיטרציה ה- k של התהליך הנ"ל, תת-המרחב שנפרס ע"י הוקטורים המקוריים עד נקודה זו זהה לתת-המרחב שנפרס ע"י הוקטורים המיוצרים בתהליך - זוהי תכונה בסיסית של תהליך גרהם שמידט התקפה גם כאן.

כפי שאנו רואים, חישוב זה הולך וגדל עם התקדמות k. כדי לחשב את כל n הוקטורים נידרש ל- $n(n-1)/2$ חישובי הקבועים c. כך שבניה של סט וקטורים כזו בד"כ אינה דבר פשוט כלל. כמו כן יש לזכור כי בגישה זו יש לזכור את סט הוקטורים המתקבל למן התחלתו - חסרון נוסף.

בסעיף הבא אנו נראה מקרה פרטי ייחודי של ה- CD אשר מניב סט וקטורים Q אורתוגונליים בפשטות.

הערה: ניתן לראות די בפשטות כי אם קבוצת וקטורים היא אורתונורמלית, אין הדבר סותר את יכולתה להיות גם **Q**-אורתוגונלית. למעשה - האפשרות היחידה לקבלת תכונה כפולה זו של אורתונוגונליות היא בוקטורים העצמיים של המטריצה **Q**. אם נסמן את הוקטורים העצמיים של **Q** כ- $\{\theta_k\}_{k=1}^{n}$, ואת הערכים העצמיים המתאימים כ- $\{\lambda_k\}_{k=1}^{n}$, אזי, בשל היותה של **Q** חיובית מוגדרת מתקבל כי קבוצה זו אורתונורמלית. לכן:

$$\theta_k^T \mathbf{Q} \theta_j \underset{\substack{\uparrow \\ \mathbf{Q}\theta_j = \lambda_j \theta_j}}{=} \theta_k^T \lambda_j \theta_j = \lambda_j \theta_k^T \theta_j = \lambda_j \delta(k-j)$$

מהאמור לעיל גם נובע כי לא תיתכן **אורתונורמליות** כפולה - למעט עבור מטריצה בה כל הערכים העצמיים הם 1.

9.2 אלגוריתם הגרדיאנטים הצמודים - CG

כבר ראינו בהקשרים קודמים כי הוקטורים $e_k = b - \mathbf{Q}x_k$ הם בחירה חכמה לכיוון התקדמות כיוון שהם מגשימים ירידה מירבית בפונקציה הריבועית. נציע אם כך שבאיטרציה ה- k+1 יחושב הוקטור $e_k = b - \mathbf{Q}x_k$, ועליו יחד עם קבוצת הכיוונים הקודמים תיעשה פעולת גרהם שמידט אשר תוארה קודם. כלומר:

$$v_k = e_k - \sum_{j=0}^{k-1} \frac{v_j^T \mathbf{Q} e_k}{v_j^T \mathbf{Q} v_j} v_j \quad ; \quad x_{k+1} = x_k + \alpha_k v_k$$

הגודל α_k יהיה כפי שקיבלנו עבור אלגוריתם ה- CD (זהו הרי מקרה פרטי של ה- CD). כאן בא החלק אשר הופך את הגישה הנוכחית לכה יעילה - אנו מיד נראה כי בניגוד למקרה גרהם-שמידט הכללי, כאן אין צורך ביותר מהאיבר האחרון בסכימה, דהיינו, מתקבל קשר רקורסיבי פשוט לבנייתו של הכיוון v_k.

למה 9.2: עבור אלגוריתם ה- CD אשר עושה שימוש בגרדיאנטים קיים הקשר הבא:

$$v_k = e_k - \sum_{j=0}^{k-1} \frac{v_j^T \mathbf{Q} e_k}{v_j^T \mathbf{Q} v_j} v_j = e_k + \frac{e_k^T e_k}{e_{k-1}^T e_{k-1}} v_{k-1}$$

הוכחה: כיוון ש- $x_{k+1} = x_k + \alpha_k v_k$, מתקבל - $e_{k+1} - e_k = \mathbf{Q}(x_{k+1} - x_k) = \alpha_k \mathbf{Q} v_k$. לכן:

$$v_j^T \mathbf{Q} e_k = \frac{1}{\alpha_j}\left(e_{j+1} - e_j\right)^T e_k$$

לפי משפט 9.1, קיבלנו כי $e_k \perp v_i \quad \forall 0 \le i \le k-1$ אך גם קיבלנו מתהליך גרהם שמידט כי $\mathrm{span}\{v_i\}_{i=0}^{k-1} = \mathrm{span}\{e_i\}_{i=0}^{k-1}$, ולכן $e_k \perp e_i \quad \forall 0 \le i \le k-1$. לכן מתקבל:

$$v_j^T \mathbf{Q} e_k = \frac{1}{\alpha_j}\left(e_{j+1} - e_j\right)^T e_k = \begin{cases} \frac{1}{\alpha_{k-1}} e_k^T e_k & j = k-1 \\ 0 & 0 \le j < k-1 \end{cases}$$

אם נחזור לביטוי המקורי של תהליך גרהם שמידט קיבלנו:

$$v_k = e_k - \sum_{j=0}^{k-1} \frac{v_j^T \mathbf{Q} e_k}{v_j^T \mathbf{Q} v_j} v_j = e_k - \frac{e_k^T e_k}{\alpha_{k-1} v_{k-1}^T \mathbf{Q} v_{k-1}} v_{k-1}$$

אך בשל הקשר אשר ראינו קודם - $e_{k+1} - e_k = \mathbf{Q}(x_{k+1} - x_k) = \alpha_k \mathbf{Q} v_k$ - נקבל כי:

$$v_j^T \mathbf{Q} v_j = \frac{1}{\alpha_j}\left(e_{j+1} - e_j\right)^T v_j = \frac{-1}{\alpha_j} e_j^T v_j$$

הגורם השני נפל בשל ניצבות. שימוש בנוסחת גרהם שמידט המקורית נותן:

$$v_{k-1}^T \mathbf{Q} v_{k-1} = \frac{-1}{\alpha_{k-1}} e_{k-1}^T v_{k-1} = \frac{-1}{\alpha_{k-1}} e_{k-1}^T \left[e_{k-1} - \sum_{j=0}^{k-2} \frac{v_j^T \mathbf{Q} e_k}{v_j^T \mathbf{Q} v_j} v_j \right] = \frac{-1}{\alpha_{k-1}} e_{k-1}^T e_{k-1}$$

ולסיכום מתקבל:
$$v_k = e_k + \frac{e_k^T e_k}{e_{k-1}^T e_{k-1}} v_{k-1}$$

נקודה אחרונה שנותרה לטיפול היא האפשרות שהוקטור e_k יימצא במרחב M_{k-1} - דהיינו -תלוי ליניארית בגרדיאנטים הקודמים לו. אפשרות זו פירושה כי לא ניתן לפתח את הוקטור ה- $\mathbf{Q}$ אורתוגונלי הבא והאלגוריתם ייתקע. ובכן - מסתבר כי אפשרות כזו קיימת רק אם מתקבל $e_k = 0$, ואז ממילא פירוש הדבר הגעה לנקודת האופטימום וניתן לעצור את האלגוריתם. נניח אם כך כי $e_k \neq 0$, ונראה כי בהכרח פירוש הדבר כי הוא לא תלוי ליניארית בקודמיו. מתכונת הניצבות מתקבל כי חייב להתקיים כי $e_k \perp v_i \quad \forall 0 \le i \le k-1$ וכמקודם $e_k \perp e_i \quad \forall 0 \le i \le k-1$, כיוון ש- $\mathrm{span}\{v_i\}_{i=0}^{k-1} = \mathrm{span}\{e_i\}_{i=0}^{k-1}$. לכן - לא יתכן שהוקטור e_k יהיה תלוי בקודמיו. ■

לסיכום נרכז את אלגוריתם ה- CG כפי שמתואר בציור 9.2. אלגוריתם זה הוא בדיוק האלגוריתם אשר התקבל בגישת ה- PARTAN (למרות שנדרשת אלגברה מייגעת כדי להראות זאת), והאלגוריתם אשר התקבל כמקרה פרטי של ההאצה הפולינומיאלית.

1. אתחל את האלגוריתם עם x_0 כלשהו.
2. קבע את כיוון החיפוש כ- $v_0 = e_0 = b - \mathbf{Q}x_0$.
3. באיטרציה ה- k חשב את $\beta_k = \frac{e_k^T e_k}{e_{k-1}^T e_{k-1}}$.
4. קבע את כיוון החיפוש כ- $v_k = e_k + \beta_k v_{k-1}$
5. חשב את המקדם $\alpha_k = \frac{v_k^T e_k}{v_k^T \mathbf{Q} v_k}$ (מינימום על ישר).
6. עדכן את הפתרון לפי - $x_{k+1} = x_k + \alpha_k v_k$.
5. עצור את האלגוריתם אם $\|\mathbf{Q}x_k - b\| \le \varepsilon$ או באיטרציה ה- n.

<u>ציור 9.2 - אלגוריתם הכיוונים הצמודים - Conjugate Directions עבור הבעיה הריבועית</u>

אלגוריתם זה דורש כמות כפולה בקירוב של חישובים וזכרון בהשוואה ל- NSD. לכן טיבעי שנדרוש שמבחינת קצב התכנסות יחול שיפור משמעותי בהשוואה לאלגוריתם זה.

9.3 התכנסות אלגוריתם ה- CG

בשל היותו של ה- CG תהליך האצה של ה- NSD, ברור כי ביצועיו טובים יותר, וברור כי הוא בעל תכונת התכנסות גלובלית. לכן אלגוריתם ה- CG הוא בעל סדר התכנסות ליניארי לפחות, ויחס התכנסות טוב מזה של ה- NSD. יתרה מזו - כיוון שאלגוריתם ה- CG מתכנס לתוצאה המדוייקת ב- n איטרציות, פירוש הדבר שהוא איננו איטרטיבי במהותו, אלא תהליך סידרתי של בניית פתרון. למרות זאת מקובל לנתחו כתהליך איטרטיבי מסיבה מרכזית אחת - במרבית המקרים אנו נסתפק בהרבה פחות מ- n איטרציות לקבלת פתרון מרוב, ונסתפק בזה.

הדרך הנוחה ביותר לדון בתכונות ההתכנסות של ה- CG היא דרך ייצוגו כתהליך האצה פולינומיאלי, בדומה לייצוג שניתן בפרק הקודם. לצורך כך נשתמש בתכונה הבאה.

הגדרה 9.2: בהינתן וקטור כלשהו $v \in R^n \;;\; v \neq 0$, ומטריצה $\mathbf{Q}$ חיובית מוגדרת, מרחב קרילוב ממימד k מוגדר כמרחב הנפרס ע"י הוקטורים $\left\{v,\ \mathbf{Q}v,\ \mathbf{Q}^2v,\ \ldots,\mathbf{Q}^{k-1}v\right\} = \left\{\mathbf{Q}^j v\right\}_{j=0}^{k-1}$.

אנו רואים כי להגדרת מרחב קרילוב יש קשר לייצוג הפולינומי שהוצג בפרק קודם. אנו נעמוד בהמשך על קשר זה. למעשה, אוסף הוקטורים שתואר כפורס את מרחב קרילוב הוא אוסף בלתי תלוי ליניארית -תחת תנאים מסויימים - נראה זאת.

למה 9.3: בהינתן וקטור <u>כלשהו</u> $v \in R^n \;;\; v \neq 0$, ומטריצה $\mathbf{Q}$ חיובית מוגדרת בעלת n ערכים עצמיים <u>שונים זה מזה</u>, קבוצת הוקטורים $\left\{v,\ \mathbf{Q}v,\ \mathbf{Q}^2v,\ \ldots,\mathbf{Q}^{k-1}v\right\} = \left\{\mathbf{Q}^j v\right\}_{j=0}^{k-1}$ בלתי תלויה ליניארית, כאשר $k \leq n$.

הוכחה: נניח כי קבוצה זו תלויה ליניארית. לכן קיימים מקדמים שונים מאפס (לפחות אחד מהם) שייתנו כי:

$$\sum_{j=0}^{k-1} \alpha_j \mathbf{Q}^j v = \left[\sum_{j=0}^{k-1} \alpha_j \mathbf{Q}^j\right] v = 0$$

קיבלנו כי פולינום של $\mathbf{Q}$ מסדר k כופל את הוקטור v. כיוון ש- v וקטור כללי כלשהו השונה מאפס, עלינו לדרוש את התאפסות המטריצה הכופלת אותו על מנת לקבל שיויון לאפס. פירוש הדבר שקיים פולינום מסדר נמוך מ- n אשר $\mathbf{Q}$ מאפס אותו - עובדה העומדת בסתירה למשפט הפולינום האופייני

(הפולינום המינימלי של מטריצות בהן כל הערכים העצמיים שונים הוא הפולינום האופייני) - לכן ברור כי וקטורים אלו בלתי תלויים. ■

משפט 9.3: תהיינה $\{v_j\}_{j=0}^{k}$ ו- $\{e_j\}_{j=0}^{k}$ סידרת הכיוונים הצמודים וסידרת הגרדיאנטים המקומיים המתקבלים באלגוריתם ה- CG. אזי, המרחבים הבאים שוים:

$$\begin{aligned}\mathrm{span}\{e_0,\ e_1,\ e_2,\ \dots\ ,e_k\} &= \mathrm{span}\{e_0,\ \mathbf{Q}e_0,\ \mathbf{Q}^2e_0,\ \dots\ ,\mathbf{Q}^ke_0\}\\ &= \mathrm{span}\{v_0,\ v_1,\ v_2,\ \dots\ ,v_k\} = \mathbf{M}_k\end{aligned}$$

הוכחה: נוכיח זאת באינדוקציה. עבור k=0 מתקבל בודאי ש- $\mathrm{span}\{e_0\} = \mathrm{span}\{e_0\}$, וכיוון שלפי ה- CG, $v_0 = e_0$, מתקבל גם $\mathrm{span}\{v_0\} = \mathrm{span}\{e_0\}$. נניח כי הקשרים הנ"ל נכונים ל- k, ונראה את נכונותם ל- k+1. כיוון שמתקיים:

$$e_{k+1} = b - \mathbf{Q}x_{k+1} = b - \mathbf{Q}[x_k + \alpha_k v_k] = e_k - \alpha_k \mathbf{Q}v_k$$

מתקבל כי e_{k+1} מורכב מצירוף ליניארי של הוקטור $e_k \in \mathbf{M}_k$, והוקטור $\mathbf{Q}v_k \in \mathbf{M}_{k+1}$, ולכן קיים כי $e_{k+1} \in \mathbf{M}_{k+1}$, ובכך הוכחנו חצי מהטענה של המשפט. קיים גם הקשר הבא באלגוריתם ה- CG - $v_{k+1} = e_{k+1} + \beta_{k+1}v_k$, כך שברור כי $v_{k+1} \in \mathbf{M}_{k+1}$ - ובכך הוכח המשפט. ■

משמעות התוצאה הנ"ל חשובה. אם תנאי למה 9.3 מתקיימים, פירוש הדבר שה- CG לא יגיע לפתרון מדויק עד לאיטרציה האחרונה (ה- n). עבור מצב בו ישנם ערכים עצמיים עם ריבוי תיתכן הגעה לפתרון תוך מספר איטרציות קטן מ- n.

נוכל להציע את האלגוריתם הבא לפתרון בעיית מינימיזציה ריבועית. אלגוריתם זה מציע צירוף ליניארי של כל הגרדיאנטים המקומיים עד כה לבניית הפתרון החדש. בבירור ניכר הקשר לתפיסת ההאצה הפולינומיאלית שהוצגה בפרק קודם - שם הצענו כי תוצאת האיטרציה הנוכחית תהיה צירוף ליניארי של כל נקודות האיטרציות עד כה - והדברים שקולים.

$$x_1 = x_0 + \gamma_0^0 e_0$$
$$x_2 = x_0 + \gamma_1^0 e_0 + \gamma_1^1 e_1$$
$$x_3 = x_0 + \gamma_2^0 e_0 + \gamma_2^1 e_1 + \gamma_2^2 e_2$$
$$\vdots$$
$$x_{k+1} = x_0 + \gamma_k^0 e_0 + \gamma_k^1 e_1 + \gamma_k^2 e_2 + \cdots + \gamma_k^k e_k$$

אך מכיוון שעל-פי משפט 9.3 קיבלנו כי המרחב הנפרס על ידי גרדיאנטים אלו הוא מרחב קרילוב מהצורה:

$$\begin{aligned}\operatorname{span}\{e_0, e_1, e_2, \ldots, e_k\} &= \operatorname{span}\left\{e_0, \mathbf{Q}e_0, \mathbf{Q}^2 e_0, \ldots, \mathbf{Q}^k e_0\right\} \\ &= \operatorname{span}\{v_0, v_1, v_2, \ldots, v_k\} = \mathbf{M}_k\end{aligned}$$

נוכל לכתוב את המשוואת של האלגוריתם המוצע בדרך שונה:

$$\begin{aligned} x_1 &= x_0 + P_0(\mathbf{Q})e_0 & x_2 &= x_0 + P_1(\mathbf{Q})e_0 \\ x_3 &= x_0 + P_2(\mathbf{Q})e_0 \quad \cdots & x_{k+1} &= x_0 + P_k(\mathbf{Q})e_0 \end{aligned}$$

where:

$$P_k(\mathbf{Q}) = \gamma_k^0 \mathbf{I} + \gamma_k^1 \mathbf{Q} + \gamma_k^2 \mathbf{Q}^2 + \cdots + \gamma_k^k \mathbf{Q}^k$$

וכעת נוכל לשאול את השאלה הבאה - עבור האיטרציה ה- k - מהם מקדמי הפולינום אשר יתנו ערך פונקציה מינימלי? התשובה לכך ניתנה כבר במשפט 9.2 - אלגוריתם במבנה של כיוונים צמודים (כללי) מגשים באיטרציה ה- k מינימום של הפונקציה על פני תת- המרחב $\mathbf{M}_k$. לכן, אלגוריתם ה- CG (אשר הינו מקרה פרטי מיוחד של גישת ה- CD) מיישם בפועל חישוב השקול לבחירה אופטימלית של מקדמי הפולינום שתואר. נשתמש בתכונה הבאה - שחלוף במכפלה בין מטריצה לפולינומ עליה (הוכחה ע"י רישום ישיר של הביטויים במפורש):

$$P(A) \cdot A = A \cdot P(A)$$

אזי נוכל לכתוב את הפונקציה באופן הבא:

$$x_{k+1}-x^*=x_0-x^*+P_k(\mathbf{Q})e_0=$$
$$=x_0-x^*+P_k(\mathbf{Q})\mathbf{Q}(x^*-x_0)=\left[I+P_k(\mathbf{Q})\mathbf{Q}\right](x_0-x^*)$$

$$f(x)=\frac{1}{2}(x_{k+1}-x^*)^T\mathbf{Q}(x_{k+1}-x^*)=$$
$$=\frac{1}{2}(x_0-x^*)^T\mathbf{Q}\left[I+P_k(\mathbf{Q})\mathbf{Q}\right]^2(x_0-x^*)$$

כשהשתמשנו בקשר - $e_0=\mathbf{Q}x_0-b$. התקבל כי הקשר בין השגיאה באיטרציה k לשגיאת האיתחול נתון ע"י פולינום הכי כללי מסגר k+1, ולכן, ע"י שליטה במקדמיו אנו יכולים להקטין שגיאה זו באופן משמעותי (עד כדי איפוסה עבור סדר n). נרכז את תוצאה זאת במשפט הבא:

משפט 9.4: (אופטימליות אלגוריתם ה- CG). הנקודה x_{k+1} המחושבת ע"י אלגוריתם ה- CG מגשימה את המינימום של $f(x)=\frac{1}{2}(x-x^*)^T\mathbf{Q}(x-x^*)$ על פני תת-המרחב $\mathbf{M}_k$, ובדרך ייצוג שונה - מגשימה את המינימיזציה הבאה:

$$f(x_{k+1})=\min_{\{\gamma_k^j\}_{j=0}^k}\left\{\frac{1}{2}(x_0-x^*)^T\mathbf{Q}\left[I+P_k(\mathbf{Q})\mathbf{Q}\right]^2(x_0-x^*)\right\}$$

כאשר $\{\gamma_k^j\}_{j=0}^k$ הם מקדמי הפולינום $P_k(x)$.

הוכחה: למעשה ההוכחה ניתנה בעיקרה במשפט 9.2 ובלמה 9.4. ■

המשפט הקודם הביא טענה באשר לאופטימליות ה- CG. רצוננו ביותר מכך - היכולת לומר משהו אנליטי על קצב ההתכנסות. הניתוח הבא מספק תוצאות יכולות אלו.

למה 9.5: יהיו $\{\lambda_j\}_{j=1}^n$ הערכים העצמיים של המטריצה $\mathbf{Q}$. אזי, אלגוריתם ה- CG מקיים את אי-השיוויון הבא:

$$f(x_{k+1})\le\max_{\{\lambda_j\}_{j=1}^n}\left\{\left[1+\lambda_jP_k(\lambda_j)\right]^2\right\}f(x_0)$$

לכל פולינום שהוא $P_k(x)$ מסדר k.

הוכחה: נשתמש בוקטורים העצמיים של המטריצה **Q** - $\{\theta_j\}_{j=1}^{n}$ - לייצוג השגיאה - $(x_0 - x^*)$:

$$(x_0 - x^*) = \sum_{j=1}^{n} \alpha_j \theta_j$$

אזי, מתוצאה שקיבלנו במסגרת המשפט הקודם, ערך הפונקציה נתון ע"י ביטוי מהצורה:

$$f(x_{k+1}) = \min_{\{\gamma_k^j\}_{j=0}^{k}} \left\{ \frac{1}{2}(x_0 - x^*)^T \mathbf{Q}[I + P_k(\mathbf{Q})\mathbf{Q}]^2 (x_0 - x^*) \right\} =$$

$$= \frac{1}{2}\left(\sum_{j=1}^{n} \alpha_j \theta_j^T \right)^T \sum_{j=1}^{n} \lambda_j \theta_j \theta_j^T \sum_{j=1}^{n} \left[1 + \lambda_j P_k(\lambda_j)\right]^2 \theta_j \theta_j^T \left(\sum_{j=1}^{n} \alpha_j \theta_j \right) =$$

$$= \frac{1}{2}\left(\sum_{j=1}^{n} \alpha_j \lambda_j \theta_j^T \right) \cdot \left(\sum_{j=1}^{n} \alpha_j \left[1 + \lambda_j P_k(\lambda_j)\right]^2 \theta_j \right) = \frac{1}{2} \sum_{j=1}^{n} \left[1 + \lambda_j P_k(\lambda_j)\right]^2 \alpha_j^2 \lambda_j$$

נשים לב כי בביטוי הנ"ל, $\mathbf{Q} = \sum_{j=1}^{n} \lambda_j \theta_j \theta_j^T$. ברור כי עבור k=0 יתקבל:

$$f(x_0) = (x_0 - x^*)^T (x_0 - x^*) = \mathbf{Q} \frac{1}{2} \sum_{j=1}^{n} \alpha_j^2 \lambda_j$$

כך שניתן לכתוב כי:

$$f(x_{k+1}) = \frac{1}{2} \sum_{j=1}^{n} \left[1 + \lambda_j P_k(\lambda_j)\right]^2 \alpha_j^2 \lambda_j \le \max_{\lambda_j} \left\{ \left[1 + \lambda_j P_k(\lambda_j)\right]^2 \right\} f(x_0)$$

כטענת הלמה. ■

על תוצאת הלמה לעיל נוכל לבסס את הטענה הבאה:

משפט 9.6: בהנחה כי למטריצה **Q** יש n-k ערכים עצמיים בתחום [a,b], וכל יתר הערכים העצמיים גדולים מ- b, אזי, לכל וקטור איתחול x_0 מתקיים אי-השיוויון הבא:

$$f(x_{k+1}) \le \left(\frac{b-a}{b+a}\right)^2 f(x_0)$$

הערה: נשים לב לכך שגובה הפונקציה כאן מתיחס לפונקציה שבמינימום מקבלת גובה אפס.

הוכחה: ברור כי טענת למה 9.5 נכונה לכל פולינום. נניח כי $\{\lambda_j\}_{j=1}^{n}$ מסודרים בסדר יורד, ונציע את הפולינום הבא:

$$1+\lambda P_k(\lambda) = \frac{2}{a+b}\cdot\left[\frac{a+b}{2}-\lambda\right]\cdot\frac{\lambda_1-\lambda}{\lambda_1}\cdot\frac{\lambda_2-\lambda}{\lambda_2}\cdots\frac{\lambda_k-\lambda}{\lambda_k}$$

הצבת 0 בביטוי הנ"ל נותנת את הערך 1, כנדרש. האינטרוול [a,b] מגדיר את תחום ערכי הערכים העצמיים שנותרו. ברור כי פולינום זה מתאפס עבור כל הערכים העצמיים $\{\lambda_j\}_{j=1}^{k}$ אשר כולם גדולים מ- b. נציב פולינום זה בתוצאת למה 9.5 ונקבל:

$$f(x_{k+1}) \le \max_{\{\lambda_j\}_{j=1}^{n}} \left[1+\lambda_j P_k(\lambda_j)\right]^2 f(x_0) \le \max_{a\le\lambda\le b}\left\{\left[\frac{2}{a+b}\right]^2\cdot\left[\frac{a+b}{2}-\lambda\right]^2\right\} f(x_0)$$

במעבר במשוואה הנ"ל השתמשנו בעובדה ש:

$$\frac{\lambda_1-\lambda}{\lambda_1}\cdot\frac{\lambda_2-\lambda}{\lambda_2}\cdots\frac{\lambda_k-\lambda}{\lambda_k} \le 1$$

מקסימיזציה של הביטוי הנ"ל נותנת:

$$f(x_{k+1}) \le \max_{a\le\lambda\le b}\left\{\left[\frac{2}{a+b}\right]^2\cdot\left[\frac{a+b}{2}-\lambda\right]^2\right\} f(x_0) = \left[\frac{b-a}{b+a}\right]^2 f(x_0)$$

כטענת המשפט. ■

לפני שנעמוד על המשמעויות השונות של חסם זה נמחיש אותו דרך דוגמה מספרית:

דוגמה 9.1: נניח כי בידינו מטריצה Q במימד 100 עם 5 ערכים עצמיים בתחום $[100-\varepsilon,100+\varepsilon]$, וכל היתר בתחום $[1-\varepsilon,1+\varepsilon]$. לפי המשפט נקבל כי (n=100, k=5):

$$f(x_6)\le\left[\frac{(1+\varepsilon)-(1-\varepsilon)}{(1+\varepsilon)+(1-\varepsilon)}\right]^2 f(x_0)=\varepsilon^2\cdot f(x_0)$$

וזאת לעומת - $f(x_6)\le\left[\frac{(100+\varepsilon)-(1-\varepsilon)}{(100+\varepsilon)+(1-\varepsilon)}\right]^{12} f(x_0)\cong 0.98^{12}\cdot f(x_0)\cong 0.79\cdot f(x_0)$

באלגוריתם ה- NSD.

עבור מטריצה Q במימד 100 עם 99 ערכים עצמיים בתחום $[100-\varepsilon,100+\varepsilon]$, ואחד אחרון השווה 1, נקבל לפי המשפט (n=100, k=90):

$$f(x_6)\le\left[\frac{(100+\varepsilon)-(1-\varepsilon)}{(100+\varepsilon)+(1-\varepsilon)}\right]^2 f(x_0)\cong 0.98^2\cdot f(x_0)$$

וזאת לעומת - $f(x_6)\le\left[\frac{(100+\varepsilon)-(1-\varepsilon)}{(100+\varepsilon)+(1-\varepsilon)}\right]^{91\cdot 2} f(x_0)\cong 0.026\cdot f(x_0)$ באלגוריתם ה- NSD.

את הפרדוקס הזה נפתור בהמשך. בשלב זה צריך לומר כי מדובר בחסם בלבד! אין אלו הביצועים המדויקים של ה- CG, וקרוב לודאי שגם עבור מקרה זה הוא יתן ביצועים טובים יותר מה- NSD.

למשפט הנ"ל מספר משמעויות מעשיות:

1. נניח כי סידרת הערכים העצמיים של **Q** נתונה בסדר יורד - $\lambda_1\ge\lambda_2\ge\lambda_3\ge\cdots\ge\lambda_n$.
באיטרציה הראשונה, שיעור הקטנת ערך הפונקציה נתון ע"י:

$$f(x_1)\le\left(\frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}\right)^2 f(x_0)$$

בצעד הבא - האיטרציה השניה נקבל באופן דומה כי:

$$f(x_2)\le\left(\frac{\lambda_2-\lambda_n}{\lambda_2+\lambda_n}\right)^2 f(x_0)$$

משמעות תוצאה זו היא שבכל איטרציה אנו מורידים בפועל מהמטריצה **Q** את הערך העצמי הגדול ביותר, וה- Condition # הנותר מכתיב את יחס ההתכנסות. יחס זה לפיכך הולך ומשתפר עם התקדמות האיטרציות.

2. באיטרציה האחרונה מתקבל כי a=b ואז מתקבלת התוצאה המדויקת.
3. על סמך תוצאה זו מקובל להציע אלגוריתם הקרוי CG חלקי. רעיון זה יוצג בסעיף הבא.
4. אם הערך העצמי הקטן ביותר של המטריצה מופיע בריבוי m, די יהיה ב- n-m+1 איטרציות קבלת התכנסות ליעד המדויק. זאת כיוון שבמקרה זה נגדיר את האינטרוול [a,b] להיות [a,a] ואז המכפלה בחסם מתאפסת. ראינו כבר תוצאה זו מקודם ובאופן רחב יותר, בהקשר של פולינומים מינימליים.

9.4 אלגוריתם ה- CG החלקי

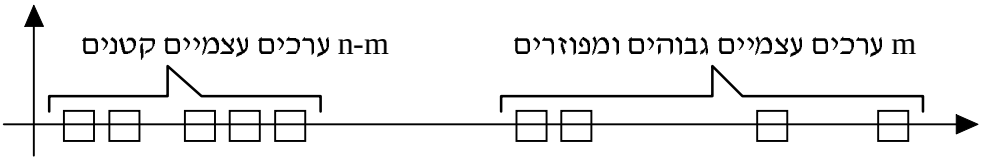
בעיות מינימיזציה ריבועיות רבות מופיעות עם מבנה ערכים עצמיים ידוע, אותו ניתן לנצל לצרכינו לקבלת התכנסות מהירה של ה- CG. במקרה בו ישנו ריכוז של ערכים עצמיים נמוכים רבים בתחום [a,b], ומיעוט m<<n של ערכים עצמיים גבוהים ניתן להציע את הרעיון הבא אשר מתבסס על תוצאת משפט 9.6. במקום לבצע n איטרציות CG, נבצע m איטרציות CG, ונאתחל מחדש. איתחול מחדש פירושו שכל היסטוריה נשכחת, והנקודה בה אנו נמצאים מהווה איתחול לאלגוריתם ה- CG. עפ"י משפט 9.6. מתקבל כי אלגוריתם זה ישיג ב- mL איטרציות את היחס הבא:

$$f(x_{mL+1}) \le \left(\frac{b-a}{b+a}\right)^2 f(x_{mL-L}) \le \cdots \le \left(\frac{b-a}{b+a}\right)^{2L} f(x_0)$$

תיאור פיזור הערכים העצמיים אשר יכול לשמש אותנו לזירוז מתואר בציור 9.3. אם אמנם b/a הינו יחס נמוך יחסית (תמיד יחס זה גדול מ- 1) בהשוואה ליחס $\lambda_{max}/\lambda_{min}$, תתקבל התכנסות מהירה הרבה יותר מאשר ב- NSD , כיוון שב- NSD מתקבל:

$$f(x_{mL+1}) \le \left(\frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}\right)^{2mL} f(x_0) = \left(\left[\frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}\right]^m\right)^{2L} f(x_0)$$

דוגמה 9.2: נניח כי בידנו בעיה ריבועית בה $\lambda_{max}/\lambda_{min}=1000$, ועבור לקיחת ארבעת הערכים הגבוהים ביותר מתקבל m=4 ו- b/a=10. במקרה כזה יתקבל כי האלגוריתמים מתכנסים אקספוננציאלית לפי:



ציור 9.3 - פיזור ערכים עצמיים אשר נותן אפשרות לשיפור קצב ההתכנסות ב- CG החלקי

$$\text{PCG:}\quad f\left(x_{mL+1}\right)\le\left(\frac{9}{11}\right)^{2L}f\left(x_0\right)=0.67^{L}f\left(x_0\right)$$

$$\text{NSD:}\quad f\left(x_{mL+1}\right)\le\left(\left[\frac{999}{1001}\right]^{4}\right)^{2L}f\left(x_0\right)=0.992^{L}f\left(x_0\right)$$

והיתרון ל- PCG ניכר.

דוגמה 9.3: נניח כי נתון לנו אות Y אשר עבר מערכת ליניארית כלשהי ידועה H וזוהם ברעש N ורצוננו לשחזרו. מודל הבעיה הוא - $Y = HX + N$. נציע את המשערך LS הבא:

$$\hat{X}=\arg\min_{X}\left\{[Y-HX]^{T}[Y-HX]\right\}=\arg\min_{X}\left\{X^{T}H^{T}HX-2X^{T}H^{T}Y+Y^{T}Y\right\}$$

כלומר, התקבלה בעיה ריבועית בה המטריצה H^TH מתפקדת כ- **Q**, ו- H^TY כ- b. בעיה מרכזית בבעיות מסוג זה היא שהמטריצה H^TH סינגולרית - שפירושו כי ישנם אינסוף פתרונות אפשריים (שקול לאמירה כי יש משטח של מינימום ולא נקודה יחידה). כדי למקד את השחזור בפתרון יחיד מקובל להוסיף מידע נוסף על היעד המבוקש. לדוגמה, נניח כי ידוע בנוסף כי ממוצעו של האות המבוקש הוא אפס. אזי נבקש פתרון לבעיה האלטרנטיבית הבאה:

$$\hat{X}=\arg\min_{X}\left\{[Y-HX]^{T}[Y-HX]+\beta\cdot\left(1^{T}X\right)^{2}\right\}$$

הוקטור 1 הוא וקטור עמודה של '1'. קל לראות כי כעת המטריצה **Q** של הבעיה הריבועית היא המטריצה - $\mathbf{Q}=H^TH+\beta\cdot\left[1\cdot1^T\right]$. מטריצה זו מקבלת בכל איבר ב- H^TH תוספת של הערך β. עבור ערך גבוה של β (יחסית לערכים העצמיים של H^TH) מתקבל כי ערך עצמי אחד של המטריצה Q יהיה גבוה - בסדר גודל של β והיתר יהיו נמוכים. (טענה זו מבוססת על משפט ה- Interlock של

ערכים עצמיים שיוזכר בפרק הבא). מבנה כזה יוצר בעיה קשה ל- NSD כיוון שה- # Condition של **Q** נפגע מאוד. לעומת זאת, הפעלת אלגוריתם ה- PCG עם צמדים של איטרציות CG ואיתחול יצור התכנסות מהירה.

9.5 אלגוריתם ה- CG לבעיות שאינן ריבועיות

בהינתן פונקציה לא ריבועית f(x) אותה רצוננו להביא למינימום, ישנן שתי גישות כלליות המאפשרות יישומו של אלגוריתם ה- CG. הדרך הראשונה היא הינה קירוב הפונקציה באופן מקומי כפונקציה ריבועית ע"י חישוב הגרדיאנט וחישוב ההסיין בכל איטרציה ושימוש בהם. ברור כי עבור בעיה ריבועית נקבל התלכדות עם טיפול ישיר כבעיה ריבועית. אבל, עבור בעיות לא ריבועיות אין כל סיבה לקבל את הפתרון הנכון תוך n איטרציות, ולכן מומלץ לבצע איתחול מחודש כל n איטרציות ע"י איטרציית SD פשוטה. תיאור של גישה זו מתואר בציור 9.4.

יתרונו וחסרונו של אלגוריתם זה בכך שבדומה לאלגוריתם ניוטון - לא נדרש כאן יישומו של אלגוריתם חיפוש מינימום על ישר, ובמקום זאת נדרש לחשב את ההסיין בכל נקודה. בבעיות בהן חישוב ההסיין קל מוטב לנקוט בגישה זו, בעוד שבבעיות במימדים שאינם קטנים בהן חישוב ההסיין הינו בלתי מעשי, מוטב לנקוט בגישה אחרת. האלגוריתם אשר תואר כאן אינו בעל תכונת התכנסות גלובלית (לא נוכיח זאת כאן) -וזהו כמובן חסרונו המרכזי.

גישה אלטרנטיבית ליישום ה- CG לבעיה שאינה ריבועית היא כאמור יישום שיטת חיפוש על ישר, ובכך להימנע מהצורך לחשב את ההסיין. בשלב 7 של האלגוריתם אשר תואר בציור 9.4 אנו רואים כי ההסיין משמש רק לקביעת גודל צעד העדכון, וביצוע חיפוש מינימום על ישר יכול לספק את α_k ללא שימוש בהסיין. האלגוריתם אשר מיישם את כל צעדי ציור 9.4 ומחליף את חישוב α_k בחיפוש על ישר נקרא אלגוריתם Fletcher-Reeves.

חסרון מרכזי של אלגוריתם Fletcher-Reeves טמון בדרך בה מחושב הפרמטר β_k. בבעיות שאינן ריבועיות מתפתחים כיוונים שאינם יוצרים משפחה **Q** אורתונורמלית (בעיקר מכיוון שהמטריצה **Q** משתנה מאיטרציה לאיטרציה), וכתוצאה מכך נוצרים מצבים בהם האלגוריתם "נתקע". פתרון היוריסטי לכך נמצא ע"י השימוש בנוסחה התחליפית הבאה:

$$\beta_k = \frac{e_k^T[e_k - e_{k-1}]}{e_{k-1}^T e_{k-1}}$$

בשל היות הגרדיאנטים העוקבים ניצבים עבור הבעיה הריבועית, נוסחה זו נכונה כמו המקורית עבור בעיות ריבועיות. נסיונות מלמדים כי בשימוש בנוסחה זו מתקבלים ביצועים טובים יותר, ונמנעת תופעת ההיתקעות. אלגוריתם זה קרוי אלגוריתם Polak-Ribiere.

1. אתחל את האלגוריתם עם x_0 כלשהו.
2. קבע את כיוון החיפוש כ- $v_0 = e_0 = -\nabla f(x_0)$.
3. באיטרציה ה- k חשב את $e_k = -\nabla f(x_k)$.
4. בהתבסס על e_k חשב את $\beta_k = \frac{e_k^T e_k}{e_{k-1}^T e_{k-1}}$.
5. קבע את כיוון החיפוש כ- $v_k = e_k + \beta_k v_{k-1}$.
6. חשב את ההסיין המקומי $\mathbf{F}(x_k)$.
7. חשב את המקדם $\alpha_k = \frac{v_k^T e_k}{v_k^T \mathbf{F}(x_k) v_k}$.
6. עדכן את הפתרון לפי - $x_{k+1} = x_k + \alpha_k v_k$.
7. לאחר סיום האיטרציה ה- n הצב $x_0 = x_n$, קבע את אינדקס האיטרציה להיות 0 וחזור ל- 1.
8. עצור את האלגוריתם אם $\|e_k\| \le \varepsilon$.

ציור 9.4 - אלגוריתם הכיוונים הצמודים - Conjugate Directions

<u>עבור בעיה שאינה ריבועית - גישת הקירוב לבעיה ריבועית</u>

נקודה אחרונה אותה יש להזכיר בהקשר זה היא שניתן ליישם את ה- CG ל- k<n איטרציות ואז לאתחל. גישה זו המזכירה את אלגוריתם ה- PCG על-פי רוב יעילה יותר מיישום n איטרציות, בעיקר עבור בעיות ממימד גדול.

<u>9.6 סיכום</u>

בפרק זה הצגנו את תפיסת הכיוונים כצמודים, והאלגוריתמים הנובעים ממנה - ה- CD וה- CG. יחודם של אלגוריתמים אלו בכך שעבור בעיה ריבועית הם מספקים פתרון מדויק תוך n איטרציות לכל היותר. ניתוח ההתכנסות של ה- CG מלמד כי ישנו כאן אפקט של ניטרול הערך העצמי הגדול ביותר של המטריצה **Q** בבעיה הריבועית בכל איטרציה - עובדה שסיפקה בסיס

תיאורתי נוח לאלגוריתם ה- CG החלקי. ראינו גם בפרק זה את הקשר להאצה הפולינומיאלית. למעשה - משפט 9.5 מספק את התוצאה החשובה ביותר בהקשר זה - וקובע כי - אלגוריתם ה- CG מספק את התוצאה הטובה ביותר האפשרית ע"י האצה פולינומיאלית (ללא צורך להכניס את הפולינומים הרקורסיביים אשר הוצגו בפרק קודם).

ישנה בעיית הבנה המתעוררת בהקשר למשפחת אלגוריתמי ה- CD, ומתיחסת לעובדת היותם של אלגוריתמי ה- CD אופטימליים. אופטימליות זו מתבטאת בכך שמעל תתי המרחבים המצטברים של הכיוונים ה- **Q**-צמודים, לא יתכן פתרון טוב יותר מזה של ה- CD. כמקרה פרטי, אלגוריתם ה- CG אופטימלי אף הוא, וזאת מעל תתי המרחבים שנוצרו מהגרדיאנטים לאחר **Q**-אורתוגונליזציה. בהינתן שני אלגוריתמי CD, לא ניתן לקבוע כי האחד טוב מהשני, כי שניהם אופטימליים - כל אחד מול וקטוריו. כשלוקחים את אלגוריתם ה- PCG מול ה- CG, נדמה כי ניתן רק להפסיד כיוון שה- CG אופטימלי וה- PCG איבד תכונה זו. כמובן, אין זה כך כיוון שה- PCG מגדיר סידרת כיוונים אחרת לגמרי אשר לא מאפשרת לתת השוואה בינו ל- CG. כל שנוכל לומר הוא שבמונחים של החסם על ה- CG, ל- PCG יכול להיות יתרון בשל מבנה הערכים העצמיים.

פרק 10 - השיטות הניוטוניות למחצה

10.1 כללי

המוטיבציה לעסוק ב- CG באה מההבנה כי בידנו יש שתי קצוות של אלגוריתמים - ה- SD (או ה- NSD) אשר מהווה אלגוריתם פשוט מבחינה חישובית אך עם ביצועים ירודים, ולעומתו - אלגוריתם ניוטון אשר הינו אלגוריתם מורכב חישובית אבל בעל ביצועי התכנסות מעולים. חיפשנו דרך אמצע, ואמנם - ה- CG מהווה דרך אמצע שכזו - ביצועי התכנסותו טובים יותר מה- SD, ומורכבותו החישובית אינה גבוהה בהרבה מזו של ה- NSD.

בפרק זה נמשיך באותו קו של חיפוש אלגוריתמי אמצע הדרך. נקודת המוצא הפעם תהיה אלגוריתם ניוטון, ותחת ההנחה שעלינו לאתר דרכים לעקוף את חישוב ההסיין נציע דרכים מקורבות שונות - אלגוריתמים אלו ייקראו לפיכך שיטות ניוטוניות למחצה. התפיסה המובילה באוסף גישות אלו היא שכתחליף לחישוב ההסיין באופן מפורש, נבנית מטריצה המתכנסת להסיין מאיטרציה לאיטרציה (תחת ההנחה שמטריצת ההסיין קבועה) בדרך חישובית זולה.

כמקודם, כל הפיתוחים ייעשו עבור בעיות ריבועיות מסדר n. הכללת האלגוריתמים לבעיות שאינן ריבועיות ישירה, ותחת ההנחה כי הפונקציה אותה רצוננו להביא למינימום ניתנת לקירוב כריבועית באופן מקומי - הרי שתכונות ההתכנסות דומות.

10.2 בניית ההסיין ע"י עדכונים מדרגה 1

נניח כי מטריצת ההסיין קבועה באיזור בו אנו פועלים - נסמנה ב- **F**. נניח כי בידינו שתי נקודות x_1, x_2 סמוכות וכן ערך הגרדיאנט של הפונצקיה f(x) בשתי נקודות אלה - e_1, e_2 (לפי סימונים מפרקים קודמים - $e_{1,2} = b - \mathbf{F}x_{1,2} = -\nabla f\left(x_{1,2}\right)$). נוכל להתבסס על קירובים לטור טיילור מהצורה:

$$f(x_2) - f(x_1) = -e_1^T[x_2 - x_1] + \frac{1}{2}[x_2 - x_1]^T \mathbf{F}[x_2 - x_1] + o\left\{\|x_2 - x_1\|^2\right\}$$
$$f(x_1) - f(x_2) = -e_2^T[x_1 - x_2] + \frac{1}{2}[x_1 - x_2]^T \mathbf{F}[x_1 - x_2] + o\left\{\|x_2 - x_1\|^2\right\}$$

וסיכום שתי משוואות אלה מניב את הביטוי:

$$-[e_2 - e_1] \cong \mathbf{F}[x_2 - x_1]$$

את הביטוי הנ"ל גם קל לקבל ע"י שימוש בקשרים $e_{1,2} = b - \mathbf{F}x_{1,2}$ ישירות - החסרתם זה מזה, ואז זה נכון במדויק לבעיה הריבועית.

אם בידנו n+1 נקודות וגרדיאנטי הפונקציה בהם - $\{x_k, e_k\}_{k=0}^{n}$ - נוכל לכתוב n משוואת כמו זו שנכתבה ולקבל:

$$-[e_k - e_{k-1}] \cong \mathbf{F}[x_k - x_{k-1}] \qquad k = 1,2,3,\ldots\ldots,n$$

ואם נגדיר את המטריצות X ו- E כך שהוקטורים $\{[x_k - x_{k-1}],[e_{k-1} - e_k]\}_{k=0}^{n}$ הן עמודות בהן, נקבל את המשוואה המטריצית הבאה - $EX^{-1} = \mathbf{F}$. כלומר, עקרונית ניתן מאוסף הוקטורים $\{x_k, e_k\}_{k=0}^{n}$ לחשב את מטריצת ההסיין **F** -כל שצריך לעשות הוא לצבור את n+1 הצמדים הללו ולחשב את **F** לפי השוואה המטריצית שקיבלנו. עם זאת, בהתיחס לאוסף הצמדים $\{x_k, e_k\}_{k=0}^{n}$ כתוצאות ביניים של אלגוריתם איטרטיבי כלשהו, רצוננו בגישה איטרטיבית אשר מקדמת את חישובו של **F** מאיטרציה לאיטרציה על סמך הצמד החדש בכל פעם. זוהי בדיוק הגישה המוצעת במשפט הבא:

משפט 10.1: תהי **F** מטריצה סימטרית כלשהי שאינה סינגולרית, ויהיו נתונים צמדי הוקטורים $\{p_j, q_j\}_{j=0}^{n-1}$, אשר מקיימים את הקשר - $q_j = \mathbf{F}p_j$ לכל j בתחום 0 עד n-1. אזי, לכל איתחול $\mathbf{H}_0$ סימטרי תקיים המשוואה האיטרטיבית הבאה:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{[p_k - \mathbf{H}_k q_k][p_k - \mathbf{H}_k q_k]^T}{q_k^T[p_k - \mathbf{H}_k q_k]}$$

את המשוואה $p_j = \mathbf{H}_{k+1} q_j$ לכל j בין 0 ל- k. אם בנוסף מתקיים כי סדרות הוקטורים $\{p_j, q_j\}_{j=0}^{n-1}$ בלתי תלויים ליניארית (כל סידרה בנפרד), יתקבל כי - $\mathbf{H}_n = \mathbf{F}^{-1}$, כלומר לאחר n איטרציות יתקבל היפוך המטריצה **F** במדויק.

הוכחה: ההוכחה תיעשה באינדוקציה. עבור $\mathbf{H}_0$ זה בודאי נכון כיוון שאין כל דרישה על מטריצה זו. נניח כי הדבר נכון ל- k, כלומר -נניח כי מתקיים כי- $p_j = \mathbf{H}_k q_j \quad \forall 0 \le j \le k-1$. עבור k+1 נפריד את ההוכחה לטיפול ב- j=k וביתר. עבור $0 \le j \le k-1$, שימוש במשוואת האיטרציה נותנת:

$$\forall 0 \le j \le k-1 \qquad \mathbf{H}_{k+1}q_j = \mathbf{H}_k q_j + \frac{\left[p_k - \mathbf{H}_k q_k\right]}{q_k^T\left[p_k - \mathbf{H}_k q_k\right]}\left[p_k^T q_j - q_k^T \mathbf{H}_k q_j\right]$$

נשים לב שבמשוואה הנ"ל מסתתרת ההנחה שהמטריצה $\mathbf{H}_0$ סימטרית כיוון ש- $\mathbf{H}_k$ צריך היה להופיע בשחלוף - משמע הנחנו ש- $\mathbf{H}_k$ סימטרי, ומתוך דרך בנייתו חייב להתקיים כי $\mathbf{H}_0$ סימטרי.
לפי הנחת האינדוקציה מתקבל כי $\mathbf{H}_k q_j = p_j$ ולכן:

$$\forall 0 \le j \le k-1 \qquad \mathbf{H}_{k+1}q_j = p_j + \frac{\left[p_k - \mathbf{H}_k q_k\right]}{q_k^T\left[p_k - \mathbf{H}_k q_k\right]}\left[p_k^T q_j - q_k^T p_j\right]$$

הגורם השני בביטוי הנ"ל נופל כיוון ש:

$$p_k^T q_j - q_k^T p_j \underset{\substack{\uparrow \\ q_j = \mathbf{F}p_j}}{=} p_k^T \mathbf{F} p_j - q_k^T p_j = \left[p_k^T \mathbf{F} - q_k^T\right] p_j \underset{\substack{\uparrow \\ \mathbf{F} = \mathbf{F}^T}}{=} \left[\mathbf{F}p_k - q_k\right]^T p_j = 0$$

ולכן - $\forall 0 \le j \le k-1 \qquad \mathbf{H}_{k+1}q_j = p_j$.

עבור j=k מתקבל:

$$\mathbf{H}_{k+1}q_k = \mathbf{H}_k q_k + \frac{\left[p_k - \mathbf{H}_k q_k\right]\left[p_k - \mathbf{H}_k q_k\right]^T}{q_k^T\left[p_k - \mathbf{H}_k q_k\right]} q_k = \mathbf{H}_k q_k + p_k - \mathbf{H}_k q_k = p_k$$

כך שהטענה הוכחה גם עבור j=k. לכן, בפרט מתקיים כי - $\mathbf{H}_n q_j = p_j \qquad \forall 0 \le j \le n-1$, ובסידור אוסף קשרים אלו למבנה מטריצי יתקבל:

$$\mathbf{H}_n \begin{bmatrix} q_0 & \cdots & q_{n-1} \end{bmatrix} = \begin{bmatrix} p_0 & \cdots & p_{n-1} \end{bmatrix} \Rightarrow \mathbf{H}_n Q = P$$

$$\Rightarrow \quad \mathbf{H}_n = PQ^{-1} = \left(QP^{-1}\right)^{-1} = \mathbf{F}^{-1}$$

אך קשר זה נכון רק אם סדרות הוקטורים בלתי תלויות. ■

על סמך משפט 10.1 נוכל להציע אלגוריתם דמוי ניוטון אשר במקום לחשב את היפוך ההסיין הוא בונה אותו במהלך האיטרציות. ישנו כאן יתרון משמעותי - המטריצה המתפתחת היא היפוך ההסיין - עובדה אשר מאפשרת יישום משוואת העדכון לפתרון באופן ישיר כיוון שמשוואה זו היא הבאה:

$$x_{k+1} = x_k + \alpha_k \mathbf{H}_{k+1} e_k$$

ערכו של הפרמטר α_k צריך להיקבע כזה שיתן מינימום על הישר $x = x_k + \alpha_k \mathbf{H}_{k+1} e_k$.

ערך זה נתון ע"י:

$$\frac{\partial f(x_{k+1})}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k}\left(x_k + \alpha_k \mathbf{H}_{k+1} e_k - x^*\right)^T \mathbf{F}\left(x_k + \alpha_k \mathbf{H}_{k+1} e_k - x^*\right) = 0$$

$$\Rightarrow \qquad e_k^T \mathbf{H}_{k+1} \mathbf{F}\left(x_k + \alpha_k \mathbf{H}_{k+1} e_k - x^*\right) = 0$$

$$\Rightarrow \qquad \alpha_k = \frac{e_k^T \mathbf{H}_{k+1} \mathbf{F}\left(x_k - x^*\right)}{e_k^T \mathbf{H}_{k+1} \mathbf{F} \mathbf{H}_{k+1} e_k} = \frac{e_k^T \mathbf{H}_{k+1} e_k}{e_k^T \mathbf{H}_{k+1} \mathbf{F} \mathbf{H}_{k+1} e_k}$$

אך נוסחה זו חסרת תכלית אם $\mathbf{F}$ אינו ידוע, ולכן, יידרש תהליך של חיפוש על ישר. תיאור האלגוריתם הכולל עבור בעיה ריבועית מובא בציור 10.1. על מנת להמירו לטיפול בבעיות שאינן ריבועיות כל שעלינו לעשות הוא חישוב גרדיאנט בצורה נאותה בכל איטרציה, וביצוע חיפוש מינימום על ישר בשלב 5 של האלגוריתם.

לאלגוריתם זה שתי חולשות מרכזיות הנובעות מהמשוואה האיטרטיבית לחישוב קירוב ההסיין.

1. ייתכנו מצבים בהם יתקבל כי $q_k^T[p_k - \mathbf{H}_k q_k] < 0$ (הדבר תלוי בבחירת מטריצת האיתחול). פירוש הדבר שבמהלך הדרך יתקבל כי המטריצה $\mathbf{H}_k$ אינה חיובית מוגדרת, ולכן כיוון העדכון לא יהיה כיוון ירידה.
2. הגורם $q_k^T[p_k - \mathbf{H}_k q_k]$ עלול להיות קרוב לאפס - עובדה אשר תגרום לאי-יציבות נומרית. חסרונות אלו הביאו לפיתוח אלטרנטיבות אשר נכיר בסעיפים הבאים.

1. בחר נקודת איתחול שרירותית x_0.
2. בחר מטריצת איתחול שרירותית חיובית מוגדרת $\mathbf{H}_0$.
3. באיטרציה ה- k חשב את הגרדיאנט המקומי $e_k = \nabla f(x_k)$.
4. חשב את גודל הצעד α_k לפי חיפוש מינימום על הישר הנתון ע"י $f(x+\alpha_k \mathbf{H}_k e_k)$.
5. בצע את העדכון לפי $x_{k+1} = x_k + \alpha_k \mathbf{H}_k e_k$.
6. בצע את חישוב העדכון למטריצת הכיוון לפי:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{[p_k - \mathbf{H}_k q_k][p_k - \mathbf{H}_k q_k]^T}{q_k^T[p_k - \mathbf{H}_k q_k]}$$

כאשר - $p_k = \alpha_k \mathbf{H}_k e_k$, $q_k = e_k - e_{k+1}$.

7. עצור את האלגוריתם כאשר $\|e_k\| = \|\nabla f(x_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 10.1 - סיכום אלגוריתם ניוטונית למחצה מבוססת עדכון מדרגה 1</u>

10.3 אלגוריתם ה- DFP

אלגוריתם אלטרנטיבי אשר הוצע ע"י Davidon, ולאחר מכן שוכלל ע"י Fletcher ו- Powel הוא האלגוריתם אשר יידון בסעיף זה. כפי שנראה, בנוסף להתגברותו על חסרונות התפיסה הקודמת, מציע אלגוריתם זה יתרון חדש לא צפוי עבור הבעיה הריבועית - יצירת כיווני התקדמות כשל ה- CG. אם בסעיף קודם דיברנו על עדכון מבוסס דרגה 1, כעת השיטה תתמקד בעדכון מבוסס דרגה 2, כפי שנראה מיד. נתאר תחילה את האלגוריתם, ואז נדון בתכונותיו. ציור 10.2 מספק תיאור של אלגוריתם ה- DFP.

נתחיל את הטיפול האנליטי בכך שנראה כי מתקבל כי תמיד המטריצה המתפתחת $\mathbf{H}_k$ תהיה חיובית מוגדרת. זהו כמובן יתרון משמעותי בהשוואה לאלגוריתם הקודם, כיוון שראינו כי עלול להיווצר מצב בו המטריצה המתקבלת תהיה לא חיובית מוגדרת.

משפט 10.2: משוואת הרקורסיה באלגוריתם ה- DFP, הנתונה ע"י:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{p_k p_k^T}{q_k^T p_k} - \frac{\mathbf{H}_k q_k q_k^T \mathbf{H}_k}{q_k^T \mathbf{H}_k q_k}$$

מבטיחה כי המטריצה המתקבלת הינה חיובית מוגדרת, אם מטריצת האיתחול חיובית מוגדרת.

הוכחה: על מנת שמטריצה זו תהיה חיובית מוגדרת נדרוש שלכל וקטור x יתקבל כי:

$$x^T \mathbf{H}_{k+1} x = x^T \mathbf{H}_k x + x^T \frac{p_k p_k^T}{q_k^T p_k} x - x^T \frac{\mathbf{H}_k q_k q_k^T \mathbf{H}_k}{q_k^T \mathbf{H}_k q_k} x > 0$$

1. בחר נקודת איתחול שרירותית x_0.
2. בחר מטריצת איתחול שרירותית חיובית מוגדרת $\mathbf{H}_0$.
3. באיטרציה ה- k חשב את הגרדיאנט המקומי $e_k = \nabla f(x_k)$.
4. חשב את גודל הצעד α_k לפי חיפוש מינימום על הישר הנתון ע"י $f(x + \alpha_k \mathbf{H}_k e_k)$.
5. בצע את העדכון לפי $x_{k+1} = x_k + \alpha_k \mathbf{H}_k e_k$.
6. בצע את חישוב העדכון למטריצת הכיוון לפי:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{p_k p_k^T}{q_k^T p_k} - \frac{\mathbf{H}_k q_k q_k^T \mathbf{H}_k}{q_k^T \mathbf{H}_k q_k}$$

כאשר - $p_k = \alpha_k \mathbf{H}_k e_k$, $q_k = e_k - e_{k+1}$.

7. עצור את האלגוריתם כאשר $\|e_k\| = \|\nabla f(x_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 10.2 - סיכום אלגוריתם DFP</u>

נראה זאת באינדוקציה. ברור כי עבור k=0 זה מתקיים כי $\mathbf{H}_0$ חיובית מוגדרת לפי הנחת המשפט. נניח כי $\mathbf{H}_k$ חיובית מוגדרת. אזי קיימת מטריצת שורש סימטרית למטריצה זו המקיימת $\mathbf{H}_k = \mathbf{H}_k^{1/2} \cdot \mathbf{H}_k^{1/2}$. נגדיר את שני הוקטורים הבאים:

$$a = \mathbf{H}_k^{1/2} x \qquad b = \mathbf{H}_k^{1/2} q_k$$

שימוש בהגדרות אלו במשוואת הרקורסיה אשר הוכפלה מימין ומשמאל בוקטור x ייתן:

$$x^T \mathbf{H}_{k+1} x = a^T a + \frac{\left[p_k^T x\right]^2}{q_k^T p_k} - a^T \frac{bb^T}{b^T b} a =$$

$$= a^T a - \frac{\left[a^T b\right]^2}{b^T b} + \frac{\left[p_k^T x\right]^2}{q_k^T p_k} = \frac{\left[a^T a\right]\left[b^T b\right] - \left[a^T b\right]^2}{b^T b} + \frac{\left[p_k^T x\right]^2}{q_k^T p_k}$$

נזכור כי מאלגוריתם ה- DFP - $p_k = \alpha_k \mathbf{H}_k e_k$, $q_k = e_k - e_{k+1}$, ולכן:

$$q_k^T p_k = \alpha_k \left[e_k - e_{k+1}\right]^T \mathbf{H}_k e_k = \alpha_k e_k^T \mathbf{H}_k e_k$$

זאת כיוון ש - $e_{k+1}^T \mathbf{H}_k e_k = 0$ - הגרדיאנט בנקודה החדשה אנך לכיוון ההתקדמות (משפט הניצבות). לכן מתקבל:

$$x^T \mathbf{H}_{k+1} x = \frac{\left[a^T a\right]\left[b^T b\right] - \left[a^T b\right]^2}{b^T b} + \frac{\left[p_k^T x\right]^2}{\alpha_k e_k^T \mathbf{H}_k e_k} \geq 0$$

החלק הראשון גדל מאפס בשל אי-שיויון קושי שוורץ ($\|a\| \cdot \|b\| \geq \langle a, b \rangle$), והחלק השני בשל היותו גורם ריבועי. נותר לנו להראות כי ביטוי זה לא מתאפס לעולם. לצורך זה נבחין כי איפוס הגורם הראשון במשוואה מתקבל כאשר a ו- b פרופורציוניים, דהיינו - $x = \beta q_k$, ובמקרה כזה מתקבל כי:

$$x^T \mathbf{H}_{k+1} x = \frac{\beta^2 \left[p_k^T q_k\right]^2}{\alpha_k e_k^T \mathbf{H}_k e_k} = \frac{\beta^2 \left[\alpha_k e_k^T \mathbf{H}_k e_k\right]^2}{\alpha_k e_k^T \mathbf{H}_k e_k} = \beta^2 \alpha_k e_k^T \mathbf{H}_k e_k > 0$$

כך שתמיד מתקיים כי $x^T \mathbf{H}_{k+1} x > 0$, והמשפט הוכח. ■

וכעת נפנה לטענה החזקה של אלגוריתם - DFP - יצירת כיווני חיפוש **F**-צמודים, וקיום המשוואות הנדרשות - $\mathbf{H}_k q_j = p_j \;\; \text{for} \;\; j = 0,1,\cdots,k-1$.

משפט 10.3: עבור בעיית מינימיזציה ריבועית עם מטריצת הסיין $\mathbf{F}$ וצמדי הוקטורים $\left\{p_j, q_j\right\}_{j=0}^{n-1}$, אלגוריתם ה- DFP באיטרציה ה- k מקיים את התכונות הבאות:

1. הכיוונים $\left\{p_j\right\}_{j=0}^{k}$ הינם $\mathbf{F}$-צמודים, דהיינו- $\forall 0 \le i < j \le k \quad p_i^T \mathbf{F} p_j = 0$
2. לכל איתחול $\mathbf{H}_0 > 0$ תקיים המשוואה האיטרטיבית הבאה:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{p_k p_k^T}{q_k^T p_k} - \frac{\mathbf{H}_k q_k q_k^T \mathbf{H}_k}{q_k^T \mathbf{H}_k q_k}$$

את המשוואה $p_j = \mathbf{H}_{k+1} q_j$ לכל j בין 0 ל- k.

3. אם מתקיים כי סדרות הוקטורים $\left\{p_j, q_j\right\}_{j=0}^{n-1}$ בלתי תלויים ליניארית (כל סידרה בנפרד), יתקבל כי - $\mathbf{H}_n = \mathbf{F}^{-1}$, כלומר לאחר n איטרציות יתקבל היפוך המטריצה $\mathbf{F}$ במדויק

הוכחה: לפני שניגש להוכחה נבחין כי מתקיים:

$$q_k = e_k - e_{k+1} = \left[b - \mathbf{F}x_k\right] - \left[b - \mathbf{F}x_{k+1}\right] = \mathbf{F}\left[x_{k+1} - x_k\right] = \mathbf{F}p_k$$

וכן מתקיים:

$$\mathbf{H}_{k+1} q_k = \mathbf{H}_k q_k + \frac{p_k p_k^T}{q_k^T p_k} q_k - \frac{\mathbf{H}_k q_k q_k^T \mathbf{H}_k}{q_k^T \mathbf{H}_k q_k} q_k = \left[1 - \frac{q_k^T \mathbf{H}_k q_k}{q_k^T \mathbf{H}_k q_k}\right] \mathbf{H}_k q_k + p_k = p_k$$

נוכיח משפט זה באינדוקציה. ברור כי קשרים אלו מתקיימים עבור k=0, כיוון שאז נדרש שהכיוון p_0 יהיה $\mathbf{F}$-צמוד למרחב ריק, וזה בודאי מתקיים. בהתיחס לדרישה 2 נדרש הקשר - $p_0 = \mathbf{H}_1 q_0$ וזה מתקבל כמקרה פרטי של המשוואה שהתקבלה לעיל.

כעת, נניח כי התכונות הנ"ל נכונות ל- k-1, דהיינו - $\forall 0 \le i < j \le k-1 \quad p_i^T \mathbf{F} p_j = 0$, וכן $\forall 0 \le j \le k-1 \quad p_j = \mathbf{H}_k q_j$. מתקיים הקשר הבא לכל $0 \le i \le k-2$:

$$\begin{aligned} e_k &= e_{k-1} - q_{k-1} = e_k - \mathbf{F}p_{k-1} = \\ &= e_{k-2} - e_{k-2} + e_{k-1} - \mathbf{F}p_{k-1} = \cdots = e_{i+1} - \mathbf{F}\left[p_{k-1} + \cdots + p_{i+1}\right] \end{aligned}$$

ע"י הכפלתו משמאל בוקטור p_i ($0 \le i \le k-2$), ושימוש בהנחת האינדוקציה יתקבל:

$$p_i^T e_k = p_i^T e_{i+1} - p_i^T \mathbf{F}\left[p_{k-1} + \cdots + p_{i+1}\right] = p_i^T e_{i+1} = 0$$

כשעשינו שימוש בעובדה ש - $p_i^T e_{i+1} = 0$ - משפט הניצבות. למעשה קשר זה נכון גם עבור k-1 כיוון ש- $p_{k-1}^T e_k = 0$ (שוב - משפט הניצבות). מהנחת האינדוקציה ידוע כי מתקיים כי - $\mathbf{H}_k q_i = \mathbf{H}_k \mathbf{F} p_i = p_i$. לכן:

$$0 = p_i^T e_k = \left[\mathbf{H}_k \mathbf{F} p_i\right]^T e_k = p_i^T \mathbf{F} \mathbf{H}_k e_k \underset{\substack{\uparrow \\ p_k = \alpha_k \mathbf{H}_k e_k}}{=} \frac{1}{\alpha_k} p_i^T \mathbf{F} \mathbf{H}_k \mathbf{H}_k^{-1} p_k = p_i^T \mathbf{F} p_k$$

כטענת המשפט. באופן זה הוכחנו את תכונת ה- $\mathbf{F}$ אורתוגונליות הנטענת לכל $0 \le i \le k-1$. שימוש במשוואת הרקורסיה של $\mathbf{H}_{k+1}$, ושימוש בהנחת האינדוקציה מניבים:

$$\mathbf{H}_{k+1} q_i = \mathbf{H}_k q_i + \frac{p_k p_k^T}{q_k^T p_k} q_i - \frac{\mathbf{H}_k q_k q_k^T \mathbf{H}_k}{q_k^T \mathbf{H}_k q_k} q_i = p_i + \left[\frac{p_k}{q_k^T p_k} - \frac{\mathbf{H}_k q_k}{q_k^T \mathbf{H}_k q_k}\right] p_k^T \mathbf{F} p_i = p_i$$

ולכן - $\mathbf{H}_{k+1} q_j = p_j \quad \forall 0 \le j \le k-1$ (עבור i=k כבר ראינו את קיום תכונה זו).

בפרט מתקיים כי $\mathbf{H}_n q_j = p_j \quad \forall 0 \le j \le n-1$, וכמקודם, פירוש הדבר הוא שעבור סדרות הוקטורים בלתי תלויות (כל אחת בנפרד) מתקבל כי:

$$\mathbf{H}_n \begin{bmatrix} q_0 & \cdots & q_{n-1} \end{bmatrix} = \begin{bmatrix} p_0 & \cdots & p_{n-1} \end{bmatrix} \Rightarrow \mathbf{H}_n Q = P$$

$$\Rightarrow \quad \mathbf{H}_n = PQ^{-1} = \left(QP^{-1}\right)^{-1} = \mathbf{F}^{-1}$$

והמשפט הוכח. ■

במשפט זה הוכחנו למעשה כי משוואת הרקורסיה המוצעת מקרבת באופן איטרטיבי את מטריצת ההסיין ומתכנסת אליה כעבור n איטרציות - תכונה רצויה אשר ראינו את קיומה גם בשיטת העדכון מדרגה 1. בנוסף, משמעות משפט זה היא שבדומה לאלגוריתם ה- CG, אלגוריתם ה- DFP יוצר כיוונים **F**-צמודים. יש לזכור כי תכונה זו באה בנוסף לכך שמתפתחת מטריצת קירוב להיפוך הסיין אשר מתכנסת אליה במדויק עבור הבעיה הריבועית. לכן צפוי כי ביצועי אלגוריתם זה מבחינת קצב התכנסות יהיו טובים לפחות כמו אלו של ה- CG. תכונה מעניינת נוספת שלא תוכח כאן היא שעבור איתחול האלגוריתם עם $\mathbf{H}_0 = I$ מתקבל אלגוריתם ה- CG במדויק.

תכונה חשובה אחרת אשר התקבלה גם כאן וגם בגירסת העדכון מדרגה 1 היא הבאה - באיטרציה ה- k מתקבל כי מתקיים הקשר:

$$\forall 0 \le i \le k-1: \quad \mathbf{H}_k \mathbf{F} p_i \underset{\mathbf{F}p_i = q_i}{\underset{\uparrow}{=}} \mathbf{H}_k q_i \underset{\mathbf{H}_k q_i = p_i}{\underset{\uparrow}{=}} p_i$$

כלומר, הוקטור $\{p_i\}_{i=0}^{k-1}$ הם וקטורים עצמיים של המטריצה $\mathbf{H}_k\mathbf{F}$, ואליהם מתיחסים ערכים עצמיים השוים זהותית ל- 1. אלגוריתמי הניוטון למחצה בונים בפועל סידרת וקטורים עצמיים אלו, ומביאים אחד אחר אחד את כל הערכים העצמיים של המכפלה $\mathbf{H}_k\mathbf{F}$ להיות 1. באיטרציה ה-n כשכולם 1 מתקבלת התכנסות מדויקת (לבעיה הריבועית). אנו נראה כי תכונה חשובה זו חושפת תולשה קשה של אלגוריתמים אלו.

10.4 אלגוריתמי Broyden

תוך שימוש בקבוצת צימדי וקטורים המקיימים - $q_i = \mathbf{F}p_i \quad 0 \le i \le k$ בנינו באופן רקורסיבי את סידרת המטריצות $\mathbf{H}_k$ המקיימות - $\mathbf{H}_{k+1}q_i = p_i \quad 0 \le i \le k$, ומתכנסות להיפוך ההסיין. ממש כשם שיכולנו לבנות משוואה רקורסיבית לעדכון $\mathbf{H}_k$, נוכל להציע בניית סידרת מטריצות $\mathbf{B}_k$ אשר יתקרבו להסיין (ולא להיפוכו), ויקיימו - $\mathbf{B}_{k+1}p_i = q_i \quad 0 \le i \le k$. קל מאוד לקבל את משוואות הרקורסיה כיוון שכל שנדרש לעשות הוא להפוך את תפקידי הוקטורים p_i, q_i. נוכל לעשות זאת עבור משוואות הרקורסיה של העדכון מדרגה 1 או 2 (ה- DFP). עבור משוואת ה- DFP מתקבל:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{p_k p_k^T}{q_k^T p_k} - \frac{\mathbf{H}_k q_k q_k^T \mathbf{H}_k}{q_k^T \mathbf{H}_k q_k} \quad \Leftrightarrow \quad \mathbf{B}_{k+1} = \mathbf{B}_k + \frac{q_k q_k^T}{p_k^T q_k} - \frac{\mathbf{B}_k p_k p_k^T \mathbf{B}_k}{p_k^T \mathbf{H}_k p_k}$$

נשאלת השאלה - מהו אם כך הרווח כאן? כל שעשינו הוא להציע נוסחה רקורסיבית של ההיפוך של המטריצה הדרושה לנו (כיוון שאנו נדרשים באלגוריתם להיפוך ההסיין). התשובה היא שאמנם, בתצורה הנ"ל עדיין אין רווח, אך ע"י מניפולציה פשוטה ניתן להמיר משוואה זו למשוואה בה מקודם היפוך הקירוב להסיין. לצורך כך נשתמש בלמה הבאה:

למה 10.1: (נוסחת שרמן-מוריסון). עבור מטריצה הפיכה כלשהי A, וצמד וקטורים כלשהם a,b מתקיימת הנוסחה הבאה:

$$\left[A+ab^T\right]^{-1}=A^{-1}-\frac{A^{-1}ab^TA^{-1}}{1+b^TA^{-1}a}$$

הוכחה: נכפול את שני אגפי המשוואה במטריצה $\left[A+ab^T\right]$ ונקבל:

$$\begin{aligned}
I&=\left[A+ab^T\right]A^{-1}-\left[A+ab^T\right]\frac{A^{-1}ab^TA^{-1}}{1+b^TA^{-1}a}=\\
&=I+ab^TA^{-1}-\frac{ab^TA^{-1}+ab^TA^{-1}ab^TA^{-1}}{1+b^TA^{-1}a}=\\
&=I+ab^TA^{-1}-\frac{I+ab^TA^{-1}}{1+b^TA^{-1}a}ab^TA^{-1}=\\
&=I+ab^TA^{-1}-\frac{1+\left(b^TA^{-1}a\right)}{1+b^TA^{-1}a}ab^TA^{-1}=I
\end{aligned}$$

כשעשינו שימוש בעובדה ש- $b^TA^{-1}a$ הוא סקלאר ולכן במכפלה הבאה הוא יוצא החוצה - $ab^TA^{-1}ab^TA^{-1}=\left(b^TA^{-1}a\right)\cdot ab^TA^{-1}$. בכך הוכחנו את הלמה. ■

אם ניקח את המשוואה אשר קיבלנו עבור $\mathbf{B}_k$, ונשתמש בלמה הנ"ל פעמיים (פעם כנגד כל גורם עדכון מדרגה 1), ונבצע מספר ניכר ומייגע של פעולות אלגבריות נקבל את המשוואה הבאה:

$$\mathbf{H}_{k+1}^{new}=\mathbf{B}_{k+1}^{-1}=\mathbf{H}_k^{new}+\left(\frac{1+q_k^T\mathbf{H}_k^{new}q_k}{q_k^Tp_k}\right)\frac{p_kp_k^T}{q_k^Tp_k}-\frac{p_kq_k^T\mathbf{H}_k^{new}+\mathbf{H}_k^{new}q_kp_k^T}{q_k^Tp_k}$$

ובדומה לנוסחת הרקורסיה באלגוריתם ה- DFP, גם נוסחה זו כוללת עדכון (כצפוי) מדרגה 2. זוהי נוסחה אלטרנטיבית הקרויה על-שם החוקרים - Broyden, Fletcher, Goldfarb & Shanno ומכאן השם - אלגוריתם BFGS.

מעניין לציין (ונקבע זאת ללא הוכחה) כי כל התכונות שהוזכרו בהקשר של ה- DFP תקפות גם כאן - המטריצה המתקבלת הינה חיובית מוגדרת, המקיימת את הקשרים הנדרשים על הוקטורים $\{p_j, q_j\}_{j=0}^{n-1}$, וחשוב מכל - הכיוונים הנוצרים הינם **F**-צמודים. גם כאן האלגוריתם מביא אחד אחרי השני את כל הערכים העצמיים של המכפלה $\mathbf{H}_k^{new}\mathbf{F}$ להיות 1. סימולציות מלמדות כי שימוש בנוסחה זו מניב על-פי רוב ביצועי התכנסות טובים יותר מאלו המתקבלים בשיטת ה-DFP - אך אין לתכונה זו בינתיים הסבר אנליטי מובהק.

על-סמך שתי הגישות שהתפתחו - ה- DFP וה- BFGS, הציע Broyden משפחת שיטות ביניים בהן המטריצה המתפתחת תהיה צירוף קמור של השתיים, דהיינו:

$$0 \le \phi \le 1 \qquad \mathbf{H}_{k+1}^{B} = \phi \mathbf{H}_{k+1}^{DFP} + (1-\phi)\mathbf{H}_{k+1}^{BFGS}$$

כאשר:

$$\mathbf{H}_{k+1}^{DFP} = \mathbf{H}_k^{DFP} + \frac{p_k p_k^T}{q_k^T p_k} - \frac{\mathbf{H}_k^{DFP} q_k q_k^T \mathbf{H}_k^{DFP}}{q_k^T \mathbf{H}_k^{DFP} q_k}$$

$$\mathbf{H}_{k+1}^{BFGS} = \mathbf{H}_k^{BFGS} + \left(\frac{1 + q_k^T \mathbf{H}_k^{BFGS} q_k}{q_k^T p_k}\right)\frac{p_k p_k^T}{q_k^T p_k} - \frac{p_k q_k^T \mathbf{H}_k^{BFGS} + \mathbf{H}_k^{BFGS} q_k p_k^T}{q_k^T p_k}$$

לאחר סידרת צעדי אלגברה מייגעים מתקבל:

$$\mathbf{H}_{k+1}^{B} = \mathbf{H}_k^{DFP} + \phi v_k v_k^T$$

$$\text{where:} \quad v_k = \left[q_k^T \mathbf{H}_k^{DFP} q_k\right]^{0.5} \cdot \left[\frac{p_k}{p_k^T q_k} - \frac{\mathbf{H}_k^{DFP} q_k}{q_k^T \mathbf{H}_k^{DFP} q_k}\right]$$

ומסתבר כי גם למשפחה פרמטרית זו מתקיימות עיקר התכונות המבורכות - המטריצה המתקבלת אמנם אינה בהכרח חיובית מוגדרת, אך היא מקיימת את הקשרים הנדרשים על הוקטורים $\{p_j, q_j\}_{j=0}^{n-1}$, הכיוונים הנוצרים הינם **F**-צמודים, והערכים העצמיים של המכפלה $\mathbf{H}_k^{B}\mathbf{F}$ מובאים ל- 1, וזאת גם עבור אלגוריתם בו הפרמטר ϕ משתנה מאיטרציה לאיטרציה. מחקר ניכר יוחד לסוגיית קביעת הפרמטר ϕ באופן אדפטיבי ואופטימלי - נושא זה לא יוצג במסגרת זו.

נקודה אחרונה אשר נזכיר על משפחת אלגוריתמי Broyden (אשר כמקרים פרטיים שלו מצויות השיטות - ה- DFP וה- BFGS), היא האפשרות לבצע m<n איטרציות של השיטה המוצעת ולאתחל מחדש, ממש כשם שהצענו זאת עבור אלגוריתם ה- CG. שיטה זו קרויה אלגוריתם Broyden החלקי. ברור כי יתרונה של שיטה זו יתמצא עבור בעיות לא ריבועיות.

10.5 תכונות התכנסות של האלגוריתמים הניוטוניים למחצה

במסגרת סעיף זה נעסוק בחקר תכונות ההתכנסות הלוקלית של האלגוריתמים אשר הוצעו בפרק זה. נתחיל במשפט אשר יבסס את כל הדיון.

משפט 10.5: תהי x^* נקודת המינימום של הבעיה הריבועית - $f(x)=\frac{1}{2}(x-x^*)^T\mathbf{F}(x-x^*)$, וניישם את האלגוריתם הבא למציאתה:

$$x_{k+1}=x_k+\alpha_k\mathbf{S}_k e_k$$

כאשר: $\alpha_k=\frac{e_k^T\mathbf{S}_k e_k}{e_k^T\mathbf{S}_k\mathbf{F}\mathbf{S}_k e_k}$; $e_k=\mathbf{F}(x^*-x_k)$ (α_k מגשים חיפוש מינימום על הישר). אזי מתקיים:

$$\frac{f(x_{k+1})}{f(x_k)}\le\left(\frac{B_k-b_k}{B_k+b_k}\right)^2$$

כאשר b_k ו- B_k הם הערכים העצמיים הקטן ביותר והגדול ביותר של המטריצה $\mathbf{S}_k\mathbf{F}$.

הוכחה: ניתן לראות פיתוח מלא של משפט דומה בפרק 7 (עבור אלגוריתם ה- SD - משפט 7.4). כאן נציג את ההוכחה בתמצות. ע"י הצבה ישירה נקבל:

$$\frac{f(x_k)-f(x_{k+1})}{f(x_k)}=\frac{\left[e_k^T\mathbf{S}_k e_k\right]^2}{\left[e_k^T\mathbf{S}_k\mathbf{F}\mathbf{S}_k e_k\right]\left[e_k^T\mathbf{F}^{-1}e_k\right]}$$

וע"י שימוש בסימונים - $v_k=\mathbf{S}_k^{1/2}e_k$; $\mathbf{T}_k=\mathbf{S}_k^{1/2}\mathbf{F}\mathbf{S}_k^{1/2}$, יתקבל:

$$\frac{f(x_k)-f(x_{k+1})}{f(x_k)}=\frac{\left[v_k^T v_k\right]^2}{\left[v_k^T \mathbf{T}_k v_k\right]\left[v_k^T \mathbf{T}_k^{-1} v_k\right]}$$

כיוון שמתקיים $\mathbf{T}_k = \mathbf{S}_k^{1/2}\mathbf{F}\mathbf{S}_k^{1/2}$, קיים כי $\mathbf{S}_k^{1/2}\mathbf{T}_k\mathbf{S}_k^{-1/2} = \mathbf{S}_k\mathbf{F}$, כלומר המטריצות $\mathbf{T}_k$ ו- $\mathbf{S}_k\mathbf{F}$ דומות, ולפיכך בעלות אותם ערכים עצמיים. ע"י נוסחת קנטורוביץ אותה הכרנו כבר בפרק 7 מתקבלת טענת המשפט. ■

המשפט שהוצג מכיל בתוכו מגוון רחב של מקרים פרטיים - נדון בהם:

1. עבור $\mathbf{S}_k = I$, האלגוריתם הנדון הוא ה- NSD, ותוצאת משפט זה הוצגה עבורו כבר בפרק 7. שם ראינו כי יחס הערכים העצמיים (הגדול לקטן) של $\mathbf{F}$ מכתיב את ביצועי האלגוריתם (לפחות כחסם עליון).
2. עבור הבחירה - $\mathbf{S}_k = \mathbf{F}^{-1}$ מתקבל אלגוריתם ניוטון. הערכים העצמיים של $\mathbf{S}_k\mathbf{F}$ הם כולם 1, ועל פי משפט זה ההתכנסות מתרחשת באיטרציה אחת - גם תוצאה זו ראינו בפרק 7.
3. עבור בחירת $\mathbf{S}_k = \mathbf{H}_k \xrightarrow[k\to n]{} \mathbf{F}^{-1}$ אנו מקבלים את אוסף הגישות הניוטוניות למחצה. כפי שראינו - באיטרציה ה- k מביאים האלגוריתמים הניוטוניים למחצה k ערכים עצמיים של המטריצה $\mathbf{S}_k\mathbf{F}$ להיות 1, כיוון ש:

$$\forall 0 \le i \le k-1 \quad \mathbf{S}_k\mathbf{F}p_i = \mathbf{S}_k q_i = p_i$$

כלומר, p_i $\forall 0 \le i \le k-1$ וקטורים עצמיים של $\mathbf{S}_k\mathbf{F}$ עם ערך עצמי 1 (ויש k כאלה). השאלה היא מה קורה ליתר? אם היתר מצויים סביב הערך 1, נקבל יחס התכנסות טוב. אם לעומת זאת היתר מצויים הרבה מעל 1 - נקבל יחס גרוע, וביצועים ירודים.

כדי להבין מה קורה ליתר הערכים העצמיים - נציג את הלמה הבאה ללא הוכחה:

למה 10.2: (משפט ה- Interlock). תהי A מטריצה סימטרית בגודל $n\times n$. נניח כי ערכיה העצמיים מסודרים ונתונים כאוסף - $\lambda_1 \le \lambda_2 \le \lambda_3 \le \cdots \le \lambda_n$. יהיה a וקטור כלשהו ממימד n. נסמן את הערכים העצמיים של המטריצה $A + aa^T$ ב- $\mu_1 \le \mu_2 \le \mu_3 \le \cdots \le \mu_n$. אזי מתקיים כי:

$$\lambda_1 \le \mu_1 \le \lambda_2 \le \mu_2 \le \lambda_3 \le \mu_3 \le \cdots \le \mu_{n-1} \le \lambda_n \le \mu_n$$

הוכחה: ראה ספרי אלגברה ליניארית. ■

על סמך תוצאה זו ניכר כי עבור מטריצת הסיין $\mathbf{F}$ בעלת ערכים עצמיים גבוהים מאוד (מ-1), עדכון המטריצה $\mathbf{H}_k$ בגורם בעל דרגה 1 או 2 יותיר ערכים עצמיים גבוהים יחד עם אלו אשר הפכו להיות 1. מצב כזה כמובן ייצור יחס ערכים עצמיים ירוד לכל אורך התהליך עד לאיטרציה האחרונה, בה כל הערכים העצמיים יהפכו להיות 1. כך שכפי שאנו רואים - האלגוריתמים הניוטוניים למחצה אינם טובים בכל מצב, ועשויים אף לתת ביצועים חלשים בהשוואה ל- NSD במצבים מסויימים.

דוגמה 10.1: נניח כי מטריצת ההסיין $\mathbf{F}$ בעלת n ערכים עצמיים בתחום [1e5,1.001e5] (כלומר - ה- Condition Number של המטריצה נמוך מאוד). במצב זה, אלגוריתם ה- SD (או ה- NSD) יתכנס במספר איטרציות מועט. לעומת זאת, אלגוריתם ה- DFP יגרום לערך עצמי אחד להיות 1 לאחר איטרציה אחת, כשכל השאר נותרים ערכים גבוהים מאוד. כתוצאה מכך, במהלך n-1 איטרציות לאחר מכן יתכנס ה- DFP לאט מאוד בשל Condition Number גרוע.

מהמתואר לעיל גם ניכר כי הבעיה היא בעיית scale בלבד, כיוון שהכפלת המטריצה בקבוע יכולה לשנות בתכלית את המצב, כיוון שהערכים העצמיים של αA הם ערכיה העצמיים של A המוכפלים ב- α. ניתן להציע לפי ראיה זו את הרעיון האלטרנטיבי הבא (ביחס ל- DFP) - באיטרציה ה- k יש לכפול את המטריצה $\mathbf{H}_k$ בפרמטר γ_k, ואז לבצע את עדכון המטריצה. המשוואה הכוללת החדשה תהיה:

$$\mathbf{H}_{k+1} = \gamma_k \mathbf{H}_k + \frac{p_k p_k^T}{q_k^T p_k} - \gamma_k \frac{\mathbf{H}_k q_k q_k^T \mathbf{H}_k}{q_k^T \mathbf{H}_k q_k}$$

אך עדיין נותרה השאלה - כיצד נקבע את ערכו של פרמטר זה כך שהערכים העצמיים של $\mathbf{H}_k\mathbf{F}$ יהיו סביב 1. לצורך זה נדרוש למעשה כי:

$$\lambda_{\min}\{\mathbf{H}_k\mathbf{F}\} \le \frac{1}{\gamma_k} \le \lambda_{\max}\{\mathbf{H}_k\mathbf{F}\}$$

נשתמש בתכונה הבאה של מטריצות:

למה 10.3: לכל מטריצה A ולכל וקטור x קיים הקשר הבא:

$$\lambda_{min}\{A\} \le \frac{x^T A x}{x^T x} \le \lambda_{max}\{A\}$$

ולמעשה: $\lambda_{min}\{A\} = \underset{x}{Min}\frac{x^T A x}{x^T x} \;\; ; \;\; \lambda_{max}\{A\} = \underset{x}{Max}\frac{x^T A x}{x^T x}$

הוכחה: ראה סיפרי אלגברה ליניארית ■

כלומר, כל הכפלה מהצורה הבאה תקיים קשר דומה:

$$\lambda_{min}\{\mathbf{H}_k\mathbf{F}\} \le \frac{x^T \mathbf{H}_k \mathbf{F} x}{x^T x} \le \lambda_{max}\{\mathbf{H}_k\mathbf{F}\}$$

אם נציב $x = p_k$ נקבל:

$$\frac{p_k^T \mathbf{H}_k \mathbf{F} p_k}{p_k^T p_k} = \frac{p_k^T \mathbf{H}_k q_k}{p_k^T p_k} \quad \Rightarrow \quad \gamma_k = \frac{p_k^T p_k}{p_k^T \mathbf{H}_k q_k}$$

ובאופן זה קיבלנו נוסחה לחישובו של הפרמטר γ_k אשר עושה שימוש רק בערכים המצויים בידינו. סימולציות מלמדות כי אלגוריתם זה הרבה יותר חסין להרכב הערכים העצמיים של $\mathbf{F}$, וכן לחיפוש לא מדויק על ישר.

10.6 שיטות ניוטוניות למחצה חסרות זכרון

חסרונם המרכזי של אלגוריתמי Broyden בכך שנדרשת בניה של מטריצה, ולשם כך נדרש זכרון של מטריצה בהפעלת האלגוריתם. נוכל להציע את הוריאציה הבאה הקרויה אלגוריתם ניוטון חסר זכרון - בכל איטרציה, במקום להשתמש ב- $\mathbf{H}_k$ לבניית המטריצה הבאה $\mathbf{H}_{k+1}$, נשתמש בהנחה כי - $\mathbf{H}_k = I$. לדוגמה, עבור אלגוריתם BFGS מתקבל כי המשוואה הרקורסיבית היא:

$$\mathbf{H}_{k+1}^{new} = I + \left(\frac{1 + q_k^T q_k}{q_k^T p_k}\right)\frac{p_k p_k^T}{q_k^T p_k} - \frac{p_k q_k^T + q_k p_k^T}{q_k^T p_k}$$

מסתבר כי האלגוריתם המתקבל הינו וריאציה מוכרת של ה- CG, הקרויה על שמם של Polak & Rebiere.

10.7 סיכום

בפרק זה הצגנו משפחה רחבה של אלגוריתמים הקרויים גישות ניוטוניות למחצה. בדומה לאלגוריתם ניוטון, מפעילות שיטות אלו מטריצה בעת קביעת כיוון החיפוש. בשונה משיטת ניוטון, מטריצה זו המופעלת אינה היפוך ההסיין אלא מטריצה המקרבת את היפוך זה. ראינו מספר דרכים לבניית קירוב זה, אשר עבור הבעיה הריבועית משיגים את בניית ההסיין במדויק כעבור n איטרציות. מגמה מעניינת אשר נחשפה במהלך בנייתם של האלגוריתמים הללו היא קירבתם לאלגוריתם ה- CG, עובדה המחזקת את התחושה כי ה- CG על סוגיו השונים (ויש וריאציות שונות עליו) מהווה אלגוריתם יעיל באופן ייחודי.

פרק 11 - בעיות LS ודרכים לפתרון מערכות של משוואות

11.1 כללי

במדעים המדויקים, בהנדסה, בכלכלה ומתמטיקה נפגשים שוב ושוב עם תבנית אשר קרויה ריבועים פחותים - Least-Squares. תבנית זו בצורתה הכללית נראית כך:

$$f(X)=\sum_{k=1}^{m}\left[r_k(X)\right]^2=\begin{bmatrix} r_1(X) & r_2(X) & \cdots & r_m(X)\end{bmatrix}\begin{bmatrix} r_1(X) \\ r_2(X) \\ \vdots \\ r_m(X)\end{bmatrix}=R^T(X)\cdot R(X)$$

כאשר $\{r_k(X)\}_{k=1}^{m}$ הן פונקציות סקלריות של וקטור המשתנים X בו יש n איברים. ראינו בעבר מספר דוגמאות לתבנית זו - ראה דוגמאות 1.3 עד 1.6, 2.2, 2.3 ועוד. מספר מקורות יכולים לשמש לקבלת תבנית מעין זו:

1. התאמת מדידות למודל - נתון בידינו אוסף מדידות $\{y_k\}_{k=1}^{m}$, אשר אמור להיות מוצא של מערכת בעלת פרמטרים לא ידועים X. דוגמה קלאסית לכך היא דוגמה 1.6 אשר עוסקת ברשתות נוירונים. דוגמה פשוטה יותר אשר נתקלים בה שוב ושוב היא הבאה:

דוגמה 11.1: התאמת צמדי נקודות לפונקציה פרמטרית - נניח כי בידינו אוסף $\{x_k,y_k\}_{k=1}^{m}$ ורצוננו למצוא את הפולינום מסדר n האופטימלי אשר עובר דרך או בסמוך לנקודות אלה. נגדיר קריטריון מרחק:

$$\varepsilon^2(a_0,a_1,\ldots,a_n)=\sum_{k=1}^{m}\left[y_k-\left(a_0x_k^0+a_1x_k^1+\cdots+a_nx_k^n\right)\right]^2$$

והתבנית שהתקבלה הינה בעלת מבנה ריבועים פחותים.

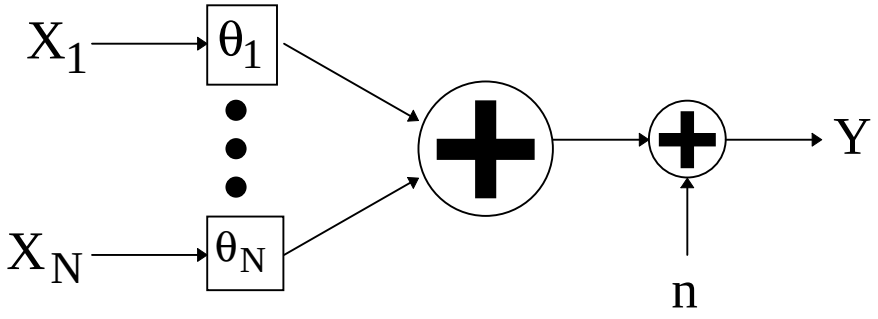
2. פתרון מערכת משוואות - עבור אוסף המשוואות- $r_k(x_1,x_2,\ldots,x_n)=0 \quad k=1,\ldots,m$, ניתן להציע את פתרונן ע"י כלים מאופטימיזציה. לצורך כך נגדיר בעיית מינימיזציה באופן הבא:

$$f(x_1,x_2,\ldots,x_n)=\sum_{k=1}^{m}r_k^2(x_1,x_2,\ldots,x_n)$$

וכך התקבלה התבנית הקלאסית של ה- LS. יש לשים לב לעובדה שלא דרשנו כי מערכת המשוואות תכלול מספר זהה של משוואות ונעלמים. במקרה הכללי, גם למצב בו יש יותר משוואות מנעלמים יש משמעות - במקרה כזה נרצה את הפתרון הטוב ביותר במובן של התקרבות לאיפוס משותף של כל המשוואות - או במקרה כמו כאן - קבלת נורמת l_2 מינימלית של הפרש אגפי המשוואות כולן במשותף. מסתבר עם זאת שיש עניין מיוחד במערכות של n משוואות ב- n נעלמים - אנו נעסוק בכך בהמשך הפרק.

3. משערך הסבירות המירבית - Maximum Likelihood - בבעיות רבות בהן מעורב רעש אקראי ונדרש שערוך של אוסף פרמטרים שאינם ניתנים למדידה ללא רעש, מקובל השימוש במשערך ה- ML. מסתבר כי משערך ה- ML בשילוב עם הסתברויות אקספוננציאליות (גאוסי, לפלסי, גאמה, גיבס ועוד) מניב את ה- LS. מכיוון שנושא זה אינו בהכרח מוכר נציג דוגמה בסיסית להמחשתו.

דוגמה 11.2 : נניח כי בידינו מכונה בעלת N כניסות אשר יסומנו כוקטור X. מכונה זו מחשבת ממוצע משוקלל של כניסותיה ואמורה לספק במוצאה ערך ממוצע זה. בפועל, בשל חיבור לא נאות מתקבל הממוצע המשוקלל כשאליו מצורף רעש גאוסי לבן בעל ממוצע אפס ושונות 1. מכונה זו מתוארת בציור 11.1.



ציור 11.1 - המכונה המתוארת בדוגמה 11.2

חוק הסתברות של רעש זה נראה כך:

$$p(n) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{n^2}{2}\right\}$$

שפירושו שההסתברות לקבל רעש בעל ערך n היא p(n). נוכל לכתוב את הקשר בין כניסת המערכת למוצאה ע"י - $y = X^T\theta + n$, כאשר y הוא מוצא המערכת המזוהם ברעש, ו- θ הוא אוסף המשקלות שאינו ידוע במוצע המשוקלל שמבצעת המערכת. על מערכת זו בוצעו m ניסיונות עם

כניסות שונות $\{X_k\}_{k=1}^m$ והתקבלו המוצאים $\{y_k\}_{k=1}^m$. רצוננו לשערך את משקלות המערכת θ. נוכל לכתוב את חוק ההסתברות הבא של הרעש באופן אלטרנטיבי:

$$p(n_k)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{n_k^2}{2}\right\}=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\left[y_k-X_k^T\theta\right]^2}{2}\right\}=p\{y_k,X_k/\theta\}$$

מכיוון שהרעש ב- m המדידות הינו לבן, פירוש הדבר הוא שההסתברות לכל מדגם הניסויים היא:

$$p\left\{\{y_k,X_k\}_{k=1}^m/\theta\right\}=\prod_{k=1}^m\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\left[y_k-X_k^T\theta\right]^2}{2}\right\}\right]=C\cdot\exp\left\{-\frac{1}{2}\sum_{k=1}^m\left[y_k-X_k^T\theta\right]^2\right\}$$

כיוון שזוהי מכפלה ישירה של ההסתברויות לכל ניסוי בנפרד. משערך הסבירות המירבית מציע כשערוך את θ אשר יביא למקסימום את ההסתברות הנ"ל. לכן:

$$\hat{\theta}_{ML}=\arg\max_{\theta}p\left\{\{y_k,X_k\}_{k=1}^m/\theta\right\}=\arg\min_{\theta}\left\{\sum_{k=1}^m\left[y_k-X_k^T\theta\right]^2\right\}$$

כך שקיבלנו את תבנית ה- LS.

בפרק זה נסקור שיטות שונות אשר נועדו לבעיות בעלות מבנה LS. מסתבר כי מבנה זה מספק אפשרויות ייחודיות אשר כדאי לנצלן לשם השגת פתרון איטרטיבי מהיר. כמו כן, פרק זה יציג את נושא הפתרון של מערכות משוואות אשר לעניינו הינו מקרה פרטי של בעיית LS.

11.2 פתרון בעיות LS ליניאריות

ראינו כי מבנה כללי של בעיית LS נתון ע"י:

$$f(X)=\sum_{k=1}^m\left[r_k(X)\right]^2=\begin{bmatrix}r_1(X) & r_2(X) & \cdots & r_m(X)\end{bmatrix}\begin{bmatrix}r_1(X)\\ r_2(X)\\ \vdots\\ r_m(X)\end{bmatrix}=R^T(X)\cdot R(X)$$

נתחיל את הדיון ב- LS בבעיות בעלות המבנה:

$$r_k(X) = y_k - h_k^T X \quad \Rightarrow \quad R(X) = Y - HX$$

כאשר H מטריצה בגודל $[m \times n]$. ע"י גזירה והשוואה לאפס מתקבל כי:

$$\frac{\partial f(X)}{\partial X} = 0 = -2H^T[Y - HX] \quad \Rightarrow \quad \left[H^T H\right]X = H^T Y$$

המשוואה הנ"ל נקראית המשוואה הנורמלית. כאשר $\left[H^T H\right]$ הינה הפיכה ניתן לפתור את המשוואה הנ"ל. מטריצה זו תהיה הפיכה כאשר המטריצה H תהיה בעלת גודל $[m \times n]$ המקיים - $m \geq n$ (כלומר - יותר משוואות מאשר נעלמים), אך בכך אין די - יש לדרוש כי המטריצה H תהיה בעלת דרגה מלאה, כלומר - שכל עמודותיה בלתי תלויות ליניארית. עבור המקרה בו $m = n$ משמעות דרישה זו היא כי הדטרמיננט של H אינו אפס (כלומר - היא אינה סינגולרית).

התוצאה הנ"ל מסגירה הרכב נוסף המוכר מאלגברה ליניארית - מושג ההיפוך המדומה - Pseudo-Inverse. בהנחה כי $\left[H^T H\right]$ אינה סינגולרית, התוצאה שהתקבלה מהמשוואה הנורמלית היא:

$$X = \left[H^T H\right]^{-1} H^T Y = H^{\#} Y$$

כאשר $H^{\#}$ הוא ההיפוך המדומה של המטריצה H. יש לשים לב לתכונות הבאות:

1. מתקיים - $H^{\#}H = \left[H^T H\right]^{-1} H^T H = I_{[n \times n]}$ - משמעות תכונה זו היא ההיפוך המדומה מתנהג כהיפוך לכל דבר כשהוא מוכפל מימין.
2. בדרך כלל - $HH^{\#} = H\left[H^T H\right]^{-1} H^T \neq I_{[m \times m]}$ - בהכפלה משמאל מתקבל כי אין תוצאה כשל היפוך. זה סביר לנוכח העובדה שבמכפלה זו מעורבות שתי מטריצות מדרגה n , כשהתוצאה הינה בגודל $[m \times m]$, כאשר m>n, ולכן לא ניתן לצפות לתוצאה בעלת דרגה מלאה כשל מטריצת יחידה.
3. עבור $m = n$ מתקבל כי $H^{\#} = \left[H^T H\right]^{-1} H^T = H^{-1} H^{-T} H^T = H^{-1}$ - כלומר ישנה התלכדות במקרה זה בין ההיפוך הרגיל להיפוך המדומה.

4. המטריצה $HH^{\#} = H\left[H^TH\right]^{-1}H^T$ הינה מטריצת היטל - כלומר מתקיים כי הפעלתה שוב ושוב שקולה להפעלה יחידה. פירוש הדבר הוא כי - $\left[HH^{\#}\right]^k = HH^{\#}$.

5. המטריצה $HH^{\#} = H\left[H^TH\right]^{-1}H^T$ סימטרית - קל לראות זאת ע"י בדיקה ישירה.

עבור הפתרון האופטימלי שהתקבל - $X = \left[H^TH\right]^{-1}H^TY = H^{\#}Y$ - מתקבל כי ערך הפונקציה עבור וקטור זה היא:

$$f(X) = [Y - HX]^T[Y - HX] = \left[Y - HH^{\#}Y\right]^T\left[Y - HH^{\#}Y\right] =$$
$$= Y^T\left[I - HH^{\#} - HH^{\#} + HH^{\#}HH^{\#}\right]Y = Y^T\left[I - HH^{\#}\right]Y$$

בטרם נעזוב את נושא זה של LS ליניארי נזכיר כי בעיה זו נתונה כתבנית ריבועית לכל דבר ולכן כל השיטות הנומריות שהוצגו עד כה - SD, NSD, CG, גישות ניוטוניות וניוטוניות למחצה - כולם יעילים כאן, וכל שיטה מתאימה למימד שונה של בעיה - עבור מימד נמוך (עד n=200) כדאי להפעיל את הגישה הניוטונית שאינה אלא פתרון ישיר כפי שהוצג כאן. עבור $200 < n < 400$ כדאי להשתמש בגישות הניוטוניות למחצה. החסם מלמעלה על גודל הבעיה נובע מכך שגישות אלה מקדמות מטריצה שגודלה $[n \times n]$ באופן רקורסיבי, ומטריצה זו גדולה וצורכת זיכרון רב. עבור בעיות מעל n=400 כדאי להשתמש ב- CG.

11.3 בעיית ה- LS הכללית - אלגוריתם Gauss-Newton

בעיית ה- LS הכללית נתונה כפי שראינו ע"י:

$$f(X) = \sum_{k=1}^{m}\left[r_k(X)\right]^2 = \begin{bmatrix} r_1(X) & r_2(X) & \cdots & r_m(X) \end{bmatrix}\begin{bmatrix} r_1(X) \\ r_2(X) \\ \vdots \\ r_m(X) \end{bmatrix} = R^T(X)\cdot R(X)$$

נחשב אנליטית את נגזרת פונקציה זו הראשונה:

$$\frac{\partial f(X)}{\partial X} = \sum_{k=1}^{m} 2\frac{\partial r_k(X)}{\partial X}r_k(X) = 2\begin{bmatrix} \frac{\partial r_1(X)}{\partial X} & \frac{\partial r_2(X)}{\partial X} & \cdots & \frac{\partial r_m(X)}{\partial X} \end{bmatrix}R(X) = 2A(X)R(X)$$

$A(X)$ היא מטריצה בגודל $[n \times m]$ הנקראית מטריצת היעקובייאן. הנגזרת השניה (ההסיין) נתונה ע"י:

$$\frac{\partial^2 f(X)}{\partial X^2} = \sum_{k=1}^{m} 2\frac{\partial r_k(X)}{\partial X}\frac{\partial r_k{}^T(X)}{\partial X} + \sum_{k=1}^{m} 2\frac{\partial^2 r_k(X)}{\partial X^2} r_k(X) =$$
$$= 2A(X)A^T(X) + 2\sum_{k=1}^{m} \frac{\partial^2 r_k(X)}{\partial X^2} r_k(X)$$

רצוננו להציע אלגוריתמים לפתרון איטרטיבי של בעיית LS כללית זו. ברור כי נוכל להציע את כל מגוון הגישות שהוצגו עד כה כגישות ישימות לבעיה החדשה, אך הרכבה המיוחד יאפשר לנו כפי שנראה גישה ייחודית אשר תניב אלגוריתם יעיל במיוחד.

הרעיון אותו נקדם כעת הוא הבא - בשל רצוננו להביא למינימום את סכום הריבועים של ערכי הפונקציות $r_k(X)$, נראה סביר כי ערכים אלו יהיו <u>קטנים מאוד</u> (1) בקרבת הפתרון. לכן נוכל להציע את הקירוב הבא עבור ההסיין:

$$\frac{\partial^2 f(X)}{\partial X^2} = 2A(X)A^T(X) + 2\sum_{k=1}^{m} \frac{\partial^2 r_k(X)}{\partial X^2} r_k(X) \cong 2A(X)A^T(X)$$

לקירוב זה משמעות נוספת עמוקה יותר - עבור $r_k(X)$ <u>ליניארי</u> (2) (כלומר - $r_k(X) = y_k - H_k X$) מתקבל כי הנגזרות השניות מתאפסות והקירוב הופך למדויק. לכן, משמעות קירוב זה הינה ליניאריזציה של $r_k(X)$.

קירוב זה אשר מוצע כאן הינו בעל ערך חשוב כיוון שאנו משיגים את חישוב ההסיין דרך חישוב נגזרות ראשונות בלבד. בעבר ראינו גישות אשר תוך שימוש בנגזרות ראשונות הפיקו את ההסיין - הגישות הניוטוניות למחצה. עם זאת, בגישות ההן נדרשו n איטרציות על מנת לקבל קירוב מלא להסיין, בעוד שכאן אנו מקבלים קירוב זה בצעד אחד. תוצאה זו נובעת ישירות מהמבנה הייחודי של ה- LS. עוד הנמקה לנכונות קירוב זה היא העובדה שערכי $r_k(X)$ מתקבלים בדרך כלל <u>חיוביים ושליליים בהסתברות אחידה</u> (3), וכך יוצא שהגורם אשר הוזנח כולל מטריצות חיוביות מוגדרות העוברות מעיין חשבון ממוצע ולכן סכום זה צפוי להיות קטן.

מהתיאור לעיל נובע כי שלוש סיבות טובות מאוד (ממוספרות ומודגשות) מביאות לקירוב בו אנו מזניחים את הגורם השני בביטוי ההסיין, המורכב מנגזרות שניות של $r_k(X)$. כעת שבידינו נגזרות מסדר ראשון ושני נוכל להשתמש בהן לשם הצעת אלגוריתם איטרטיבי לפתרון בעיית LS

כללית. האלגוריתם אשר מתואר בציור 11.2 קרוי אלגוריתם Gauss-Newton. נשים לב כי משוואת האיטרציה המוצעת הינה:

$$X_{k+1} = X_k - \left[A(X_k)A^T(X_k)\right]^{-1} A(X_k)R(X_k) = X_k - \left[A^T(X_k)\right]^{\#} R(X_k)$$

ולו היה קירוב ההסיין מדויק היתה זו השיטה הניוטונית אשר הצגנו בפרק 7. כמובן כי במצב כזה היה מתקבל כי הבעיה הינה ריבועית (הראינו זאת קודם), ולכן די באיטרציה אחת כדי להגיע לפתרון.

1. בחר נקודת איתחול שרירותית X_0.
2. באיטרציה ה- k חשב את מטריצת הגרדיאנטים $A(X)$.
3. פתור את מערכת המשוואות למציאת הכיוון d - $A(X_k)A^T(X_k)d_k = -A(X_k)R(X_k)$.
4. בצע את העדכון לפי $X_{k+1} = X_k + d_k$.
5. עצור את האלגוריתם כאשר $\|d_k\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 11.2 - סיכום אלגוריתם Gauss-Newton לבעיית LS כללית</u>

נדון כעת בתכונות ההתכנסות של אלגוריתם זה.

משפט 11.1: עבור בעיית ה- LS הכללית נסמן את הפתרון ע"י X^*. אזי, אם אלגוריתם Gauss-Newton מאותחל קרוב דיו לנקודה X^* מתקיים אי השיויון הבא:

$$\|X_{k+1} - X^*\| \le \left\| \left[A(X^*)A(X^*)\right]^{-1} \cdot \left[\sum_{k=1}^{m} r_k(X^*) \frac{\partial^2 r_k(X^*)}{\partial X^2}\right] \right\| \cdot \|X_k - X^*\| + O\left\{\|X_k - X^*\|^2\right\}$$

הוכחה: ההוכחה דומה להוכחה אשר ניתנה עבור אלגוריתם ניוטון בפרק 7 - נחזור על עיקריה. ממשוואת האיטרציה נובע:

$$X_{k+1}-X^*=X_k-X^*-\left[A(X_k)A^T(X_k)\right]^{-1}A(X_k)R(X_k)$$

נוכל להכניס את הקשר הבא העושה שימוש בעובדה שהגרדיאנט בנקודה X^* מתאפס:

$$X_k-X^*=\left[A(X_k)A^T(X_k)\right]^{-1}\left[A(X^*)R(X^*)+\left[A(X_k)A^T(X_k)\right](X_k-X^*)\right]$$

וצירוף שתי משוואות אלו יחדיו נותן:

$$\begin{aligned}X_{k+1}-X^*&=\left[A(X_k)A^T(X_k)\right]^{-1}\left[A(X^*)R(X^*)+\left[A(X_k)A^T(X_k)\right](X_k-X^*)\right]-\\&\quad-\left[A(X_k)A^T(X_k)\right]^{-1}A(X_k)R(X_k)=\\&=\left[A(X_k)A^T(X_k)\right]^{-1}\left[A(X^*)R(X^*)-A(X_k)R(X_k)+\left[A(X_k)A^T(X_k)\right](X_k-X^*)\right]\end{aligned}$$

הביטוי בסוגריים הינו קירוב טיילור של הגרדיאנט ומתקיים:

$$\begin{aligned}A(X^*)R(X^*)&=A(X_k)R(X_k)+\left[A(X_k)A^T(X_k)\right](X^*-X_k)+\\&\quad+\left[\sum_{k=1}^{m}r_k(X_k)\frac{\partial^2 r_k(X_k)}{\partial X^2}\right](X^*-X_k)+O\left\{(X_k-X^*)^T(X_k-X^*)\right\}\end{aligned}$$

הצבת תוצאה זו במשוואה הקודמת נותנת:

$$\begin{aligned}X_{k+1}-X^*&=\left[A(X_k)A^T(X_k)\right]^{-1}\left[\sum_{k=1}^{m}r_k(X_k)\frac{\partial^2 r_k(X_k)}{\partial X^2}\right](X^*-X_k)+\\&\quad+\left[A(X_k)A^T(X_k)\right]^{-1}O\left\{(X_k-X^*)^T(X_k-X^*)\right\}\end{aligned}$$

כיוון שמשוואה זו סימטרית בהתיחסותה ל- X^* ול- X_k, נוכל להחליף את תפקידיהם, והפעלת נורמה על שני אגפי המשוואה תיתן:

$$\|X_{k+1}-X^*\|\le\left\|\left[A(X^*)A^T(X^*)\right]^{-1}\sum_{k=1}^{m}r_k(X^*)\frac{\partial^2 r_k(X^*)}{\partial X^2}\right\|\|X_k-X^*\|+O\left\{\|X_k-X^*\|^2\right\}$$

והמשפט הוכח ■

משמעויות משפט זה הן הבאות:

1. סדר ההתכנסות הינו לכל הפחות ליניארי.
2. כאשר בעיית ה- LS מתיחסת למערכת של n משוואות ב- n נעלמים וקיים עבורן פתרון מתקבל כי $\{r_k(X^*)=0\}_{k=1}^{n}$ ואז הגורם הראשון באי-השיויון הנ"ל נופל ומתקבל סדר התכנסות ריבועי.
3. במצבים בהם הגורם שהוזנח בהסיין אינו קטן ייתכנו מצבים בהם אלגוריתם זה יתבדר אפילו אם אנו סמוכים לפתרון. במצבים אלו מתקבל כי מקדם הגורם $\|X_k - X^*\|$ באי-השיויון הנ"ל גדול מ- 1 ואז משוואה רקורסיבית זו אינה מתארת התכנסות.
4. מסקנה מכל האמור לעיל - אלגוריתם Gauss-Newton כפי שתואר כאן אינו מומלץ לשימוש כיוון שהוא נסמך על שתי הנחות אשר אינן מתקיימות בהכרח - ההנחה כי האיתחול סמוך לפתרון האופטימלי, וההנחה כי הגורם שהוזנח בהסיין קטן יחסית.

בשל חסרונותיו אלה של אלגוריתם Gauss-Newton הוצעו במהלך השנים גרסאות בהן היוה אלגוריתם זה בסיס ונעשו בו שינויים שונים. הגירסה הרווחת הינה אלגוריתם Gauss-Newton עם חיפוש על ישר - הרעיון פשוט למדי - האלגוריתם נותר כמקודם, אך הוקטור d_k מגדיר כיוון חיפוש ולא גודל עדכון מדויק, ולפי כיוון חיפוש זה מבוצע חיפוש על ישר לאיתור נקודה אופטימלית. סיכום אלגוריתם זה מובא בציור 11.3.

1. בחר נקודת איתחול שרירותית X_0.
2. באיטרציה ה- k חשב את מטריצת הגרדיאנטים $A(X)$.
3. פתור את מערכת המשוואות למציאת הכיוון d - $A(X_k)A^T(X_k)d_k = -A(X_k)R(X_k)$.
4. חפש α שיתן מינימום על הישר $f(X_k+\alpha d_k)$.
5. בצע את העדכון לפי $X_{k+1} = X_k + \alpha_k d_k$.
6. עצור את האלגוריתם כאשר $\|d_k\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

<u>ציור 11.3 - סיכום אלגוריתם Gauss-Newton עם חיפוש על ישר לבעיית LS כללית</u>

באלגוריתם זה, כמקודם, קירוב ההסיין מתקבל בצעד אחד - יתרון משמעותי בהשוואה לשיטות הניוטוניות למחצה. תכונה חשובה נוספת היא שהקירוב להסיין לפי הגדרתו הוא חיובי חצי מוגדר לפחות - עובדה המבטיחה כיוון ירידה בפונקציה. בנוסף, נסיון אמפירי מלמד כי שיטה זו עולה בהרבה על השיטות הניוטוניות למחצה עבור בעיות LS כלליות מבחינת קצב ההתכנסות. עם זאת, צריך לציין כי ישנן בעיות LS בהן קירוב ההסיין המוצע גרוע, ומתעוררות בעיות התכנסות. במקרים אלה מומלץ להשתמש בגישות ניוטוניות למחצה (BFGS) כשההסיין מאותחל להיות הניחוש המוצע כאן.

גישה נוספת אותה פגשנו בהקשר לשיטות הניוטוניות היא שיטת Levenbeg-Marquardt. שיטה זו ישימה גם כאן - נוכל להציע כי הקירוב להסיין תהיה המטריצה:

$$A(X_k)A^T(X_k)+\lambda I$$

כאשר λ הינו פרמטר חיובי אותו יש לבחור. ישנו מגוון גישות העושות שימוש במוטיב זה בשילובים שונים - לא נסקור זאת כאן.

11.4 בעיית LS לא ליניארית - מבנים ייחודיים

נתיחס לבעיית המינימיזציה הבאה - $f(X)=[Y-S(X)]^T[Y-S(X)]$. במקרים מסויימים קיים מבנה נוח של $S(X)$ אשר מאפשר המרת הבעיה במבנה קל יותר לטיפול. במסגרת סעיף זה נציג שני מצבים כאלה, והדרכים לפתרונם.

מקרה 1 - נניח כי קיים קשר חד-חד ערכי $Z=G(X)$ באופן שנותן כי:

$$S(X)=S\left(G^{-1}(Z)\right)=HZ$$

במקרה כזה נוכל להמיר את הבעיה הכללית שבידינו לבעיה ליניארית, שפתרונה קל לאין ערוך. בשל המעבר החד-חד ערכי בין X ל- Z מתקבל כי מציאת ה- Z האופטימלי וחישוב X מתוכו מבטיח את מציאת X האופטימלי. הפתרון ייעשה תחילה לבעיה הבאה:

$$f(X)=[Y-S(X)]^T[Y-S(X)]=[Y-HZ]^T[Y-HZ]$$
$$\Rightarrow Z_{opt}=\left[H^TH\right]^{-1}H^TY$$

והוקטור הנעלם המקורי X יימצא פשוט ע"י הקשר - $Z = G(X)$ (כיוון ש- G הפיכה יש תוצאה חד משמעית).

דוגמה 11.3 : נניח כי בידינו אוסף צמדים $\{x_k, y_k\}_{k=1}^{m}$ המהוים דגימה רועשת של האות ההרמוני - $A\cos\{2\pi f_0 x_k + \phi\}$. רצוננו לשערך את האמפליטודה והפאזה של הרמוניה זו - נניח כי התדר ידוע. לשם כך נציע את הקריטריון הבא למינימיזציה:

$$\varepsilon^2(A,\phi) = \sum_{k=1}^{m}\left[y_k - A\cos(2\pi f_0 x_k + \phi)\right]^2$$

בבירור ניכר כי הבעיה הנתונה הינה LS לא ליניארי. אבל נוכל לכתוב כי:

$$A\cos\{2\pi f_0 x_k + \phi\} = A\cos(\phi)\cos(2\pi f_0 x_k) - A\sin(\phi)\sin(2\pi f_0 x_k)$$

לכן נוכל להציע את צמד הנעלמים הבאים כתחליף - $\alpha_1 = A\cos(\phi)$, $\alpha_2 = A\sin(\phi)$, והבעיה המתקבלת תהיה:

$$\varepsilon^2(\alpha_1,\alpha_2) = \sum_{k=1}^{m}\left[y_k - \alpha_1\cos(2\pi f_0 x_k) + \alpha_2\sin(2\pi f_0 x_k)\right]^2$$

וזו בעיה ליניארית בנעלמים. ההתמרה הפולרית אשר העבירה אותנו מהמשתנים המקוריים לחדשים הפיכה כך שכל התנאים מתקיימים לשם קבלת פתרון מדויק ובקלות לבעיה אשר במבט ראשון נראית מסובכת. הפתרון יהיה הבא:

$$\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix} = \left[H^T H\right] H^T Y$$

כאשר:

$$H = \begin{bmatrix}\cos(2\pi f_0 x_1) & -\sin(2\pi f_0 x_1)\\ \cos(2\pi f_0 x_2) & -\sin(2\pi f_0 x_2)\\ \vdots & \\ \cos(2\pi f_0 x_m) & -\sin(2\pi f_0 x_m)\end{bmatrix} \qquad Y = \begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_m\end{bmatrix}$$

ומתוך פתרון זה מתקבל: $A=\sqrt{\alpha_1^2+\alpha_2^2},\ \ \phi=tg^{-1}\left\{\frac{\alpha_1}{\alpha_2}\right\}$.

מקרה 2 - במקרים רבים מתקבל כי ניתן להפריד את $S(X)$ למכפלת שני גורמים פרידים במשתנים, דהיינו - $S(X)=H(\alpha)\cdot\beta$ כאשר α ו- β הם תת-וקטורים המייצגים את משתני הוקטור X או כאלו המחליפים אותם באופן חד-חד-ערכי (כמקודם). H הינה מטריצה התלויה באופן לא ליניארי ב- α.

במצב כזה ניתן לפתור תחילה את הבעיה עבור וקטור הנעלמים β באופן הבא:

$$f(\alpha,\beta)=[Y-H(\alpha)\beta]^T[Y-H(\alpha)\beta]$$
$$\Rightarrow\beta_{opt}(\alpha)=\left[H^T(\alpha)H(\alpha)\right]H^T(\alpha)Y$$

והצבת פתרון זה בפונקציה המקורית נותנת:

$$f\left(\alpha,\beta_{opt}(\alpha)\right)=Y^T\left[I-\left[H^T(\alpha)H(\alpha)\right]H^T(\alpha)\right]Y$$

ולכן הבעיה אשר נותרה הינה מקסימיזציה של הפונקציה:

$$f\left(\alpha,\beta_{opt}(\alpha)\right)=Y^T\left[H^T(\alpha)H(\alpha)\right]H^T(\alpha)Y$$

דוגמה 11.4: נניח כי נתונה לנו המשפחה הפרמטרית הבאה של פונקציות בסיס חד-מימדיות- $\psi_k(x,\alpha)=e^{-\alpha kx}\quad k=1\ldots K_0$. בהינתן לנו פונקציה $g(x)$ רצוננו לייצגה כצירוף ליניארי של פונקציות בסיס אלה. כלומר, נחפש ייצוג מהצורה:

$$g(x)=\sum_{k=1}^{K_0}\beta_k\psi_k(x,\alpha)=\sum_{k=1}^{K_0}\beta_k\exp\{-\alpha kx\}$$

נציע את איתור פרמטרי הייצוג המוצלחים ביותר ע"י פתרון בעיית LS הבאה:

$$\varepsilon^2\left(\alpha,\beta_1,\ \ldots\ ,\beta_{k_0}\right)=\sum_{j=1}^{m}\left[g\left(x_j\right)-\sum_{k=1}^{K_0}\beta_k\exp\left\{-\alpha kx_j\right\}\right]^2$$

כאשר דגמנו את ציא ה- x ב- m נקודות פזורות בתחום ההגדרה, ואנו מחפשים התאמה מירבית. רואים מיד כי המבנה המתקבל ספרבילי במשתנים, והנעלמים β_k מצריכים פתרון בעיה ליניארית. מתקבל:

$$\beta(\alpha)=\begin{bmatrix}\beta_1(\alpha)\\ \vdots \\ \beta_m(\alpha)\end{bmatrix}=\left[H^T(\alpha)H(\alpha)\right]H^T(\alpha)Y$$

כאשר:

$$H=\begin{bmatrix}\exp\{-\alpha x_1\} & \exp\{-2\alpha x_1\} & \cdots & \exp\{-K_0\alpha x_1\}\\ \exp\{-\alpha x_2\} & \exp\{-2\alpha x_2\} & \cdots & \exp\{-K_0\alpha x_2\}\\ \vdots & \vdots & \ddots & \vdots \\ \exp\{-\alpha x_m\} & \exp\{-2\alpha x_m\} & \cdots & \exp\{-K_0\alpha x_m\}\end{bmatrix} \qquad Y=\begin{bmatrix}g(x_1)\\ g(x_2)\\ \\ g(x_m)\end{bmatrix}$$

ואת α נמצא ע"י אלגוריתמי חיפוש מקסימום על הישר - $f(\alpha)=Y^T\left[H^T(\alpha)H(\alpha)\right]H^T(\alpha)Y$.

11.5 בעיות LS והקשר לפתרון מערכות של משוואות

הזכרנו קודם כי ניתן להסתכל על תבנית ה- LS כנסיון המרה של בעיית פתרון מערכת משוואות בבעיית אופטימיזציה. נניח כי אנו מקבלים מערכת של n משוואות (לאו דוקא ליניאריות), ב- n נעלמים. אוסף משוואות אלה ובעיית ה- LS המוגדרת מתוכן יהיו:

$$\left\{r_k(x_1,x_2,\ \dots\ ,x_n)=0\right\}_{k=1}^{n} \quad\Leftrightarrow\quad R(X)=0 \quad\Leftrightarrow\quad f(X)=\sum_{k=1}^{n}r_k^2(X)=R^T(X)R(X)$$

הגישה המוכרת ביותר בהקשר לפתרון מערכות של משוואות הינה ליניארצ'יזה מקומית ופתרון באופן איטרטיבי. נניח כי בידינו פתרון מוצע X_k ורצוננו לעדכנו. ליניאריזציה של מערכת המשוואות נותנת:

$$R(X)=R(X_k)+A^T(X_k)(X-X_k)$$

כאשר A היא מטריצת היעקובייאן אותה ראינו בסעיפים קודמים. כיוון שרצוננו לאפס את R(X) מתקבלת משוואת העדכון הבאה:

$$X_{k+1} = X_k - A^{-T}(X_k)R(X_k)$$

כזכור, משוואת העדכון של אלגוריתם Gauss-Newton היתה:

$$X_{k+1} = X_k - \left[A^T(X_k)\right]^{\#} R(X_k)$$

וכיוון שכאן m=n כלומר A ריבועית, קיבלנו למעשה אותה משוואת עדכון, כלומר - ליניאריזציה עבור מערכות של משוואות מניבות את אלגוריתם Gauss-Newton. למעשה, גירסה חד-מימדית של אלגוריתם זה מוכרת בשם אלגוריתם Newton-Raphson, המתואר באופן הבא:

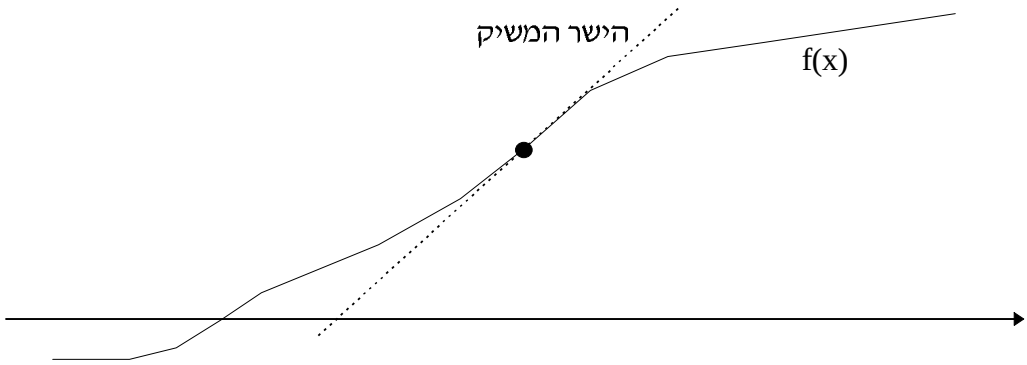
נתונה הפונקציה f(x), ורצוננו למצוא את נקודות חצייתה את ציר ה- x, דהיינו, הנקודות המקיימות f(x)=0. אלגוריתם Newton-Raphson מציע לבחור נקודה כלשהי x_0, למצוא את משוואת המשיק לפונקציה בנקודה זו - אשר יהיה הביטוי הליניארי:

$$m(x) = f'(x_0)\cdot[x - x_0] + f(x_0)$$

ולקבוע את הפתרון המוצע כנקודת חציית המשיק את ציר ה- x, דהיינו:

$$m(x) = f'(x_0)\cdot[x - x_0] + f(x_0) = 0 \qquad \Rightarrow \qquad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

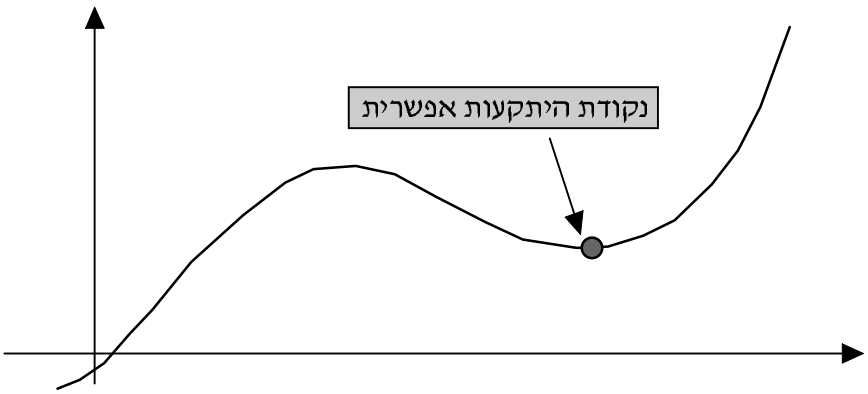
והדמיון למשוואה הרב-מימדית שניתנה לעיל ניכר. לפיכך, נכנה גם את השיטה הרב-מימדית שתוארה כאן בשם אלגוריתם Newton-Raphson. ציור 11.4 מתאר את אלגוריתם זה לחד-מימד.



ציור 11.4 - תיאור של אלגוריתם Newton-Raphson

מסתבר כי ניתן להוכיח (ואנו לא נעשה זאת) כי קצב ההתכנסות של אלגוריתם זה לפתרון מערכת המשוואות הינו מסדר שני.

חסרון מהותי של שיטה זו היא החשש שהמטריצה A תהיה סינגולרית או קרובה לכזו, עובדה אשר יוצרת אי-יציבות נומרית בתהליך הפתרון. חסרון נוסף הוא האפשרות להינעל על מינימה לוקלית בה מערכת המשוואות לא מתאפסת. קל לראות אפשרות זו על פונקציה חד-מימדית, כמתואר בציור 11.5.



ציור 11.5 - מינימה לוקלית באלגוריתם Newton-Raphson

גישה אחרת לפתרון מערכת משוואות היא גישת ההחלפה. נניח כי את $R(X)$ ניתן לפרק באופן הבא - $R(X) = X - \Phi(X) = 0$. אזי, ברור כי פתרון מערכת המשוואות חייב לקיים את המשוואה - $X^* = \Phi(X^*)$. נציע אם כך את האלגוריתם הבא אשר מתואר בציור 11.6.

1. בחר נקודת איתחול שרירותית X_0.
2. באיטרציה ה- k בצע עדכון לפי $X_{k+1} = \Phi(X_k)$.
3. עצור את האלגוריתם כאשר $\|X_{k+1} - \Phi(X_k)\| \le \varepsilon$, כאשר ε הוא ערך סף שנקבע מראש.

ציור 11.6 - אלגוריתם ההחלפה לפתרון מערכות של משוואות

דוגמה 11.5: נניח כי רצוננו לפתור את המשוואה - $r(x)=1+x-\exp(2x)$. לפנינו שתי גישות ישירות ליישום אלגוריתם ההחלפה. אפשרויות אלה הן:

$$x_{k+1}=\exp\{2x_k\}-1 \qquad x_{k+1}=\frac{1}{2}\log\{x_k+1\}$$

גישת Newton-Raphson לפתרון משוואה זו תהיה המשוואה הבאה:

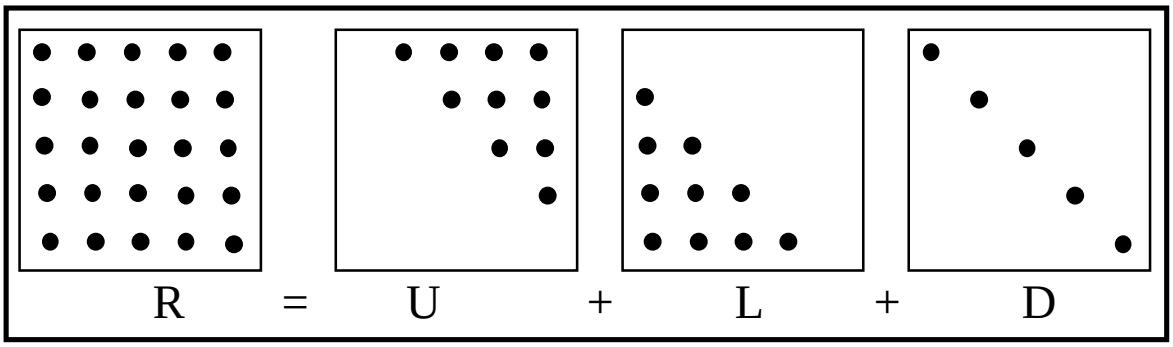
$$x_{k+1}=x_k-\frac{1+x_k-\exp(2x_k)}{1-2\exp(2x_k)}=\frac{(1-2x_k)\exp(2x_k)-1}{1-2\exp(2x_k)}$$

נמחיש את הפעלת שלושת משוואות אלה ע"י הטבלה הבאה המתארת את הערכים שהתקבלו בשלושת השיטות עבור איתחול בערך $x_0=1$.

מספר איטרציה	$x_{k+1}=\exp\{2x_k\}-1$	$x_{k+1}=\frac{1}{2}\log\{x_k+1\}$	Newton-Raphson
1	6.389	0.3466	0.6089
2	3.5e5	0.1488	0.3014
3	התבדרות	0.0693	0.1033
4		0.0335	0.0168
5		0.0165	0.0005
6		0.0082	1e-7

<u>טבלה 11.1 - תוצאות הרצה עבור דוגמה 11.5</u>

מקרה פרטי מעניין של שיטת ההחלפה מוצעת עבור פתרון מערכת משוואות ליניארית. בהינתן מערכת מהצורה RX=P, כאשר R מטריצה ריבועית שאינה סינגולרית, ניתן להציע את פיצולה של המטריצה לשלושה גורמים - מטריצה משולשת עליונה ממש U (שכוללת את כל איברי R אשר מצויים מעל האלכסון הראשי), מטריצה משולשת תחתונה ממש, ומטריצה אלכסונית D אשר כוללת את האלכסון הראשי - ראה ציור 11.7.



ציור 11.7 - פיצול המטריצה R לשם הפעלת שיטת ההחלפה

המשוואה ניתנת לרישום כעת בהתבסס על הפירוק - $[U+L+D]X=P$. ניתן אם כך להציע את את שתי האפשרויות הבאות אשר מוכרות היטב בהקשר לפתרון נומרי של משוואות דיפרנציאליות חלקיות:

1. $X_{k+1}=D^{-1}\left[P-(U+L)X_k\right]$ - שיטה זו קרויה אלגוריתם Jacoby.
2. $X_{k+1}=[D+L]^{-1}\left[P-UX_k\right]$ - שיטה זו קרויה אלגוריתם Gauss-Siedel (GS).

נשים לב לתכונה המעניינת הבאה - כאשר המטריצה R חיובית מוגדרת, ראינו בפרק 7 כי ניתן להתיחס למערכת המשוואות RX=P כגרדיאנט של הבעיה הריבועית הבאה:

$$\varepsilon^2(X)=\frac{1}{2}X^TRX-P^TX+C$$

ואלגוריתם ה- DNSD אשר הצגנו בפרק 7 הציע את משוואת האיטרציה הבאה:

$$X_{k+1}=X_k+H^{-1}\left[P-RX_k\right]$$

כאשר H מטריצה אלכסונית הכוללת על אלכסונה את הגרדיאנטים השניים, שאינם אלא איברי האלכסון הראשי של המטריצה R, כלומר - H=D. הצבת קשר זה נותנת כי משוואת ה- DNSD היא:

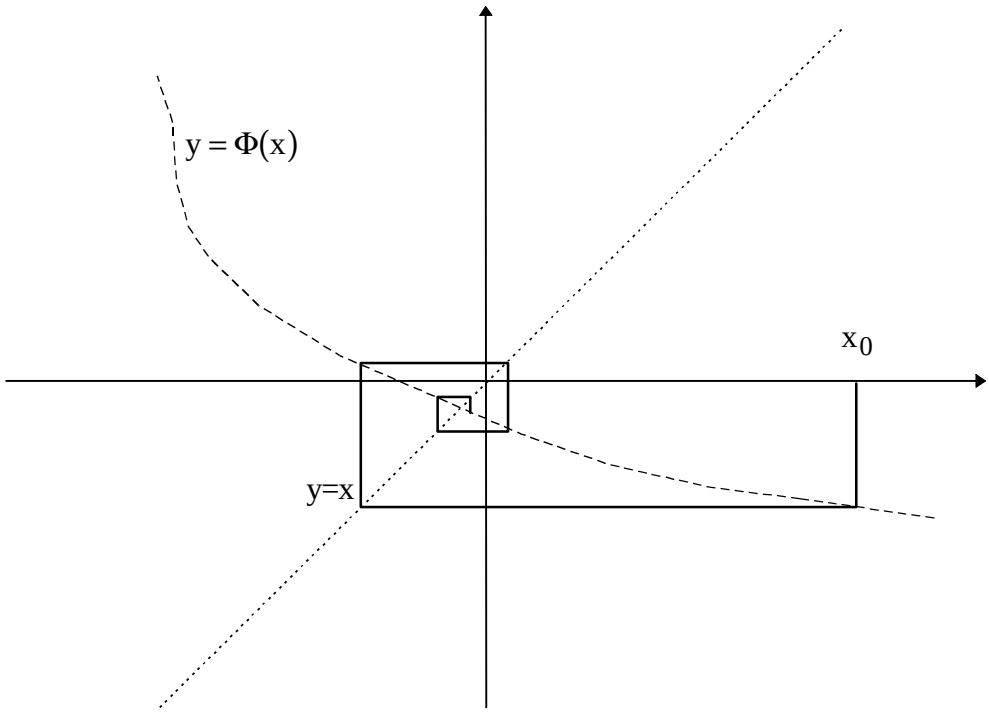
$$X_{k+1}=X_k+D^{-1}\left[P-(U+L+D)X_k\right]=D^{-1}\left[P-(U+L)X_k\right]$$

וכך קיבלנו התלכדות של אלגוריתם Jacoby עם אלגוריתם ה- DNSD - שני אלגוריתמים עם יסודות שונים לחלוטין מניבים אותה משוואת עדכון.

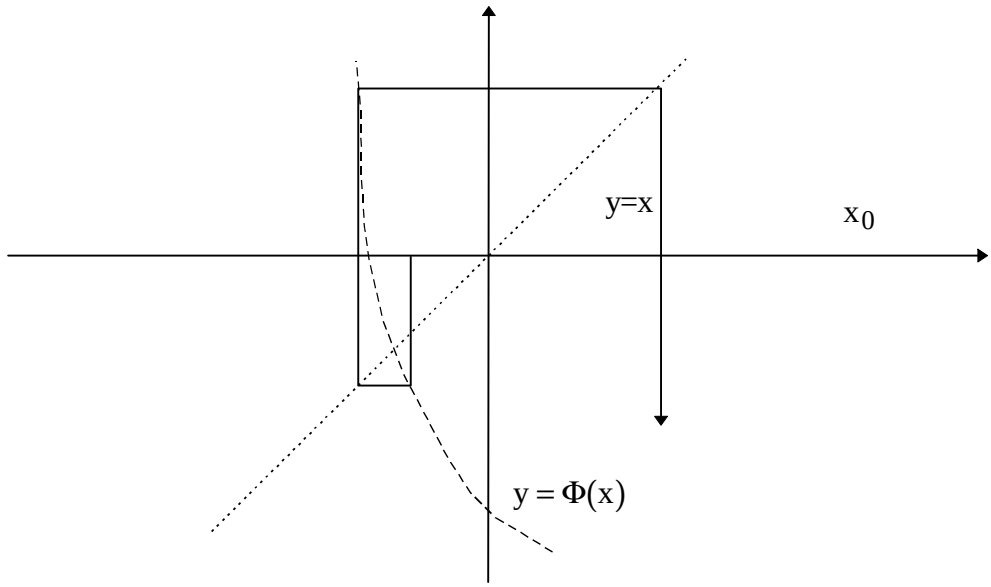
עוד נקודה חשובה מתיחסת לאלגוריתם GS. מהמשוואה $X_{k+1} = [D+L]^{-1}[P - UX_k]$ נדמה כי נדרש היפוך של מטריצה בעייתית בכל איטרציה. בפועל אין זה כך כלל כיוון שעדכון הפתרון לא נעשה באופן וקטורי כמקשה אחת, אלא באופן סידרתי על איברי וקטור הנעלמים. חישוב תוצאת האיטרציה הבאה נתון כפתרון של המשוואה - $[D+L]X_{k+1} = [P - UX_k]$. תחילה מחושב $X_{k+1}(1)$ כשכל שנדרש הוא היפוך סקלר. לשם חישוב $X_{k+1}(2)$ יש להשתמש ב- $X_{k+1}(1)$, ושוב יש היפוך של סקלר בלבד. היתרון החישובי נובע מהיות המטריצה אתה נדרש להפוך מטריצה משולשת תחתונה.

מה באשר להתכנסות אלגוריתמים מבוססי עקרון ההחלפה? אנו ראינו בדוגמה 11.5 כי בחירה מסויימת של הפרדה נותנת התכנסות בעוד שבחירה אחרת מביאה התבדרות. מהו אם כך הכלל - כיצד נכוון להפרדה אשר תניב התכנסות? באופן אינטואיטיבי ניתן לומר כי בהפרדה המוצעת $R(X) = X - \Phi(X) = 0$, רצוי שיאותרו המרכיבים הדומיננטיים של $R(X)$ והם אשר יבודדו לתת את ה- X בביטוי. לדוגמה - עבור אלגוריתם Jacoby התכנסות תהיה מהירה יותר אם איברי האלכסון של R גדולים משמעותית מיתר איברי המטריצה.

על מנת להבין (תחילה אינטואיטיבית) את תכונות ההתכנסות של אלגוריתם ההחלפה, נציע דרך גרפית לתיאור האלגוריתם, ונשתמש בבעיות חד-מימדיות לצורך זה. המשוואה $x = \Phi(x)$ פירושה כי אנו מחפשים חיתוך של שתי פונקציות - x ו- $\Phi(x)$. ניתן לצייר במישור את שתי פונקציות אלו ולהתבונן בחיתוכן. קל לראות כי האלגוריתם $x_{k+1} = \Phi(x_k)$ מתואר על גרפים אלו במסלול ספירלי, כמתואר בציור 11.8 ו- 11.9 (פעם עבור התכנסות ופעם עבור התבדרות).



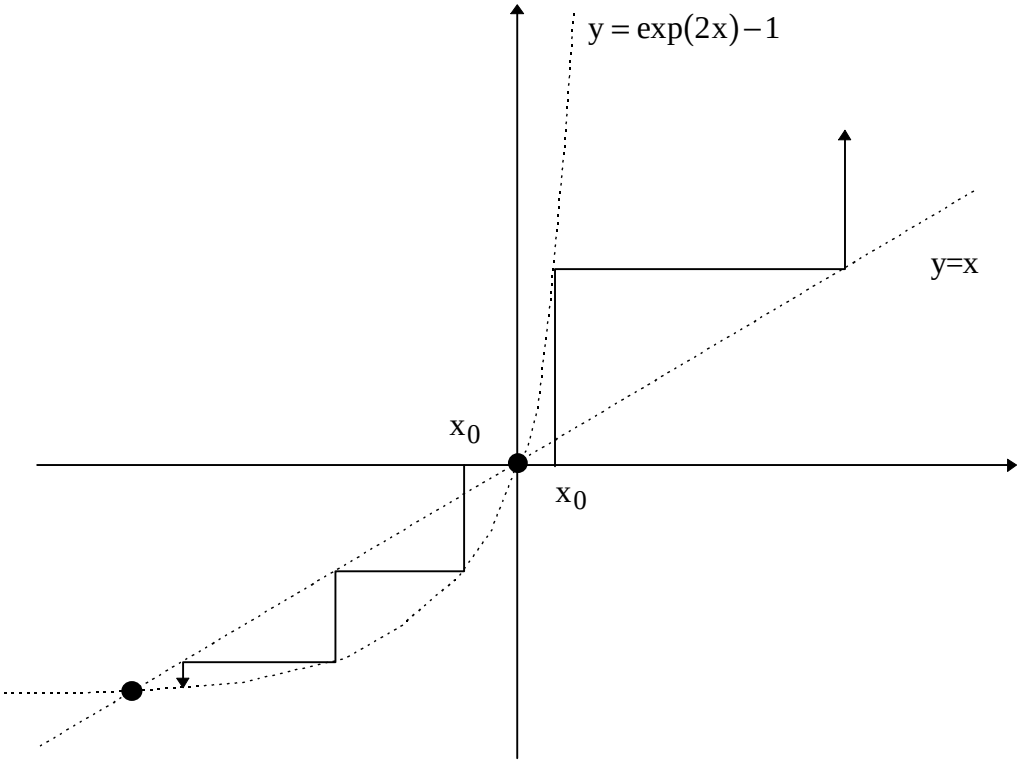
ציור 11.8 - תיאור גרפי של שיטת ההחלפה למקרה חד-מימדי



ציור 11.9 - תיאור גראפי של שיטת ההחלפה למקרה חד מימדי - התבדרות

דוגמה 11.6 : נחזור לדוגמה 11.5, אך הפעם עם תיאור גרפי של המשוואות :

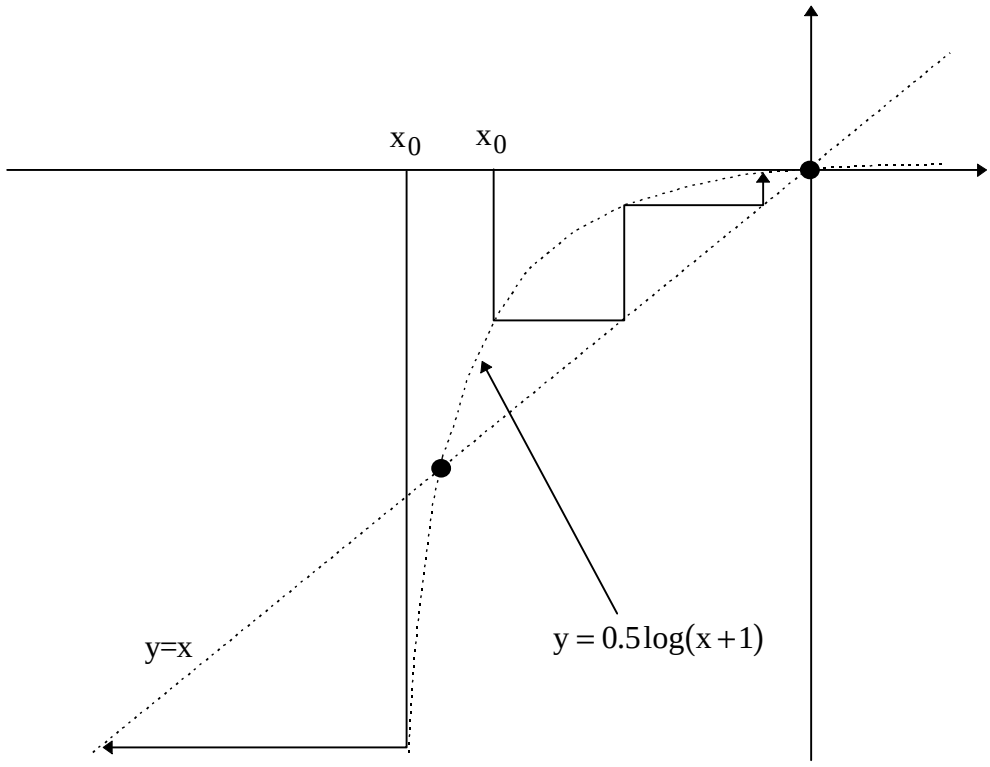
$$(1)\ \ x_{k+1} = \exp\{2x_k\} - 1 \ ; \quad (2)\ \ x_{k+1} = \frac{1}{2}\log\{x_k + 1\}$$



ציור 11.10 - משוואה (1) -תיאור גרפי של התפתחות האלגוריתם לשתי נקודות איתחול אפשריות

ציור 11.10 מתאר את משוואת האלגוריתם (1). שתי הפונקציות נחתכות ב- $x^* = 0,-0.8$. עבור נקודת איתחול x_0 חיובית **כלשהי** ברור כי האלגוריתם מתבדר. באותה מידה, עבור איתחול שלילי כלשהו חלה התכנסות לנקודה $x^* = -0.8$.

ציור 11.11 מתאר את משוואת האלגוריתם (2). הפונקציה $0.5\log(x+1)$ הינה תמונת ראי של הפונקציה $\exp(2x)+1$ סביב הציר האלכסוני y=x, ולכן, שתי הפונקציות נחתכות באותן נקודות- $x^* = 0,-0.8$. עבור נקודת איתחול x_0 חיובית **כלשהי** ברור כי האלגוריתם מתכנס לאפס. באותה מידה תחול התכנסות אם האיתחול גדול מ- (0.8-) לנקודה אפס. עבור ערך איתחול נמוך מזה חלה התבדרות.



ציור 11.11 - משוואה (2) -תיאור גרפי של התפתחות האלגוריתם לשתי נקודות איתחול אפשריות

משני גרפים אלו קיבלנו תוצאות המשלימות זו את זו. אלגוריתם אחד מתכנס לנקודת חיתוך אחת- $x^* = -0.8$, והאחר לנקודה האחרת - $x^* = 0$. תחומי ההתכנסות שונים, ואיחודם מכסה את כל הציר. תופעה שניתנת לזיהוי מגרפים אלה היא שבנקודות חיתוך בהן שיפוע הפונקציה $\Phi(x)$ נמוך חלה התכנסות. המשפט הבא מתאר את הדרישה הזו באופן מדויק יותר.

משפט 11.2: בהינתן מערכת משוואות R(X)=0 הניתנת להפרדה כ - $R(X) = X - \Phi(X) = 0$, ובהנחה שהפונקציה $\Phi(X)$ גזירה ברציפות, אם מתקיים כי $\|\nabla\Phi(X^*)\| < 1$, אזי קיימת סביבה בקרבת X^* בה האלגוריתם $X_{k+1} = \Phi(X_k)$ יתכנס.

הוכחה: תוך שימוש בנוסחת העדכון ובפיתוח לטור טיילור מתקבל:

$$X_{k+1} = \Phi(X_k) = \Phi(X^*) + \nabla\Phi(X^*)(X_k - X^*) + o\{\|X_k - X^*\|\}$$

שימוש בקשר $X^* = \Phi(X^*)$ נותן:

$$\|X_{k+1} - X^*\| = \|\nabla\Phi(X^*)\| \cdot \|X_k - X^*\| + o\{\|X_k - X^*\|\}$$

ומשוואה זו משקפת התכנסות רק אם $\|\nabla\Phi(X^*)\| < 1$. ■

קיים משפט (דניס ומור) אשר טוען כי מכל שיטות ההפרדה האפשריות, זו בעלת הגודל $\|\nabla\Phi(X^*)\|$ המינימלי שואפת אסימפטוטית לשיטת Newton-Raphson. קביעה זו פירושה כי אלגוריתם Newton-Raphson מייצג את האלגוריתם האופטימלי מבין חבורת האלגוריתמים המושתתים על הפרדה.

לפני שנסיים דיון זה נציג רעיון אשר נועד להתגבר על מינימות לוקליות באלגוריתם Newton-Raphson. שיטה זו קרויה מסלול דווידסקו או שיטת הרצף. רעיון זה משרת גם לשיטת אחרות, אך הועלה בראשונה לצורך שיפור אלגוריתמי חיפוש על ישר. בהינתן מערכת משוואות R(X)=0, מוצע להגדיר בעיה חדשה שהינה אוסף של רצף בעיות $H(X,\theta) = 0$, כך שלכל ערך של θ יש בידינו מערכת משוואות. נדרוש מ- $H(X,\theta) = 0$ שיקיימו שתי תכונות:

1. $H(X,0) = R(X)$ - כך שפתרון הבעיה H(X,0)=0 שקול לפתרון הבעיה שעל הפרק.
2. $H(X_0,1) = 0$ - כלומר - הבעיה H(X,1) בעלת פתרון ידוע מראש X_0.

הרעיון להתחיל ב- X_0 כפתרון של H(X,1), לסחוף אט-אט את הפרמטר θ לכיוון האפס כשכל הזמן אנו משתמשים בתוצאה הקודמת לשם איתור המינימה בערך של θ החדש. הרעיון המרכזי אם כן הוא החלפת בעיה אחת עם איתחול לא ידוע ברצף בעיות עם איתחול ידוע. הצעת בעיה $H(X,\theta) = 0$ מתאימה יכולה להיות הבאה:

$$H(X,\theta) = R(X) - \theta R(X_0)$$

וזו תבנית המקיימת את כל הדרישות. אוסף הפתרונות כפונקציה של הפרמטר - $X(\theta)$ - נקרא מסלול דוידסקו.

11.6 סיכום

בפרק זה הצגנו את תבנית ה- LS, מקורותיה והדרכים לפתרו בעיות אופטימיזציה הנתונות בתבנית זו. ראינו תחילה את המקרה הליניארי, וממנו עברנו למקרה הכללי אותו פותר אלגוריתם Gauss-Newton. הוצגו מבנים ייחודיים של ה- LS אשר מאפשרים הקלה נוספת בפתרון בעיית המינימיזציה הנתונה. בסיומו של פרק זה סטינו מתחום האופטימיזציה לדיון בדרכים לפתרו מערכות של משוואות. כפי שראינו יש קשר הדוק בין התחומים הללו הן בתבניות המתקבלות, והן בשיטות לפתרונן.

פרק 12 - בעיות תכנות ליניארי ופתרונן

12.1 מבנה קלאסי של בעיות תכנות ליניארי

בעיית התכנות הליניארי הקלאסית נתונה כבעיית מינימיזציה של צירוף ליניארי של מספר נעלמים, כשעל נעלמים אלה להיות חיוביים, ונתונות מספר משוואות אילוץ ליניאריות אף הן הנכפות על הפתרון. השם תכנות ליניארי בא מההסתכלות על בעיות אלה כתוכניות מחשב בהן ישנה פונקצית יעד ואילוצים ליניאריים. מבנה הבעיה ברישום פורמלי נראה כך:

$$\underset{\{x_k\}_{k=1}^{n}}{\text{minim.}} : \sum_{k=1}^{n} c_k x_k \qquad \text{subj. to:} \quad \begin{aligned} &\sum_{k=1}^{n} a_{1,k} x_k = b_1 \\ &\vdots \\ &\sum_{k=1}^{n} a_{m,k} x_k = b_m \\ &x_1 \geq 0, \cdots, x_n \geq 0 \end{aligned}$$

או ברישום יותר קומפקטי:

$$\underset{X}{\text{minim.}} : C^T X \qquad \text{subj. to:} \quad \mathbf{A}X = B \quad \& \quad X \geq 0$$

במבנה הנ"ל אנו נניח תמיד כי $b_k \geq 0$ - ניתן להשיג זאת ע"י הכפלת כל משוואת אילוץ שאינה מקיימת זאת ב- (-1). בבעיית תכנות ליניארי אופיינית יתקיים - $m<n$. מצב בו יש n אילוצים (בלתי תלויים ליניארית) פירושו שהאילוצים לבדם קובעים את הפתרון. מצב בו יש יותר מ- n אילוצים פירושו שיתכן שחלקם סותרים זה את זה, או שחלקם חוזרים על צירופים ליניאריים של אילוצים אחרים, ולכן הם מיותרים - ובמצב זה יתקבל שוב כי יש פתרון אחד ויחיד לאילוצים ואין כאן בעיית מינימיזציה כלל. אנו נניח כי דרגת המטריצה A היא m בדיוק - שפירושו שכל האילוצים הנתונים הינם בלתי תלויים, ולכן אין אילוץ אשר סותר את האחרים, או אילוץ שמופיע בסמוי באחרים.

למעשה, בעיית התכנות הליניארי הינה כניסתנו הראשונה לתחום הטיפול בבעיות עם אילוצים. יכולנו להציג את נושא הבעיות עם האילוצים בכללותו, ולטפל בבעיה זו כמקרה פרטי, אך גישה זו כלל לא כדאית. לקבוצת הבעיות אשר ניתנות לתיאור בתבנית הנ"ל מוקדש טיפול נפרד בשל סיבה עיקרית אחת - פשטותה היחסית של הבעיה מניבה יכולת פתרון "קלה במיוחד", והכלי אשר מבצע את מלאכת הפתרון הוא אלגוריתם ה- Simplex, אשר מהווה את עיקר עיסוקו של פרק זה. יש להבחין בין אלגוריתם זה ובין אלגוריתם ה- Simplex אשר תואר בפרק 6 - אין כל קשר ביניהם, למעט העובדה ששניהם עושים שימוש בפוליגון רב-מימדי הקרוי Simplex.

העניין במשפחת הבעיות של התכנות הליניארי בא מכיוונים שונים. סוגיות רבות בכלכלה, הקצאת משאבים, ניהול, ובחשמל - פתרון של מעגלים ליניאריים, תכנון מערכות ליניאריות - כל אלה ניתנים לייצוג כבעיות במבנה הנ"ל.

אחת הסוגיות הראשונות בהן יש לטפל מתייחסת למבנה הבעיה. אלגוריתם ה- Simplex אותו נכיר בהמשך מניח את המבנה שתואר לעיל בבואו לפתור את הבעיה. 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-שיוויון - נתונה הבעיה הבאה:

$$\underset{\{x_k\}_{k=1}^{n}}{\text{minim.}}: \quad \sum_{k=1}^{n} c_k x_k \qquad \text{subj. to:} \quad \begin{aligned} &\sum_{k=1}^{n} a_{1,k} x_k \le b_1 \\ &\vdots \\ &\sum_{k=1}^{n} a_{m,k} x_k \le b_m \\ &x_1 \ge 0, \cdots, x_n \ge 0 \end{aligned}$$

במקרה כזה נציע את הבעיה השקולה הבאה:

$$\underset{\{x_k\}_{k=1}^{n}}{\text{minim.}}: \quad \sum_{k=1}^{n} c_k x_k \qquad \text{subj. to:} \quad \begin{aligned} &\sum_{k=1}^{n} a_{1,k} x_k + y_1 = b_1 \\ &\vdots \\ &\sum_{k=1}^{n} a_{m,k} x_k + y_m = b_m \\ &x_1 \ge 0, \cdots, x_n \ge 0, y_1 \ge 0, \cdots, y_m \ge 0 \end{aligned}$$

כלומר, המרנו את הבעיה הקודמת בת n הנעלמים בבעיה חדשה בת n+m נעלמים, עם m משתני דמה. ברור כי עבור מצב בו חלק מהאילוצים הוא עם שיוויון וחלקו האחר עם אי-שיוויון, נוסיף משתני דמה רק לאילוצי האי-שיוויון.

2. אילוצי אי-שיוויון - נתונה הבעיה הבאה:

$$\underset{\{x_k\}_{k=1}^{n}}{\text{minim.}}: \quad \sum_{k=1}^{n} c_k x_k \qquad \text{subj. to:} \quad \begin{aligned} &\sum_{k=1}^{n} a_{1,k} x_k \ge b_1 \\ &\vdots \\ &\sum_{k=1}^{n} a_{m,k} x_k \ge b_m \\ &x_1 \ge 0, \cdots, x_n \ge 0 \end{aligned}$$

במקרה כזה נציע את הבעיה השקולה הבאה:

$$\underset{\{x_k\}_{k=1}^{n}}{\text{minim.}}: \quad \sum_{k=1}^{n} c_k x_k \qquad \text{subj. to:} \quad \sum_{k=1}^{n} a_{1,k} x_k - y_1 = b_1$$

$$\vdots$$

$$\sum_{k=1}^{n} a_{m,k} x_k - y_m = b_m$$

$$x_1 \geq 0, \cdots, x_n \geq 0, y_1 \geq 0, \cdots, y_m \geq 0$$

כלומר, גם הפעם המרנו את הבעיה הקודמת בת n הנעלמים בבעיה חדשה בת n+m נעלמים, עם m משתני דמה. וכמקודם, ברור כי עבור מצב בו חלק מהאילוצים הוא עם שיויון וחלקו האחר עם אי-שיויון, נוסיף משתני דמה רק לאילוצי האי-שיויון.

3. נתונה הבעיה הקלאסית למעט העובדה שלא נדרש $x_k \geq 0$. 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 - $u \geq 0, v \geq 0$. כעת הבעיה בעלת מבנה קלאסי, אך ממימד של n+1.

האלטרנטיבה 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.

4. 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 $y_k = d - x_k$.

5. עבור הבעיה הבאה:

$$\underset{\{x_k\}_{k=1}^{n}}{\text{minim.}}: \quad \sum_{k=1}^{n} c_k |x_k| \qquad \text{subj. to:} \quad \sum_{k=1}^{n} a_{1,k} x_k = b_1$$

$$\vdots$$

$$\sum_{k=1}^{n} a_{m,k} x_k = b_m$$

$$x_1 \geq 0, \cdots, x_n \geq 0$$

ניתן פשוט להשמיט את סימני הערך המוחלט כיוון שממילא נדרש כי כל ערכי x_k יהיו אי-שליליים.

6. נניח כי הבעיה הנתונה זהה במבנה לבעיה הקלאסית למעט גורם המינימיזציה אשר נתון בערכו המוחלט:

$$\underset{X}{\text{minim.}}: \ C^T|X| \qquad \text{subj. to:} \quad \mathbf{A}X = B$$

במקרה זה נציע את ההצבה הבאה - X=U-V. עבור הצבה זו נקבל כי האילוץ יהיה AU-AV=B, ונדרוש בנוסף $U,V \geq 0$. פונקציית המחיר תומר להיות $C^TU + C^TV$, והבעיה הכוללת תהיה:

$$\underset{X}{\text{minim.}} \ : \ [C^T, C^T]\begin{bmatrix} U \\ V \end{bmatrix} \qquad \text{subj. to:} \quad [A,-A]\begin{bmatrix} U \\ V \end{bmatrix} = B \quad \& \quad U,V \geq 0$$

נושא האילוצים בבעיה הנ"ל ברור - נושא פונקציית המחיר תמוה במקצת ומחייב התעמקות מסויימת. מסתבר (ואנו נראה זאת בהמשך פרק זה) שבעיות תכנות ליניארי מציעות פתרון עם מספר מצומצם ככל האפשר של איברים השונים מאפס (אנו נקרה לתכונה זו - פתרון בסיסי). השלכה ישירה של תכונה זו היא שבפתרון הבעיה הנ"ל, אם איבר כלשהו מ- X צריך להיות חיובי, האיבר המתאים לו ב- U יקבל את הערך הדרוש, והאיבר המתאים ב- V יהיה אפס. באופן דומה, עבור איבר שלילי, האיבר ב- V יקבל את הערך הדרוש, והאיבר המתאים ב- U יהיה אפס. אם איבר כלשהו ב- X צריך להיות אפס, כך גם יהיו שני האיברים המתאימים ב- U ו- V. לא יקרה מצב בו שני איברים תואמים ב- U ו- V יהיו שונים מאפס - זה סותר את התכונה הנ"ל. לכן, **<u>חיבור</u>** של U ו-V שקול לפעולת ערך מוחלט, כנדרש.

ניתן בצורה דומה להציע עד מגוון וריאציות לבעיות אשר ניתנות להמרה לבעיות תכנות ליניארי קלאסיות. כעת נפנה לדיון אשר יתחיל את בניית דרך הפתרון לבעיות אלו.

<u>12.2 המשפט היסודי של התכנות הליניארי</u>

נגדיר תחילה את המושגים "פתרון בסיסי" ו- "פתרון בסיסי חוקי". נסתכל על מערכת המשוואות AX=B. למערכת זו מספר אינסופי של פתרונות, אשר מגדירים תת-מרחב.

הגדרה 12.1: "פתרון בסיסי" הוא פתרון של המשוואה B=AX בו לכל היותר m איברים שונים מאפס. "פתרון בסיסי חוקי" יהיה פתרון בסיסי אשר כל ערכיו השונים מאפס גדולים מאפס.

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 $\{\forall\ m+1 \leq k \leq n \quad x_k \equiv 0\}$ נקבל את מערכת המשוואות הבאה:

$$AX = \begin{bmatrix} a_1 & \cdots & a_m & a_{m+1} & & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \\ x_m \\ x_{m+1} \\ \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 & \cdots & a_m \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = B$$

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 זו אין 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 12.1: (המשפט היסודי של התכנות הליניארי). בהינתן בעיית תכנות ליניארי בעלת המבנה הקלאסי:

$$\underset{X}{\text{minim.}} : \; C^T X \qquad \text{subj. to:} \quad \mathbf{A}X = B \;\; \& \;\; X \geq 0$$

מתקבל כי:

1. אם קיים פתרון חוקי (כלומר וקטור X אשר עונה לאילוצים - עובדה המעידה שקבוצת האילוצים לא ריקה), אזי קיים בהכרח גם פתרון בסיסי חוקי.
2. אם קיים פתרון חוקי אופטימלי (כלומר פתרון חוקי אשר משיג את המינימום לפונקציה $C^T X$) אזי בהכרח גם קיים פתרון בסיסי חוקי אופטימלי.

הוכחה: נוכיח תחילה את הטענה הראשונה. נניח כי נתון בידנו פתרון חוקי X בו יש p איברים שונים מאפס. נניח ללא אובדן הכלליות כי אלה הם p איבריו הראשונים של X. אזי, אם $p \le m$ פירוש הדבר שבידנו פתרון בסיסי והטענה נכונה.

נניח אם כן כי p>m. לכן, p העמודות המתייחסות לאיברים אלה (אלה הן p העמודות הראשונות לפי הנחתנו קודם) חייבות להיות תלויות ליניארית. כלומר, קיימים וקטור Y כך ש:

$$y_1a_1 + y_2a_2 + \cdots + y_pa_p = Y^T \cdot \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} = 0$$

ומצד שני מתקיים:

$$x_1a_1 + x_2a_2 + \cdots + x_pa_p = X_{[1:p]}^T \cdot \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} = B$$

ע"י הכפלת המשוואה הראשונה ב- ε והחסרתה מהשניה מתקבל:

$$(x_1 - \varepsilon y_1)a_1 + (x_2 - \varepsilon y_2)a_2 + \cdots + (x_p - \varepsilon y_p)a_p = \left(X_{[1:p]} - \varepsilon Y\right)^T \cdot \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} = B$$

נזכור כי X הוא פתרון חוקי ולכן כל איבריו השונים מאפס חיוביים. הוקטור 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 x_k / y_k היא הנמוכה ביותר. עבור $\varepsilon_0 = x_k / y_k$ 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 X הינו פתרון חוקי אופטימלי, ולו p איברים שונים מאפס. אם $p \le m$ פירוש הדבר שבידנו פתרון בסיסי אופטימלי והטענה נכונה. נניח אם כן כי p>m. כמקודם מתקבל כי עבור ε קטן מאוד (חיובי ושלילי) מתקבל כי $\left(X_{[1:p]} - \varepsilon Y\right)$ הינו פתרון חוקי. ערך הפונקציה לפתרונות אלו הוא:

$$C^T\left(X_{[1:p]}-\varepsilon Y\right)=C^TX-\varepsilon C^TY$$

אם הגורם C^TY שונה מאפס תתקבל סתירה - נראה זאת - נניח כי C^TY גדול מאפס. אזי יתקבל כי עבור ערך חיובי של ε מתקבל ערך נמוך יותר לפונקציה, ולכן X אינו אופטימלי. לכן בהכרח נדרוש כי $C^TY=0$. 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 אופטימלי. נמשיך בדרך זו של איפוס איברים עד לקבלת p=m, והפתרון החוקי אשר יתקבל יהיה בסיסי ואופטימלי. באופן זה הוכחנו את הטענה השניה 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 (כבר ראינו שלמספר משתנים נמוך יכולים לקבל כמות אינסופית בפועל של פתרונות בסיסיים אפשריים) - אנו נראה גישה יעילה הרבה יותר אשר דורשת בממוצע כ-m ניסיונות לשם איתור האופטימום - שיטת ה- Simplex.

המשפט הנ"ל אינו צריך להפתיענו. את עיקריו כבר ראינו בפרק 3, שם נדונו תכונות של פונקציות קמורות מעל תחום הגדרה קמור. כפי שראה בהמשך, בעיית התכנות הליניארית היא כוללת פונקציה שהיא קמורה וקעורה בו-זמנית (בהיותה ליניארית), ותחום האילוצים הוא קבוצה קמורה. לכן, הפתרון צריך להתקבל בקצוות. בהמשך נראה שקצוות האילוצים אינם אלא פתרונות בסיסיים.

12.3 הקשר בין אילוצי בעיית תכנות ליניארי ותחום קמור

מסתבר כי בעיית התכנות הליניארי ניתנת לתיאור נוח כבעיה עם אילוץ קמור, וככזו ניתן להבין את המשפט היסודי שהוצג קודם בגישה גיאומטרית. תחילה נשים לב לכך שקבוצת ה- X המקיימת את האילוצים $AX=B\ ,X\geq 0$ היא קבוצה קמורה. אם שתי נקודות X ו- Y שייכות לקבוצה זו, קל לראות כי כל צירוף קמור שלהן שייך לקבוצה.

הגדרה 12.2: נקודה X בקבוצה קמורה C תיקרא נקודת קיצון אם לא קיימות שתי נקודות ב- C אשר X מצויה על הישר אשר נמתח ביניהן. לדוגמה, עבור משולש, נקודות הקיצון הן שלושת פינותיו. באופן דומה, במצולע רב צלעות, כל פינה הינה נקודת קיצון.

משפט 12.2: יהי K המצולע הרב מימדי הקמור המוגדר ע"י - $AX = B \ , X \geq 0$. הוקטור X הוא נקודת קיצון של מצולע רב-מימדי זה אם ורק אם הוא פתרון בסיסי חוקי של המערכת $AX = B \ , X \geq 0$.

הוכחה: נניח כי 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 X. נניח כי קיימות שתי נקודות כאלה - Y1, Y2. אזי - $\exists 0 < \alpha < 1 \quad X = \alpha Y_1 + (1-\alpha) Y_2$. מכיוון שאיבריו של X הם אפס בתחום , ומכיוון שכל איברי צריכים להיות חיוביים, בהכרח נקבל כי גם איברי בתחום הם אפסים. כיוון ש- m הערכים של כל פתרון אפשרי הם למעשה מקדמי הצירוף אשר ייתן:

ועמודות אלה בלתי תלויות ליניארית, נקבל כי בהכרח , ולכן לא קיימות צמד נקודות אשר יתנו את X. בכך הראינו כיוון הוכחה אחד. נניח כי X היא נקודת קיצון של K, ונניח כי נקודה זו כוללת ערכים שונים מאפס. נראה כי פירוש הדבר הוא שהיא אינה נקודת קיצון. כבר ראינו קודם כי ניתן במקרה זה להציע פתרונות חדשים מהצורה:

כאשר הם מקדמי הצירוף הליניארי שנובע מהיותן של עמודות אלו תלויות. עבור קטן דיו שני פתרונות אלו חוקיים, וצירופם נותן את X. לכן לא יתכן כי X היא נקודת קיצון. לכן לא יתכן כי X כנקודת קיצון תכלול יותר מ- m איברים שונים מאפס. ■

מהמשפט הנ"ל נובע כי לקבוצה K יש לכל היותר נקודת קיצון (קבוצת החוקיים שבהם קטנה יותר).

משפט 12.2: יהי K המצולע הרב מימדי הקמור המוגדר ע"י - $AX = B \ , X \geq 0$. הוקטור X הוא נקודת קיצון של מצולע רב-מימדי זה אם ורק אם הוא פתרון בסיסי חוקי של המערכת $AX = B \ , X \geq 0$.

הוכחה: נניח כי 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 X. נניח כי קיימות שתי נקודות כאלה - $Y1, Y2$. אזי - $\exists 0 < \alpha < 1 \quad X = \alpha Y_1 + (1-\alpha) Y_2$. מכיוון שאיבריו של X הם אפס בתחום $m+1 \div n$, ומכיוון שכל איברי $Y1, Y2$ צריכים להיות חיוביים, בהכרח נקבל כי גם איברי $Y1, Y2$ בתחום $m+1 \div n$ הם אפסים. כיוון ש- m הערכים של כל פתרון אפשרי הם למעשה מקדמי הצירוף אשר ייתן:

$$\begin{bmatrix} a_1 & \cdots & a_m \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = B$$

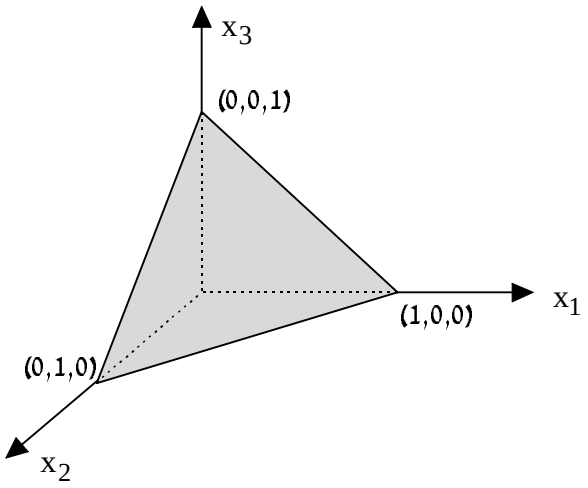
ועמודות אלה בלתי תלויות ליניארית, נקבל כי בהכרח $Y1 = Y2 = X$, ולכן לא קיימות צמד נקודות אשר יתנו את X. בכך הראינו כיוון הוכחה אחד. נניח כי X היא נקודת קיצון של K, ונניח כי נקודה זו כוללת $p > m$ ערכים שונים מאפס. נראה כי פירוש הדבר הוא שהיא אינה נקודת קיצון. כבר ראינו קודם כי ניתן במקרה זה להציע פתרונות חדשים מהצורה:

$$(x_1 - \varepsilon y_1)a_1 + (x_2 - \varepsilon y_2)a_2 + \cdots + (x_p - \varepsilon y_p)a_p = B$$
$$(x_1 + \varepsilon y_1)a_1 + (x_2 + \varepsilon y_2)a_2 + \cdots + (x_p + \varepsilon y_p)a_p = B$$

כאשר y_k הם מקדמי הצירוף הליניארי שנובע מהיותן של עמודות אלו תלויות. עבור ε קטן דיו שני פתרונות אלו חוקיים, וצירופם נותן את X. לכן לא יתכן כי X היא נקודת קיצון. לכן לא יתכן כי X כנקודת קיצון תכלול יותר מ- m איברים שונים מאפס. ■

מהמשפט הנ"ל נובע כי לקבוצה K יש לכל היותר $\#_s$ נקודת קיצון (קבוצת החוקיים שבהם קטנה יותר).

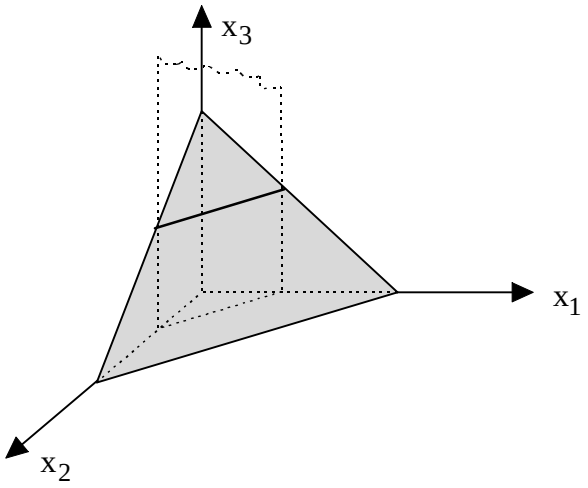
דוגמה 12.1: עבור מערכת האילוצים - $\begin{array}{c} x_1 + x_2 + x_3 = 1 \\ x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0 \end{array}$ קבוצת הפתרונות נראית כך:



ציור 12.1 - תיאור קבוצת הפתרונות

וכצפוי, יש שלוש נקודת קיצון ל- K, המתייחסות לשלוש נקודת פתרון בסיסיות חוקיות.

דוגמה 12.2: נניח כי מערכת האילוצים כוללת בנוסף את המשוואה $2x_1 + 2x_2 = 1$. הקבוצה K מתוארת בציור הבא:



ציור 12.2 - תיאור קבוצת הפתרונות

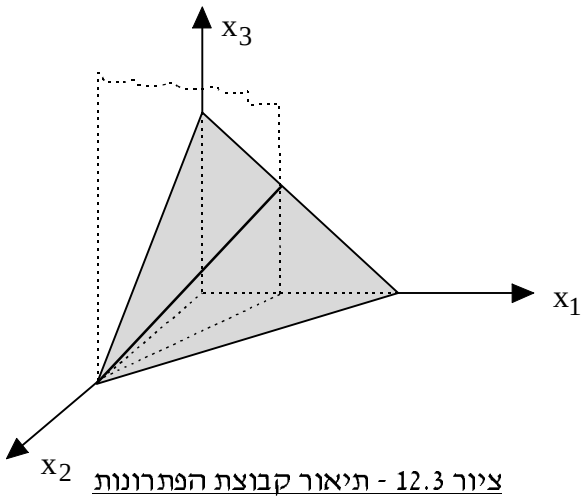
הפתרונות הבסיסיים הפעם הם:

$$\begin{bmatrix}1&1&1\\2&2&0\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{array}{l}\begin{bmatrix}1&1\\2&0\end{bmatrix}\begin{bmatrix}x_1\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_1\\x_3\end{bmatrix}=\begin{bmatrix}0.5\\0.5\end{bmatrix}\\ \begin{bmatrix}1&1\\2&0\end{bmatrix}\begin{bmatrix}x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_2\\x_3\end{bmatrix}=\begin{bmatrix}0.5\\0.5\end{bmatrix}\\ \begin{bmatrix}1&1\\2&2\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_1\\x_2\end{bmatrix}=?\end{array}$$

ובפתרון 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.

דוגמה 12.3: נניח כי מערכת האילוצים כוללת במקום את המשוואה $2x_1+x_2=1$. הקבוצה K מתוארת בציור הבא:



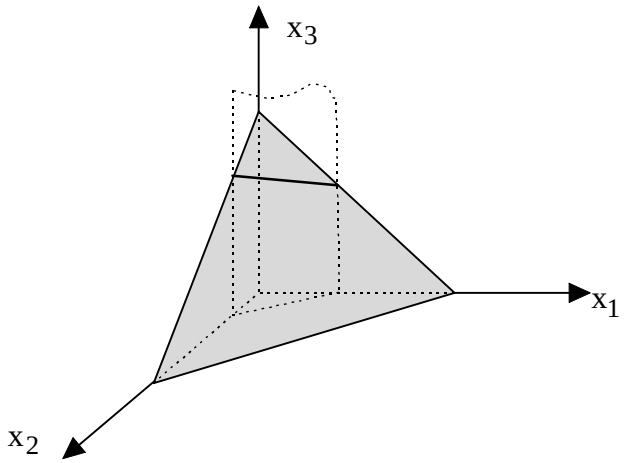
ציור 12.3 - תיאור קבוצת הפתרונות

הפתרונות הבסיסיים הפעם הם:

$$\begin{bmatrix}1&1&1\\2&1&0\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{array}{l}\begin{bmatrix}1&1\\2&0\end{bmatrix}\begin{bmatrix}x_1\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_1\\x_3\end{bmatrix}=\begin{bmatrix}0.5\\0.5\end{bmatrix}\\ \begin{bmatrix}1&1\\1&0\end{bmatrix}\begin{bmatrix}x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\0\end{bmatrix}\\ \begin{bmatrix}1&1\\2&1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}0\\1\end{bmatrix}\end{array}$$

והפתרון הבסיסי השלישי מתלכד עם השני, כך שקיבלנו שני פתרונות בסיסיים חוקיים, וכן שתי נקודות קיצון ל- K.

דוגמה 12.4: נניח כי מערכת האילוצים כוללת במקום את המשוואה $2x_1+3x_2=1$. הקבוצה K מתוארת בציור הבא:



ציור 12.4 - תיאור קבוצת הפתרונות

הפתרונות הבסיסיים הפעם הם:

$$\begin{bmatrix}1&1&1\\2&3&0\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{array}{l}\begin{bmatrix}1&1\\2&0\end{bmatrix}\begin{bmatrix}x_1\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_1\\x_3\end{bmatrix}=\begin{bmatrix}0.5\\0.5\end{bmatrix}\\ \begin{bmatrix}1&1\\3&0\end{bmatrix}\begin{bmatrix}x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_2\\x_3\end{bmatrix}=\begin{bmatrix}0.33\\0.67\end{bmatrix}\\ \begin{bmatrix}1&1\\2&3\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}\Rightarrow\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}2\\-1\end{bmatrix}\end{array}$$

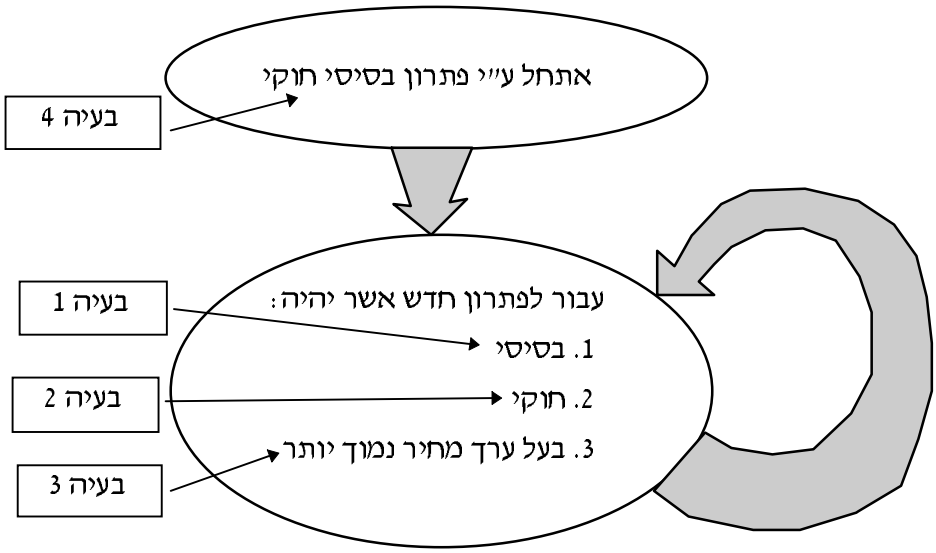
והפעם, הפתרון הבסיסי השלישי אינו חוקי, כך שקיבלנו שני פתרונות בסיסיים חוקיים, וכן שתי נקודות קיצון ל- K.

לסיכום עד כה, ראינו כי פתרון בעיית תכנות ליניארית כרוך באיתור פתרון בסיסי וחוקי אשר יגשים את המינימום של הפונקציה. במונחים גאומטריים, נקודת פתרון זו הינה נקודת קצה של קבוצת הפתרונות החוקיים. נקודה אחרונה היא שישנו מספר סופי (אך יכול להיות גדול) של נקודות כאלה - לכן, אין צורך, כבאלגוריתמים קודמים, לחפש מעל הרצף, ודי יהיה לסרוק בנקודות מסויימות את מרחב הפתרונות החוקיים - עובדה הצפויה להקל את פתרון הבעיה.

12.4 אלגוריתם ה- Simplex

אלגוריתם ה- Simplex הוצע לראשונה ע"י Dantzig בשנת 1947, ומאז קמו מאות חוקים שהציעו שיפורים ווריאציות שונות על אלגוריתם זה. הישגו של Dantzig עצום על רקע העובדה שב-1947 עדיין לא ידעו כיצד פותרים אנליטית בעיות עם אילוצי אי-שיוויון. ה- Simplex היא טכניקה מוגדרת היטב ולא כל כך מסובכת אשר מובילה לפתרון האופטימלי (הגלובלי!!!) של הבעיה.

אלגוריתם ה- Simplex עושה שימוש בתכונה החשובה ביותר שנחשפה עד כה על בעיות תכנות ליניאריות - העובדה כי הפתרונות מוגשמים בקצוות (בפתרונות הבסיסיים). הרעיון המרכזי של האלגוריתם הוא הבא - התחל בפתרון בסיסי חוקי אחד כלשהו (אנו נראה כיצד מקבלים איתחול כזה בסעיף הבא), ועבור לפתרון בסיסי אחר סמוך (שקו מתוח ביניהם בקבוצה K) באופן איטרטיבי כך שיובטח כי הפתרון החדש חוקי, והוא מקטין את ערך פונקצית יעד המינימיזציה. באופן זה נמשיך עד שנגיע למינימום. ציור 12.5 מתאר את האלגוריתם כסכימת בלוקים, ומציין את הנקודות הבעיתיות בו מבחינתנו.



ציור 12.5 - סכימת בלוקים של אלגוריתם ה- Simplex

השאלות שעלינו לשאול אם כך הן הבאות:

1. בהינתן פתרון בסיסי אחד - איך עוברים לפתרון בסיסי סמוך לו?
2. בהנחה כי אנו יודעים את התשובה ל- 1, כיצד נחליט (ללא חישוב מפורש) כי הפתרון החדש יהיה חוקי?
3. בהנחה כי אנו יודעים את התשובה ל- 1 ו- 2, כיצד נחליט (ללא חיפוש מלא) מיהו הוקטור אשר צפוי לתת ירידה הכי גדולה בערך הפונקציה?

4. וכמובן - שאלת האיתחול - כיצד נמצא פתרון התחילי בסיסי חוקי? (כאמור, שאלה זו תידון בסעיף הבא בנפרד, ואנו נניח לכל אורך הדיון בסעיף זה כי בידנו פתרון חוקי בסיסי כאיתחול).

לפני שנתחיל את תיאור האלגוריתם נעיר כי התפיסה הנ"ל מבטיחה התכנסות למינימום הגלובלי כיוון שהבעיה קמורה ותחום האילוצים קמור, ולפיכך המינימה הלוקלית שתימצא באלגוריתם זה חייבת להיות גם הגלובלית.

אלגוריתם ה- Simplex פועל ע"י בניית טבלה וביצוע מניפולציות עליה. בתיאור שיובא להלן, נראה כיצד נבנית טבלה זו, וכיצד המניפולציות המוצעות משיגות את יעדן.

בהתייחס לשאלה הראשונה, פתרון סמוך פירושו שמתוך m האיברים השונים מאפס אחד יאופס, ומתוך n-m המאופסים אחד יוכנס למשחק. הדרך לביצוע פעולה זו היא שימוש בציר - Pivot. נניח כי בידנו פתרון בסיסי חוקי הכולל את m האיברים הראשונים של X. נכתוב את הטבלה הבאה אשר מתייחסת לאילוצים:

$$\left[\begin{array}{cccccc|c} 1 & & 0 & a_{1,m+1} & \cdots & a_{1,n} & b_1 \\ & 1 & & a_{2,m+1} & \cdots & a_{2,n} & b_2 \\ & & \ddots & \vdots & & \vdots & \vdots \\ 0 & & 1 & a_{m,m+1} & \cdots & a_{m,n} & b_m \end{array}\right]$$

נשים לב לכך שהפתרון המוצע לפי טבלה זו הוא הוקטור B הנמצא בעמודה הימנית. טבלה זו מתייחסת למטריצה A ולוקטור B לאחר לכסון אוסף העמודות הראשון. באופן מטריצי נאמר את הדבר כך - במקור נתון כי AX=B כאשר A ניתנת לרישום בלוקי כ- $A=\left[A_1|A_2\right]$ כאשר A_1 כוללת את m העמודות הראשונות. באופן דומה, $X^T=\left[X_1^T|X_2^T\right]=\left[X_1^T|0^T\right]$, כיוון ש- X פתרון בסיסי. נוכל לכתוב:

$$AX=\left[A_1|A_2\right]\begin{bmatrix}X_1\\X_2\end{bmatrix}=B \;\Rightarrow\; \left[I\middle|A_1^{-1}A_2\right]\begin{bmatrix}X_1\\X_2\end{bmatrix}=\begin{bmatrix}X_1\\0\end{bmatrix}=A_1^{-1}B$$

וכך קיבלנו את הטבלה שתוארה. הדוגמה הבאה תמחיש את צעד זה של קנוניזציה של הטבלה, לפני תחילת האלגוריתם.

דוגמה 12.5 : נתונה בעיית LP עם אילוצי השיוויון המאופיינים ע"י הצמד B,A הבאים :

$$A=\begin{bmatrix}2 & -1 & -1 & -1 & -9 & 4\\ 0 & 1 & 3 & 5 & 17 & 4\\ 1 & 1 & 1 & 4 & 6 & 5\end{bmatrix} \qquad B=\begin{bmatrix}6\\6\\9\end{bmatrix}$$

נניח כי רצוננו במבנה קאנוני בו שלושת האיברים הראשונים הם איברי הפתרון הבסיסי. לכן :

$$A_1=\begin{bmatrix}2 & -1 & -1\\ 0 & 1 & 3\\ 1 & 1 & 1\end{bmatrix} \Rightarrow A_1^{-1}=\frac{1}{6}\begin{bmatrix}2 & 0 & 2\\ -3 & -3 & 6\\ 1 & 3 & -2\end{bmatrix} \qquad A_2=\begin{bmatrix}-1 & -9 & 4\\ 5 & 17 & 4\\ 4 & 6 & 5\end{bmatrix}$$

הטבלה המוצעת תהיה :

$$\left[I\middle|A_1^{-1}A_2\middle|A_1^{-1}B\right]=\left[\begin{array}{ccc:ccc:c}1 & 0 & 0 & 1 & -1 & 3 & 5\\ 0 & 1 & 0 & 2 & 2 & 1 & 3\\ 0 & 0 & 1 & 1 & 5 & 1 & 1\end{array}\right]$$

וקיבלנו את המבנה הנדרש להתחלת התהליך. את התהליך הנ"ל יכולנו לעשות גם בדרך אחרת - ע"י פעולות שורה חוקיות (הכפלת שורה בקבוע, חיבור וחיסור שורות) - וזה שקול לחלוטין. כאמור, הפתרון נמצא בעמודה האחרונה וכפי שרואים - הוא חוקי (כל איבריו חיוביים). תוצאה זו היא מזל ולא יותר! בד"כ לא ניתן להבטיח כי עבור בחירה שרירותית של עמודות נקבל תוצאה חוקית.

נחזור לתיאוריה : נניח כי ברצוננו להחליף בפתרון הבסיסי הקיים את x_k (עבור $1\le k\le m$) ב- x_j ($j>m$). הדרך לביצוע פעולה זו היא חלוקת השורה ה- k בגודל $a_{k,j}$ (כלומר - כאן נכנסת הדרישה כי ערך זה אינו שלילי על מנת לאפשר החלפה זו). ערך זה קרוי הציר. כן נבצע שינוי כל אחת מהשורות i למעט השורה ה- k ע"י החסרת השורה ה- k כשהיא מוכפלת ע"י $a_{i,j}$. פעולות אלו יסתכמו בכך שהעמודה ה- j תכלול '1' במקום ה- (k,j), וכל היתר אפסים. העמודות $1,2,\ldots,m$ למעט k לא ישתנו (חיברנו והחסרנו אפסים בלבד בעמודות אלו). התוצאה הסופית תהיה פתרון בסיסי חדש בו יש m איברים שוני מאפס - באינדקסים - $1,2,3,\ldots,k-1,k+1,\ldots,m,j$.

דוגמה 12.6 : נתונה הטבלה הבאה אשר לקוחה מתוצאת הדוגמה הקודמת :

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & -1 & 3 & 5 \\ 0 & 1 & 0 & 2 & 2 & 1 & 3 \\ 0 & 0 & 1 & 1 & 5 & 1 & 1 \end{bmatrix} \end{array}$$

בהינתן הטבלה, עלינו להחליט את איזה מהאיברים בפתרון הבסיסי הנתון רצוננו להחליף (- שקול לקביעת השורה המטופלת), ועם מי להחליפו (- שקול לקביעת העמודה). שתי החלטות אלה קובעות את איבר המנוף. נראה כיצד אנו מחליפים את x_2 ב- x_5. המנוף מסומן בתיאור הבא :

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & -1 & 3 & 5 \\ 0 & 1 & 0 & 2 & 2 & 1 & 3 \\ 0 & 0 & 1 & 1 & 5 & 1 & 1 \end{bmatrix} \end{array}$$

נחלק תחילה את השורה ב- 2 ונקבל :

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & -1 & 3 & 5 \\ 0 & 0.5 & 0 & 1 & 1 & 0.5 & 1.5 \\ 0 & 0 & 1 & 1 & 5 & 1 & 1 \end{bmatrix} \end{array}$$

נשנה את השורה הראשונה ע"י חיבורה לשניה, ונשנה את השלישית ע"י החסרת השורה השניה כשהיא מוכלת ב- 5. נקבל באופן זה :

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} \boxed{1} & 0.5 & \boxed{0} & 2 & \boxed{0} & 3.5 & 6.5 \\ \boxed{0} & 0.5 & \boxed{0} & 1 & \boxed{1} & 0.5 & 1.5 \\ \boxed{0} & -2.5 & \boxed{1} & -4 & \boxed{0} & -1.5 & -6.5 \end{bmatrix} \end{array}$$

והעמודות המוקפות מציגות את הפתרון הבסיסי החדש (שאינו חוקי, כפי שניכר מהעמודה האחרונה!).

באופן הנ"ל הראינו כיצד מפתרון בסיסי אחד נעבור למשנהו. כעת נדון בשאלה השניה - כיצד נוודה כי אנו עוברים לפתרון בסיסי חדש אשר הינו חוקי? לצורך זה נחזור לייצוג בו עשינו שימוש בסעיף קודם, ונציג את השאלה באופן רחב יותר - כיצד נחליט את מי לנפות, ואת מי להכניס לפתרון הבסיסי החדש תוך שמירת חוקיות?

נניח כי בידינו פתרון בסיסי חוקי בו m האיברים הראשונים שונים מאפס. לכן מתקיים:

$$x_1a_1 + x_2a_2 + \cdots + x_ma_m = b$$

(צירוף ליניארי של m העמודות הראשונות של A המקורית). נניח כי ברצוננו להביא למשחק את $x_q \quad q > m$. נוכל לייצג את העמודה a_q כצירוף ליניארי של m העמודות לפי:

$$y_1a_1 + y_2a_2 + \cdots + y_ma_m = a_q$$

ועל-ידי צירוף ליניארי של שתי המשוואות האחרונות מתקבל:

$$(x_1 - \varepsilon y_1)a_1 + (x_2 - \varepsilon y_2)a_2 + \cdots + (x_m - \varepsilon y_m)a_m + \varepsilon a_q = B$$

נניח כי לפחות אחד מאיברי הוקטור Y חיובי, נתחיל מהערך $\varepsilon = 0$ ונגדיל אותו אט אט. בשל מסוים יתקבל כי אחד המקדמים יתאפס ונקבל פתרון בסיסי חוקי חדש.

נשים לב לכך שכיוון שכחלק מהתהליך מתקבל כי העמודות המעורבות $(a_1 \div a_m)$ הינן עמודות קנוניות (אשר יוצרות מטריצת יחידה), מתקבל כי למעשה מקדמי הצירוף הליניארי של a_q הם איברי עמודה זו, דהיינו:

$$a_{1,q}a_1 + a_{2,q}a_2 + \cdots + a_{m,q}a_m = a_q$$

לקבלת ההתאפסות הראשונה תתקבל עבור:

$$\varepsilon = \min_i \left\{ x_i / a_{i,q} \mid a_{i,q} > 0 \right\}$$

הערה: יתכן בהחלט מצב בו כל איברי העמודה ה- q שליליים, ואז ניתן בצירוף הליניארי הנ"ל להגדיל את ε עוד ועוד ללא הגבלה - פירוש הדבר הוא שהקבוצה הקמורה K המייצגת את האילוצים אינה חסומה.

דוגמה 12.7: נניח כי בידינו הטבלה הבאה:

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & -1 & 3 & 5 \\ 0 & 1 & 0 & 2 & 2 & 1 & 3 \\ 0 & 0 & 1 & 1 & 5 & 1 & 1 \end{bmatrix} \end{array}$$

ורצוננו לבחון את אפשרות החלפת העמודה החמישית באחת מהקיימות (1,2,3). ע"י חלוקת איבר איבר בין עמודת ה- X (קיצונית מימין) והעמודה החמישית מתקבל וקטור המנות:

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & -1 & 3 & 5 \\ 0 & 1 & 0 & 2 & 2 & 1 & 3 \\ 0 & 0 & 1 & 1 & 5 & 1 & 1 \end{bmatrix} \end{array} \qquad \begin{bmatrix} 5/-1 \\ 3/2 \\ 1/5 \end{bmatrix} = \begin{bmatrix} -5 \\ 1.5 \\ 0.2 \end{bmatrix}$$

וכפי שרואים, ההחלפה צריכה להיעשות בין העמודה החמישית והשלישית. בדוגמה הקודמת ביצענו החלפה עם העמודה השניה, ולכן קיבלנו פתרון בסיסי שאינו חוקי. ביצוע החלפה זו ייעשה ע"י איבר הציר המוקף:

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & -1 & 3 & 5 \\ 0 & 1 & 0 & 2 & 2 & 1 & 3 \\ 0 & 0 & 1 & 1 & 5 & 1 & 1 \end{bmatrix} \end{array}$$

תחילה נחלק את השורה השלישית ב- 5 ונקבל:

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & -1 & 3 & 5 \\ 0 & 1 & 0 & 2 & 2 & 1 & 3 \\ 0 & 0 & 0.2 & 0.2 & 1 & 0.2 & 0.2 \end{bmatrix} \end{array}$$

ולסיום נחסיר שורה זו משורות 1 ו- 2 על מנת לקבל עמודה חמישית קנונית:

$$\begin{array}{c} \begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & \end{array} \\ \begin{bmatrix} 1 & 0 & 0.2 & 1.2 & 0 & 3.2 & 5.2 \\ 0 & 1 & -0.4 & 1.6 & 0 & 0.6 & 2.6 \\ 0 & 0 & 0.2 & 0.2 & 1 & 0.2 & 0.2 \end{bmatrix} \end{array}$$

וכצפוי קיבלנו כי הפתרון הבסיסי החדש חוקי.

לסיכום ביניים, ראינו כי ארגון הבעיה בטבלה מקילה על הטיפול. נדרש שלב ראשון של איתור פתרון בסיסי חוקי אחד וייצוג טבלאי בו העמודות המעורבות הופכות לקנוניות. אם רצוננו להחליף עמודה בעמודה בפתרון הבסיסי תוך שמירה על החוקיות, מבחן פשוט של מנות ומציאת הערך החיובי המינימלי שלהן מספק בטחון כי המעבר יניב פתרון חוקי. הנקודה שלא טופלה בינתיים היא - כיצד נדאג שבמעבר שניקח נקבל גם ירידה בערך הפונקציה עבורה רצוננו להשיג מינימום? נניח כי נתונה הטבלה הבאה, כפי שכבר ראינו:

$$\left[\begin{array}{cccccc|c} 1 & & & 0 & a_{1,m+1} & \cdots & a_{1,n} & b_1 \\ & 1 & & & a_{2,m+1} & \cdots & a_{2,n} & b_2 \\ & & \ddots & & \vdots & & \vdots & \vdots \\ 0 & & & 1 & a_{m,m+1} & \cdots & a_{m,n} & b_m \end{array}\right]$$

אזי, לפתרון X כללי נוכל לכתוב את הקשר הבא:

$$\sum_{k=1}^{n} x_k a_k = \sum_{k=1}^{m} x_k e_k + \sum_{k=m+1}^{n} x_k a_k = B$$

הכפלת שני אגפי המשוואה הנ"ל בוקטור $C_B^T = \begin{bmatrix} c_1 & c_2 & \cdots & c_m \end{bmatrix}^T$ נותנת:

$$\sum_{k=1}^{m} x_k C_B^T e_k = C_B^T B - \sum_{k=m+1}^{n} x_k C_B^T a_k \quad \Rightarrow \quad \sum_{k=1}^{m} x_k c_k = C_B^T B - \sum_{k=m+1}^{n} x_k C_B^T a_k$$

והוספת הגורם - $\sum_{k=m+1}^{n} x_k c_k$ לשני האגפים נותנת:

$$\sum_{k=1}^{n} x_k c_k = C^T X = C_B^T B + \sum_{k=m+1}^{n} \left[c_k - C_B^T a_k \right] x_k$$

משמעות משוואה זו היא שערך הפונקציה $C^T X$ לפתרון מוצע X שווה לערך הפונקציה עבור הפתרון הבסיסי הנתון בידינו (דהיינו הערך -$C_B^T X$) ותוספת של גורם בעל מבנה מאוד נוח לשימוש. כיוון שחייב להתקיים לכל k כי $x_k \geq 0$, פירוש הדבר כי אם ל- k מסוים בתחום $[m+1, n]$ הגורם $c_k - C_B^T a_k = r_k$ שלילי, הרי שכדאי לשנות את העמודה ה- k, ולקבל $x_k \geq 0$ אשר יגרום לירידה בערך הפונקציה אותה רצוננו להביא למינימום.

נוכל לומר אם כך את הדברים הבאים - אם במצב בו נתון בידינו פתרון בסיסי חוקי קיים גורם r_k שלילי (כאשר k אינו אחד האינדקסים של איברי הפתרון הבסיסי אשר שונים מאפס), ניתן להוריד עוד את ערך הפונקציה ע"י מציאת פתרון בסיסי חדש. במקביל נוכל לומר כי אם כל הערכים r_k חיוביים ניתן לעצור את האלגוריתם בידיעה כי אנו בפתרון האופטימלי. מסתבר כי ל- r_k משמעות אינטואיטיבית ביחס לבעיה אותה רצוננו לפתור. המשמעות של c_k הינה עלותו הסגולה (ליחידה) של המשתנה x_k. הערך $C_B^T a_k$ מודד את המחיר הסגולי של אי השימוש במשתנה x_k בפתרון, כיוון שאי-שימוש בו מחייב בשל האילוצים שימוש במשתנים אחרים לשם "חיפוי עליו". כך יוצא כי אם המחיר של שימוש ישיר גבוה מהמחיר של שימוש עקיף ב- x_k, הגורם הנ"ל חיובי, ולא כדאי להחליף את הפתרון שבידינו, ולהפך.

מסתבר כי נוח מאוד להוסיף לטבלת ה- Simplex שורה נוספת המתייחסת לפונקצית המחיר. הטבלה הקנונית תיראה כך, כאשר השורה התחתונה בטבלה הנ"ל מתייחסת למשוואה אותה ראינו קודם:

$$\left[\begin{array}{cccccccc|c} 1 & & & 0 & a_{1,m+1} & \cdots & a_{1,n} & & b_1 \\ & 1 & & & a_{2,m+1} & \cdots & a_{2,n} & & b_2 \\ & & \ddots & & \vdots & & \vdots & & \vdots \\ 0 & & & 1 & a_{m,m+1} & \cdots & a_{m,n} & & b_m \\ \hline 0 & 0 & \cdots & 0 & r_{m+1} & \cdots & r_n & & -C_B^T X_B \end{array}\right]$$

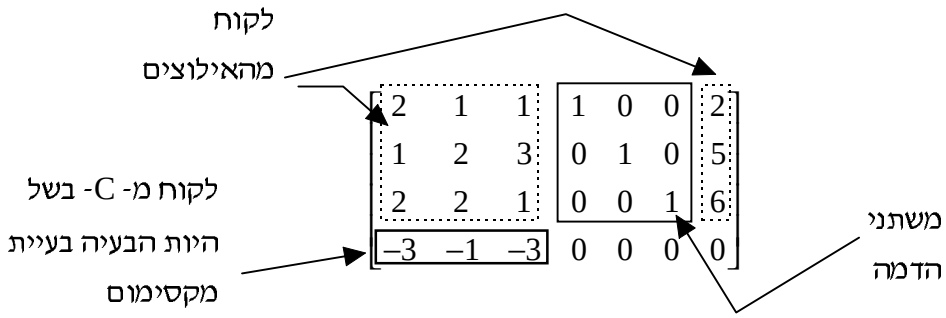
וסדר הדברים יהיה כך - יש לבחון את הערכים r_k - אם כולם חיוביים ניתן לעצור ולקבוע את הפתרון הבסיסי שבידינו כאופטימלי. אם אחד מהם שלילי יש לבחור את הנמוך ביותר שביניהם והוא קובע איזו עמודה תשולב בפתרון החדש. על סמך מנות בין איברי עמודה זו והעמודה הימנית נקבע מיהו האיבר מהפתרון הקודם אשר עומד להיות מוחלף תוך הבטחת חוקיות הפתרון החדש. כעת בידינו כל הדרוש לביצוע עדכון הפתרון - ומובטח כי הפתרון החדש חוקי ומניב ערך פונקציה נמוך יותר. ציור 12.6 נותן סיכום של האלגוריתם במלואו. מקובל לקבל התכנסות תוך m עד $2m$ איטרציות, אם כי ניתן לתפור מקרים שונים לכאן ולכאן.

הערה: נדרשת הוכחה (אשר לא תסופק כאן) לכך שבפינה הימנית תחתונה של הטבלה מתפתח במהלך האלגוריתם ערך הפונקציה.

דוגמה 12.8 : נפתור את הבעיה הבאה:

$$\text{Max.}\quad 3x_1 + x_2 + 3x_3 \qquad \text{S.T.}\quad \begin{aligned} 2x_1 + x_2 + x_3 &\le 2 \\ x_1 + 2x_2 + 3x_3 &\le 5 \\ 2x_1 + 2x_2 + x_3 &\le 6 \\ x_1 \ge 0, x_2 &\ge 0, x_3 \ge 0 \end{aligned}$$

תחילה, בשל היות האילוצים אילוצי אי שיוויון, נוסיף שלושה משתני עזר ונכתוב את הטבלה באופן הבא:



$$\begin{bmatrix} 2 & 1 & 1 & 1 & 0 & 0 & 2 \\ 1 & 2 & 3 & 0 & 1 & 0 & 5 \\ 2 & 2 & 1 & 0 & 0 & 1 & 6 \\ -3 & -1 & -3 & 0 & 0 & 0 & 0 \end{bmatrix}$$

למעשה, בזכות משתני הדמה יש בידינו פתרון בסיסי חוקי:

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 2 & 5 & 6 \end{bmatrix}$$

וקיים כי $C_B = 0$, ולכן איברי השורה התחתונה לקוחים מהוקטור C ישירות. נוכל לחשב את הגדלים הבאים ישירות מהנוסחה שפותחה:

$$r_1 = -3 - [0,0,0]\begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = -3 \qquad r_2 = -1 - [0,0,0]\begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = -1 \qquad r_1 = -3 - [0,0,0]\begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = -3$$

ניכר מהטבלה (העובדה כי הערכים בשורה התחתונה שליליים) כי כל שינוי של אחד ממשתנים אלה במשתנים המקוריים של הבעיה יביא לשיפור בערך הפונקציה. נסמן בעיגול בעמודות 1 עד 3 את האיבר אשר צריך לשמש כציר:

$$\begin{bmatrix} 2 & 1 & 1 & 1 & 0 & 0 & 2 \\ 1 & 2 & 3 & 0 & 1 & 0 & 5 \\ 2 & 2 & 1 & 0 & 0 & 1 & 6 \\ -3 & -1 & -3 & 0 & 0 & 0 & 0 \end{bmatrix}$$

נבחר לשנות את העמודה השלישית כיוון שערך r_3 הוא הכי נמוך. הטבלה לאחר סידרת פעולות שורה תהיה:

$$\begin{bmatrix} 1.67 & 0.33 & 0 & 1 & -0.33 & 0 & 0.33 \\ 0.33 & 0.67 & 1 & 0 & 0.33 & 0 & 1.67 \\ 1.67 & 1.33 & 0 & 0 & -0.33 & 1 & 4.33 \\ -2 & 1 & 0 & 0 & 1 & 0 & 5 \end{bmatrix}$$

כעת העמודה הראשונה הינה היחידה בה אפשר לטפל. איבר הציר מסומן בעיגול. החלפת העמודה הראשונה ברביעית תיתן:

$$\begin{bmatrix} 1 & 0.2 & 0 & 0.6 & -0.2 & 0 & 0.2 \\ 0 & 0.6 & 1 & -0.2 & 0.4 & 0 & 1.6 \\ 0 & 1 & 0 & -1 & 0 & 1 & 4 \\ 0 & 1.4 & 0 & 1.2 & 0.6 & 0 & 5.4 \end{bmatrix}$$

אנו רואים כי כעת אין כל עמודה בה כדאי לבצע החלפה. הפתרון אם כך הוא אופטימלי:

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{bmatrix} = \begin{bmatrix} 0.2 & 0 & 1.6 & 0 & 0 & 4 \end{bmatrix}$$

12.5 איתחול אלגוריתם ה- Simplex

כזכור, אלגוריתם ה- Simplex נחייב איתחול בפתרון בסיסי חוקי. ישנן בעיות בהן פתרון ראשוני כזה הינו מיידי כמו בדוגמה 12.8 - בשל היות כל האילוצים אילוצי אי-שיוויון, נדרשים משתני עזר נוספים אשר מהווים את הפתרון הבסיסי הראשוני, וזהו פתרון חוקי כיוון שאלמנטי הוקטור B חיוביים. נראה זאת למקרה הכללי - נתונה הבעיה:

$$\text{Min. } C^T X \quad \text{S.T.} \quad AX \le B \quad \& \quad X \ge 0$$

בעיה זו מומרת להיות:

$$\text{Min. } \begin{bmatrix} C^T & 0^T \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix} \quad \text{S.T.} \quad \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix} = B \quad \& \quad \begin{bmatrix} X \\ U \end{bmatrix} \ge 0$$

ואיתחול ע"י $X = 0,\ U = B$ הוא חוקי כיוון ש- B חיובי, ובסיסי כיוון שמספר איברי U כמספר האילוצים. השאלה היא מה עושים במקרה הכללי בו פתרון כזה אינו מיידי. נניח כי בידינו הבעיה הבאה:

$$\text{Min. } C^T X \quad \text{S.T.} \quad AX = B \quad \& \quad X \ge 0$$

הצעה ראשונה היא להגדיר את הבעיה החדשה הבאה:

$$\text{Min.} \begin{bmatrix} C^T & k\cdot \mathbf{1}^T \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix} \quad \text{S.T.} \quad \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix} = B \quad \& \quad \begin{bmatrix} X \\ U \end{bmatrix} \geq 0$$

כאשר k סקלאר חיובי וגדול מספיק. כיוון שנכפה על איברי U להיות אי-שליליים, וכיוון שבפונקציה המובאית למינימום ערכים גבוהים של איברי U "משלמים" מחיר k כל אחד, נקבל כי הבעיה "תעדיף" לקבוע את איברי U לאפס. במקרה זה נקבל באופן טיבעי כי בעיה זו מתנוונת לבעיה המקורית ולכן הפתרון הסופי יהיה כשל הבעיה המקורית.

בעיה זו תאותחל ע"י $X = 0,\ U = B$ ושוב - זהו פתרון בסיסי וחוקי. החסרון היחד של גישה זו הוא שלא ידוע כיצד לקבוע את k. בהחלט יתכן שהבעיה תנותב לבחור איברי U שונים מאפס אם המחיר של בחירת איבר כלשהו מ- X עולה יותר. במקרה כזה, כשמתקבלת התוצאה הסופית ובה איבר U שונה מאפס, יש לקבוע את k גבוה יותר ולהתחיל מהתחלה.

גישה טובה יותר לפתרון הבעיה הכללית היא להגדיר בעיית עזר חדשה מהצורה:

$$\text{Min.}\ \mathbf{1}^T Y \quad \text{S.T.} \quad AX + Y = B \quad \& \quad X, Y \geq 0$$

אם לבעיה חדשה זו קיים פתרון אזי פתרון זה הינו וקטור אפסים ל- Y, ווקטור ה- X אשר יתקבל יהיה וקטור אשר מקיים את האילוצים של הבעיה המקורית, והינו חוקי. כלומר - בעיה זו הינה בעיית תכנות ליניארי מדומה אשר פתרונה יכול לשמש כאיתחול לבעיה המקורית. יש לשים לב לעובדה כי פתרון בעיה זו קל כיוון שהאיתחול עבורה הוא Y=B, X=0, כלומר - לבעיה זו האיתחול מיידי.

משמעות הגישה הנ"ל הינה שכל בעיית תכנות ליניארי נפתרת בשני שלבים - בשלב ראשון מוגדרת בעיית תכנות ליניארי חדשה שכל ייעודה הוא לספק איתחול לפתרון הבעיה במקורית. לאחר השגת פתרון לבעיה הראשונה בידינו איתחול נאות. בשלב שני של הטיפול מופעל אלגוריתם ה- Simplex בשנית, והפעם על הבעיה המקורית עם האיתחול שהושג.

1. התחל באיתור פתרון בסיסי חוקי כלשהו.
2. עבור פתרון זה בנה את טבלת ה- Simplex ע"י רישום המטריצה A, העמודה B לצידה, והשורה מתחתיהן באופן הבא:

$$\begin{bmatrix} A & B \\ C^T & -Z \end{bmatrix}$$

כאשר Z היא ערך הפונקציה בפתרון מוצע. הפעל פעולות שורה לשם הבאת העמודות של הפתרון הבסיסי למבנה קנוני. (השורה האחרונה עוברת אותן מניפולציות ללא הבדל).

3. בחר את הערך העמודה בה הערך הנמוך ביותר של r_k - אם הוא חיובי - עצור - בידך הפתרון האופטימלי. אם הוא שלילי - המשך.
4. העמודה בה נמצא r_k הנ"ל תוחלף באחת מעמודות הפתרון הבסיסי הנתון. לשם כך חשב את היחס בין איברי עמודה זו ועמודת B. בחר את היחס החיובי הנמוך ביותר. אם כל היחסים שליליים - בידינו בעיה בה הפתרון אינו חסום, וניתן להגשים מינימום נמוך כרצוננו.
5. בהנחה כי נמצא ערך יחס חיובי נמוך ביותר, מיקומו מגדיר את האיבר מהפתרון הבסיסי הקודם אשר יוחלף. בצע החלפה ע"י פעולות שורה על המטריצה שמטרתם הבאת העמודה ה- k למבנה קנוני.
6. חזור ל- 3.

<u>ציור 12.6 - אלגוריתם ה- Simplex</u>

12.6 אלגוריתם ה- Simplex עבור ערכים חסומים מלמעלה

נניח כי לפנינו בעיית תכנות ליניארי בעלת המבנה הבא:

$$\text{Min.}\quad C^T X \quad \text{S.T.}\quad AX = B \quad \& \quad 0 \le X \le H$$

כאשר H הוא וקטור ערכים חיוביים אשר חוסמים מלמעלה את ערכי X. השאלה בה נעסוק בסעיף זה היא כיצד נפתור בעיה זו ע"י אלגוריתם ה- Simplex. בגישה שכבר הוצגה במסגרת פרק זה נוכל להציע בעיה אלטרנטיבית עם משתני עזר באופן הבא:

$$\begin{aligned} \text{Min.}\quad C^T X \quad \text{S.T.}\quad & AX = B \quad \& \\ & X + Y = H \quad \& \\ & 0 \le X, \quad 0 \le Y \end{aligned}$$

אך גישה זו בזבזנית בחישובים כיוון שבעיה בה טבלת ה- Simplex הינה בגודל $[(m+1)\times(n+1)]$ הומרה בבעיה בה טבלת ה- Simplex היא בגודל - $[(m+n+1)\times(2n+1)]$. נציג לפיכך גישה אלטרנטיבית יעילה יותר. לשם כך נציג את ההגדרה והמשפט הבאים:

הגדרה 12.3: פתרון עבור בעיית החסם הכפול ייקרא "פתרון בסיסי חוקי מורחב" אם זהו פתרון חוקי (העונה לכל האילוצים) ואשר n-m איברים בו שווים לערכי הגבול (אפס או איברי H), ויתר m האיברים מתייחסים ל- m עמודות בלתי תלויות ליניארית של A.

ההבדל בין פתרון בסיסי מורחב ופתרון בסיסי רגיל הוא בכך שבעוד שבעבר n-m איברים של הפתרון היו אפס, הרי שהפעם הם יכולים להיות גם בעלי ערך החסם העליון.

משפט 12.2: פתרון בסיסי חוקי מורחב לבעיית החסם הכפול יהיה אופטימלי אם עבור n-m איברי X שעל הגבולות מתקיים:

$$\begin{aligned} r_j &\ge 0 \quad \text{if} \quad x_j = 0 \\ r_j &\le 0 \quad \text{if} \quad x_j = h_j \end{aligned}$$

הוכחה: אנו ראינו כבר את המשוואה הבאה:

$$C^T X = C_B^T B + \sum_{k=m+1}^{n} r_k x_k$$

נניח כי בידינו פתרון בסיסי חוקי מורחב. כזכור, אם עבור עמודה כלשהי שאינה קנונית מתקבל כי r_j שלילי אזי פירוש הדבר ששימוש ב- x_j צפוי להוריד את ערך פונקצית היעד. במקרה כזה, אם x_j היה אפס, ניתן להגדילו עוד ועוד עד להגעה לחסם העליון. אם ערכו היה בחסם העליון - אין כל רווח בשינויו - טענה זו מתייחסת לשורה הראשונה בתנאי של המשפט.

באופן דומה, אם ערכו של x_j בחסם העליון והתקבל r_j חיובי, פירוש הדבר הוא שכדאי להוריד את ערכו של x_j. ניתן להביאו עד לאפס ובכך להשיג ירידה מירבית בפונקצית היעד. אם לעומת זאת r_j שלילי, מתקבל כי אין כל טעם בהפחתתו - ובכך הוכח המשפט. ■

הדרך ליישם תוצאה זו באלגוריתם ה- Simplex כוללת שינויים ספורים באלגוריתם המקורי אשר לא ייסקרו כאן. נזכיר כי פתרון הבעיה מושג בטבלה בגודל $[(m+2)\times(n+1)]$, כך שהמורכבות החישובית של האלגוריתם כמעט זהה לזו של הבעיה ללא אילוץ כפול.

12.7 בעיית התכנות הליניארי הדואלית ומאפייניה

בהינתן בעיית התכנות הליניארי הבאה:

$$\text{Min.} \quad C^T X \quad \text{S.T.} \quad AX \geq B \ \ \& \ \ X \geq 0$$

ניתן לבנות מתוכה בעיה דואלית מהצורה:

$$\text{Max.} \quad B^T Y \quad \text{S.T.} \quad A^T Y \leq C \ \ \& \ \ Y \geq 0$$

אשר בה הנעלם הוא הוקטור Y. יש לשים לב לעובדה שבבעיה הדואלית מינימיזציה הפכה למקסימיזציה, תפקידי B ו- C הוחלפו, ו- A הפכה ל- A^T. כן נשים לב לעובדה שבחרנו מבנה עם אילוצי אי-שיוון לבעיה המקורית, כיוון שעבורה הבעיה הדואלית בעלת סימטרייה מירבית אליה. עבור בעיה בעלת מבנה קלאסי:

$$\text{Min.} \quad C^T X \quad \text{S.T.} \quad AX = B \ \ \& \ \ X \geq 0$$

נוכל לייצגה ע"י:

$$\begin{aligned} \text{Min.} \quad C^T X \quad \text{S.T.} \quad & AX \geq B \\ & -AX \geq -B \\ & \& \ \ X \geq 0 \end{aligned}$$

וכעת נוכל להגדיר בעיה דואלית ע"י:

$$\begin{aligned} &\text{Max.} \quad B^T U - B^T V = B^T (U - V) \\ &\text{S.T.} \quad A^T U - A^T V = A^T (U - V) \le C \\ &\qquad \& \quad U, V \ge 0 \end{aligned}$$

או לחילופין באופן מתומצת יותר:

$$\text{Max.} \quad B^T Y \quad \text{S.T.} \quad A^T Y \le C$$

כאשר $(U - V) = Y$, ואין אילוצי חיוביות - וכפי שאנו רואים אין כאן סימטרייה מלאה כמקודם.

מהו עניינינו בבעיה הדואלית? מסתבר כי בעיה זו ופתרונה קשורים באופן הדוק לבעיה המקורית ופתרונה. קשר זה משמש לא אחת לשם בניית אלגוריתמים אשר "מתנדנדים" בין שתי הבעיות לשם פתרון מהיר יותר של הבעיה המקורית - גישה זו קרויה The Primal-Dual Algorithm. במסגרת פרק זה לא נציג את אלגוריתם זה ואת התפיסה שמאחריו, אך נציג מספר תוצאות חשובות אשר מבססות את הקשר בין שתי הבעיות.

למה 12.1: אם X ו- Y הם פתרונות חוקיים לשתי הבעיות המקורית (עם אילוץ שיוון) והדואלית (ללא אילוץ חיוביות) בהתאמה, אזי מתקיים:

$$C^T X \ge B^T Y$$

אם X_0 ו- Y_0 פתרונות חוקיים מקיימים $C^T X_0 = B^T Y_0$ אזי בהכרח אלו הם הפתרונות האופטימליים של שתי הבעיות בהתאמה.

הערה - למה זו נכונה גם לבעיה עם - $AX \ge B$

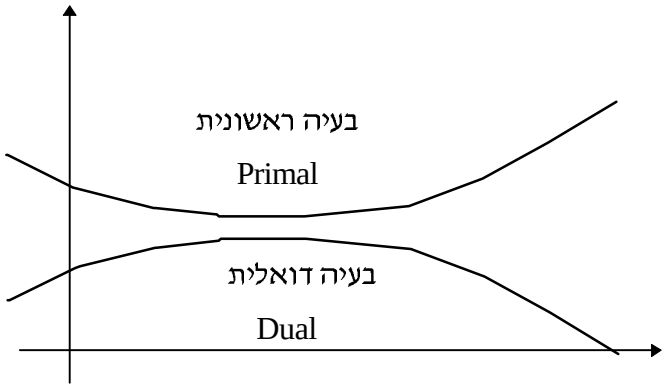
הוכחה: נוכל לכתוב:

$$B^T Y \underset{\substack{\uparrow \\ B=AX}}{=} X^T A^T Y \underset{\substack{\uparrow \\ A^T Y \le C}}{\le} X^T C = C^T X$$

לכן, אם נתונים שני פתרונות חוקיים אשר מקיימים $C^T X_0 = B^T Y_0$, מכיוון שלא יתכן פתרון X בעל ערך נמוך יותר ברור כי X_0 אופטימלי לבעיה הראשונה, ובאופן דומה - Y_0 אופטימלי לבעיה השניה, ובכך הלמה הוכחה. ■

משמעות למה זו היא שערכי הפונקציה אשר עוברת מינימיזציה חוסמים מלמעלה אל ערכי הפונקציה אשר עוברים מקסימיזציה, כמתואר בציור 12.7.

משפט 12.3: עבור בעיית התכנות הליניארי המקורית עם אילוצי שיוויון, נניח כי X הינו הפתרון האופטימלי המתייחס ל- m עמודות A אותן נסמן כמטריצה A_1. אזי, הוקטור $Y=A_1^{-T}C_1$ הוא הפתרון האופטימלי לבעיה הדואלית כאשר C_1 הוא חלק מ- C המתייחס לערכים השונים מאפס.



ציור 12.7 - פונקציית המחיר בבעיות Primal - Dual

הוכחה: נפצל את המטריצה A לפי - $A=\left[A_1|A_2\right]$. אזי הפתרון האופטימלי מקיים - $X_1=A_1^{-1}B$, וקטור הערכים R המתייחס לאיברי X שאופסו נתון ע"י:

$$R_2=\begin{bmatrix} c_{m+1}-a_{m+1}^T C_B \\ c_{m+2}-a_{m+2}^T C_B \\ \\ c_n-a_n^T C_B \end{bmatrix}=C_2-\tilde{A}_2^T C_B$$

הסימון $\tilde{A}_2$ מתיחס לעובדה שזהו אינו A_2 המקורי, אלא לאחר קנוניזציה מהצורה:

$$\left[A_1,A_2\right]\Rightarrow\left[I,A_1^{-1}A_2\right]=\left[I,\tilde{A}_2\right]$$

ולכן:

$$R_2=C_2-A_2^T A_1^{-T}C_1\geq 0$$

וקטור זה גדול מאפס כיוון שאנו בפתרון האופטימלי - ואין כל ערך שלילי המבשר כי יש דרך לרדת עוד בפונקציה. אזי:

$$A^TY=\begin{bmatrix}A_1^TY\\A_2^TY\end{bmatrix}\underset{\substack{\uparrow\\Y=A_1^{-T}C_1}}{=}\begin{bmatrix}C_1\\A_2^TA_1^{-T}C_1\end{bmatrix}\le\begin{bmatrix}C_1\\C_2\end{bmatrix}=C$$

ולכן הפתרון המוצע Y הוא חוקי (הוא מקיים את האילוץ של הבעיה הדואלית). מצד שני:

$$B^TY=B^TA_1^{-T}C_1\underset{\substack{\uparrow\\B^TA_1^{-T}=X_1^T}}{=}X_1^TC_1=C^TX$$

לכן, על פי הלמה 12.1 מתקבל כי בהכרח Y שהתקבל הינו האופטימלי. ■

בסעיף זה הצגנו את מושג הבעיה הדואלית, והקשר בינה ובין הבעיה המקורית. התוצאה המרכזית הינה שמתוך הפתרון של האחת ניתן לקבל את הפתרון של הבעיה האחרת. תוצאות רבות ואלגוריתמים רבים מצויים בהקשר לנושא זה של הבעיה הדואלית. אנו בחרנו להתמקד בבסיס בלבד.

12.8 סיכום

בפרק זה הצגנו את נושא בעיות התכנות הליניארי ואלגוריתם ה- Simplex אשר נועד לפתרונן. ראינו כי מבנה בעיות אלו יוצר מצב נוח בו הפתרונות מוגשמים בקצוות של אזור האילוץ. כתוצאה מכך יש מלכתחילה מספר פתרונות אפשריים סופי, וכל שנדרש הוא לסרוק אותם ולמצוא את זה אשר מניב מינימיזציה של פונקצית היעד. אלגוריתם ה- Simplex מבצע משימה זו באופן יעיל בשל איתחול ע"י פתרון בסיס חוקי, ובשל העבודה שבמעבר מפתרון לפתרון מוודא האלגוריתם כי הפתרון החדש יובטח להיות חוקי, וכזה אשר מקטין את ערך הפונקציה. ראינו במסגרת פרק זה כי בעיות תכנות ליניארי אשר באות במבנה שונה מהקלאסי ניתנות להמרה בדרכים אלו ואחרות לבעיה הקלאסית. את הפרק סיים נושא הבעיה הדואלית - החומר שהוצג בהקשר זה הינו קצהו של קרחון! ישנו שפע של תוצאות בהקשר זה אשר בחרנו שלא להציג במסגרת זו.

לסיום נאמר את הדבר הבא - כמעט לכל אורך הדיון בבעיות אופטימיזציה כלליות הצענו קירוב ריבועי ויישום של אלגוריתמים אשר הוכחו עבור הבעיה הריבועית. כעת נוכל לומר כי ניתן להשתמש באלגוריתם ה- Simplex באותה מידה - נתחיל בניחוש ראשוני, נבצע לפונקצית היעד ולאילוצים ליניאריזציה סביב נקודה זו ונפתור בעיית תכנות ליניארי. סביב הפתרון החדש תבוצע ליניאריציזה, וחוזר חלילה. גישה זו יעילה במיוחד לבעיות מורכבות עם אילוצים מורכבים.

פרק 13 - תנאים הכרחיים ומספיקים בבעיות עם אילוצים

13.1 כללי

בכל הפרקים הקודמים (למעט האחרון) עסקנו בבעיות נטולות אילוצים. אנו ראינו כי גרדיאנט הפונקציה ומטריצת ההסיין שלה בנקודה נתונה משמשות ככלים חשובים בקביעת אופיה של הנקודה, ובדרך ההתקדמות ממנה עבור אלגוריתמים שונים. בפרק זה נכליל תוצאות אלה עבור בעיות עם אילוצים. הבעיה הכללית ביותר בה נרצה לטפל היא הבאה:

$$\begin{aligned} \text{Min } f(x) \quad \text{S.T.} \quad & x \in \Omega \\ & h_k(x) = 0 \quad 1 \le k \le m \\ & g_j(x) \le 0 \quad 1 \le j \le p \end{aligned}$$

כאשר x הוא וקטור של n נעלמים, ומתקיים כי m<n (אחרת האילוצים לבדם מכתיבים את התוצאה ללא מינימיזציה).

כלומר - נרצה לטפל באילוצים משלושה סוגים - אילוצי שייכות לקבוצה, אילוצי שיוויון של פונקציות ואילוצי אי-שיוויון של פונקציות. כבר כעת נאמר כי לאורך כל פרק זה נתעלם מאילוצי שייכות לקבוצה - זהו הקל שבחבורת האילוצים וישנן גישות שונות לטיפול בו - אנו נניח לכל אורך פרק זה כי אין אילוץ כזה או לחילופין - אם קיים אילוץ כזה אזי הפתרון מתקבל בתוך הקבוצה Ω ממש, ולכן כל התכונות אשר נוכיח תקיפות במידה זהה.

במסגרת פרק זה נפגוש מושג אשר בודאי הוצג בעבר בקורסים יסודיים יותר של מתמטיקה - כופלי לגרנז'. בפרק זה נוכיח את משפט כופלי לגרנז' ונמחיש את דרך השימוש ברעיון זה. עם זאת, פרק זה מתאר תוצאות תיאורתיות בלבד - אלגוריתמים נומריים לפתרון בעיות אופטימיזציה עם אילוצים יידונו במסגרת פרק הבא, כשהם עושים שימוש נרחב בהישגים התיאורתיים של פרק זה.

מעניין לציין כי משפט לגרנז' הוצע בשנת 1800 בקירוב, והציע דרך אנליטית לטיפול בבעיות אופטימיזציה עם אילוצי שיוויון. רק לאחר 165 שנה, בשנת 1965 הורחב משפט זה לאילוצי אי-שיוויון - ע"י Kuhn ו- Tucker. עובדה זו גם ממחישה את עוצמת ההישג של Dantzig שהציע את אלגוריתם ה- Simplex ב- 1947 - הרבה לפני שהוצעה דרך כללית לפתרון בעיות עם אילוצי אי-שוויון.

מטעמי נוחות נתחיל את הדיון בבעיות עם אילוצי שיוויון בלבד. רק לאחר ביסוס תוצאות עבור מקרה זה נרחיב את היריעה לטיפול בבעיות הכוללות אילוצי אי-שיוויון.

13.2 תנאים הכרחיים מסדר ראשון - אילוצי שיוויון

נתחיל בטיפול בבעיה הבאה:

$$\text{Min } f(x) \quad \text{S.T.} \quad h_k(x)=0 \quad 1\le k\le m$$

או ברישום מקוצר נאמר - $\text{Min } f(x) \quad \text{S.T.} \quad h(x)=0$, כאשר $h(x)=\left[h_1(x) \quad \cdots \quad h_m(x)\right]^T$.
אנו נניח כי הפונקציה f וכן פונקציות האילוצים רציפות וגזירות ברציפות פעם אחת לפחות - $f(x),\{h_k(x)\}_{k=1}^m \in C^1$. עבור בעיה זו תקף המשפט הבא:

הגדרה 13.1: נקודה X^* תיקרא רגולרית אם מטריצת היעקוביאן של האילוצים:

$$\nabla h(X^*)=\left[\frac{\partial h_1(X^*)}{\partial x} \quad \frac{\partial h_2(X^*)}{\partial x} \quad \cdots \quad \frac{\partial h_m(X^*)}{\partial x}\right]$$

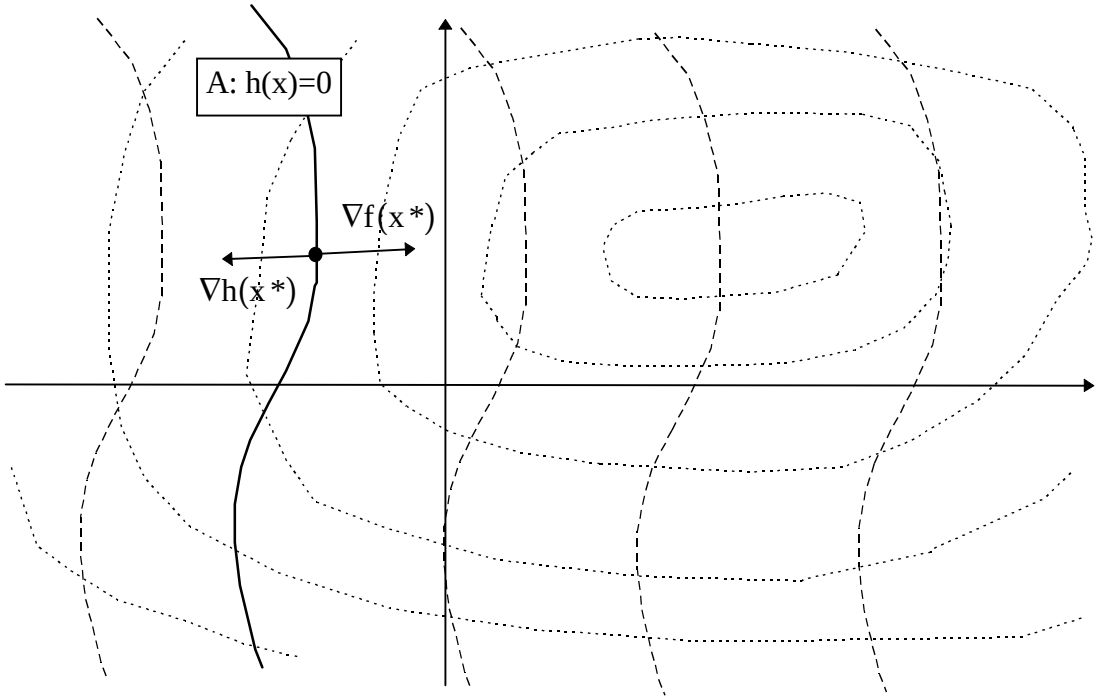
שהינה בגודל $[n\times m]$ היא בעלת דרגה מלאה.

משפט 13.1: (משפט כופלי לגרנז'). תהי X^* נקודת מינימום רגולרית של $f(x)$ המקיימת את האילוץ $h(x)=0$. אזי קיים וקטור אחד ויחיד $\lambda^*=\left[\lambda_1^* \quad \cdots \quad \lambda_m^*\right]$ הנקרא וקטור כופלי לגרנז' כך שמתקיים:

$$\nabla f(X^*)+\sum_{k=1}^{m}\lambda_k^*\nabla h_k(X^*)=\nabla f(X^*)+\nabla h(X^*)\lambda^*=0$$

לפני שניגש להוכחה, ננסה להבין את טענת המשפט באופן אינטואיטיבי.

בציור 13.1 מתוארים קוים שווי גובה של פונקציית המחיר $f(x)$ ופונקציית האילוץ $h(x)$ למקרה דו-מימדי. קו הגובה בו האילוץ מתקיים - $h(x)=0$ - מסומן בקו עבה ונקרה A. על קו זה ורק עליו אנו מחפשים את נקודת הפתרון x^*. מחד, כיוון ש- A הוא קו שווה גובה של פונקציית האילוץ, בכל נקודה בו יתקיים כי הגרדיאנט $\nabla h(x)$ ניצב לקו. מאידך, אנו מסתכלים על ערכי הפונקציה $f(x)$ לאורך A ומחפשים מינימום. בשל משפט הניצבות, בנקודה בה יושג המינימום נקבל כי מרכיב הגרדיאנט $\nabla f(x)$ לאורך A הוא אפס - עובדה שפירושה שגם $\nabla f(x)$ ניצב לקו A. לכן, שני הגרדיאנטים הנ"ל תלויים ליניארית זה בזה - טענת המשפט.



ציור 13.1 - משפט כופלי לגרנז' - הסבר אינטואיטיבי. הגרף מתאר קוים שוי גובה
קו מנוקד - פונקצית המחיר, קו מקווקוו - פונקציית האילוץ

הוכחה: הרעיון בהוכחה יהיה לבנות בעיה תחליפית ללא אילוצים, או למעשה אוסף בעיות כאלה אשר הינן שקולות (כלומר בעלות אותו פתרון אופטימלי בגבול) לבעיה הנתונה. נגדיר את משפחת הבעיות הבאות עבור $j=1,2,3, \ldots$:

$$F_j(x) = f(x) + \frac{j}{2}\|h(x)\|^2 + \frac{\alpha}{2}\|x - X^*\|^2$$

כאשר α פרמטר בעל ערך חיובי כלשהו.

בשל היותה של X^* נקודת מינימום לוקלית של $f(x)$, קיים כדור סגור של נקודות העונות לאילוצים $S_\varepsilon = \{x|\ h(x) = 0,\ \|x - X^*\| \le \varepsilon\}$ ועבור כל נקודה בו מתקיים - $f(x) \ge f(X^*)$. נגדיר את הסידרה x_j כפתרון סידרת הבעיות:

$$\min.\quad F_j(x) \quad S.T. \quad x \in S_\varepsilon$$

וברור כי קיים פתרון לכל אחת מבעיות אלה בשל התיחסותן לפונקציות רציפות והיותן מעל קבוצה קומפקטית (משפט ויירשטרס). כמו כן ברור כי לנקודות אלה יש גבול בשל היותן בתוך קבוצה קומפקטית - נסמן גבול זה ב- $\lim_{j\to\infty} x_j = \hat{X}$. נראה כי גבול זה הינו X^*. מתקיים:

$$F_j(x_j) = f(x_j) + \frac{j}{2}\left\|h(x_j)\right\|^2 + \frac{\alpha}{2}\left\|x_j - X^*\right\|^2 \le F_j(X^*) = f(X^*)$$

הסימן "קטן מ-" נכנס בשל היות x_j נקודת מינימום של הפונקציה הנתונה. בשל היות הנורמה של פונקציות האילוץ חיובית בלבד מתקבל כי צריך להתקיים אי השיויון:

$$f(x_j) + \frac{\alpha}{2}\left\|x_j - X^*\right\|^2 \le f(X^*)$$

ע"י השאפת j לאינסוף מתקבל כי נדרש - $f(\hat{X}) + \frac{\alpha}{2}\left\|\hat{X} - X^*\right\|^2 \le f(X^*)$. בנוסף, מתוך אי-השיויון הקודם עם השאפת j לאינסוף מתקבל כי הגורם שמכפיל את $\left\|h(x_j)\right\|^2$ גדל לאינסוף. נשים לב כי בשל היות f פונקציה רציפה, היא חסומה בתוך התחום S_ε. לכן, חייב להתקיים כי:

$$\lim_{j\to\infty}\left\|h(x_j)\right\|^2 = \left\|h(\hat{X})\right\|^2 = 0$$

קיבלנו אם כך שהנקודה $\hat{X}$ מקיימת את האילוצים והיא בתוך S_ε, ולכן צריך להתקיים (מתוך הגדרת S_ε) כי - $f(\hat{X}) \ge f(X^*)$. מתוך אי-שיויון זה ואי-השיויון ההפוך לו שקיבלנו קודם נובע כי בהכרח מתקבל - $\left\|\hat{X} - X^*\right\|^2$, דהיינו - $\hat{X} = X^*$.

נשתמש כעת בעובדה שעבור בעיות אופטימיזציה ללא אילוצים ידוע כי נקודת מינימום לוקלית חייבת לקיים כי הגרדיאנט בה מתאפס. לכן צריך להתקיים:

$$0 = \nabla F_j(x_j) = \nabla f(x_j) + j\cdot\nabla h(x_j)\cdot h(x_j) + \alpha\left[x_j - X^*\right]$$

כיוון שהמטריצה $\nabla h(X^*)$ בעלת דרגה מלאה, קיים j גדול דיו שייתן כי גם $\nabla h(x_j)$ בעלת דרגה מלאה. מתקבל עבור j זה כי המטריצה $\left[\nabla h(x_j)\right]^T\left[\nabla h(x_j)\right]$ אינה סינגולרית. לכן נוכל לכפול את שני האגפים של המשוואה הנ"ל במטריצה $\left[\nabla h(x_j)\right]^{\#}$ (היפוך מדומה) ולקבל:

$$\left[\left[\nabla h(x_j)\right]^T\left[\nabla h(x_j)\right]\right]^{-1}\left[\nabla h(x_j)\right]^T\cdot\left[\nabla f(x_j)+\alpha\left[x_j-X^*\right]\right]=-j\cdot h(x_j)$$

נשאיף את j לאינסוף ונקבל:

$$-\left[\left[\nabla h(X^*)\right]^T\left[\nabla h(X^*)\right]\right]^{-1}\left[\nabla h(X^*)\right]^T\cdot\nabla f(X^*)=\lim_{j\to\infty}\left\{j\cdot h(x_j)\right\}\overset{\Delta}{=}\lambda^*$$

נציב זאת במשוואת הגרדיאנט המקורית של הפונקציה כש- j שואף לאינסוף ונקבל:

$$0=\nabla f(X^*)+\nabla h(X^*)\cdot\lambda^*$$

והמשפט הוכח. ■

תוצאת המשפט הנ"ל יכולה להתפרש באופן הבא - בנקודת המינימום הלוקלית מתקבל כי וקטור הגרדיאנט של פונקציית המחיר f(x) נפרס ליניארית ע"י וקטורי גרדיאנט פונקציות האילוצים. בכל אופן - תוצאה זו מכלילה את התוצאה אשר התקבלה בפרק 2 ואשר התיחסה לתנאים הכרחיים מסדר ראשון לבעיות ללא אילוצים. למעשה קיבלנו כי בבעיה עם m אילוצי שיויון עלינו לפתור מערכת של m+n משוואות ב- m+n נעלמים כיוון שהמערכת לפתרון כעת היא:

$$\begin{aligned}0&=\nabla f(x)+\nabla h(x)\cdot\lambda\\0&=h(x)\end{aligned}$$

נוח להגדיר פונקציה הקרויה לגרז'יאן - $L(x,\lambda)=f(x)+\lambda h(x)$. פונקציה זו מהווה ריכוז של בעיית האופטימיזציה לפתרון, והיא כוללת בתוכה את האילוצים. במקרה זה התנאים ההכרחיים לקיומה של נקודת מינימום הם:

$$\frac{\partial L(x,\lambda)}{\partial x} = 0 = \nabla f(x) + \nabla h(x)\cdot\lambda$$
$$\frac{\partial L(x,\lambda)}{\partial \lambda} = 0 = h(x)$$

דוגמה 13.1: 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:

$$\frac{\partial L(x_1,x_2,x_3,\lambda)}{\partial x_1} = x_2 + x_3 + \lambda = 0 \;\; ; \quad \frac{\partial L(x_1,x_2,x_3,\lambda)}{\partial x_2} = x_1 + x_3 + \lambda = 0$$
$$\frac{\partial L(x_1,x_2,x_3,\lambda)}{\partial x_3} = x_1 + x_2 + \lambda = 0 \;\; ; \quad \frac{\partial L(x_1,x_2,x_3,\lambda)}{\partial \lambda} = x_1 + x_2 + x_3 - 3 = 0$$

כך שהתקבלה מערכת משוואות ליניארית:

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 3 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -2 \end{bmatrix}$$

דוגמה 13.2: נפתור את הבעיה הבאה, המנסה להביא למקסימום נפחה של תיבה בהינתן כי שטח פניה נתון:

$$\begin{aligned} &\text{Min.} \quad xyz \\ &\text{S.T.} \quad xy + yz + zx = 3 \end{aligned}$$

כמקודם, הגדרת הלנגרז'יאן היא - $L(x,y,z,\lambda) = xyz + \lambda[xy + yz + zx - 3]$. המשוואות המתקבלות הן:

$$yz + \lambda y + \lambda z = 0 \qquad xz + \lambda x + \lambda z = 0$$
$$xy + \lambda x + \lambda y = 0 \qquad xy + yz + xz = 3$$

ואחרי מספר צעדים אלגבריים מתקבל כי הפתרון הוא - $\lambda = -0.5$, $x = y = z = 1$.

דוגמה 13.3 : נפתור את הבעיה הבאה אשר מתיחסת לאנטרופיה מקסימלית כפי שזו מוגדרת בתורת האינפורמציה. בהינתן חוק הסתברות בדיד - $\{p_k\}_{k=1}^N$ לסט ערכים אפשריים $\{x_k\}_{k=1}^N$, (דהיינו - $\Pr ob\{x_k\} = p_k$) מוגדרת כ:

$$\varepsilon = -\sum_{k=1}^{N} p_k \log[p_k]$$

רצוננו להביא למקסימום את האנטרופיה תחת שלושה האילוצים הבאים - הראשון והשני מתיחסים לכך שחוק ההסתברות חוקי ולכן סכום ההסתברויות הוא 1, וכל ההסתברויות חיוביות. האילוץ השלישי מתיחס לכך שרצוננו בממוצע m. נגדיר את הלנגרז'יאן באופן הבא:

$$L\left[\{p_k\}_{k=1}^N, \lambda_1, \lambda_2\right] = -\sum_{k=1}^{N} p_k \log[p_k] + \lambda_1\left[-1 + \sum_{k=1}^{N} p_k\right] + \lambda_2\left[-m + \sum_{k=1}^{N} x_k p_k\right]$$
$$= \sum_{k=1}^{N} p_k\left[-\log(p_k) + \lambda_1 + \lambda_2 x_k\right] - \lambda_1 - m\lambda_2$$

בחרנו להתעלם מאילוץ החיוביות בשל היותו אילוצי אי-שיויון. גזירה לפי אוסף ההסתברויות נותנת:

$$k = 1,2, \ldots ,N: \quad -\log(p_k) + \lambda_1 + \lambda_2 x_k - 1 = 0$$
$$\Rightarrow \quad p_k = \exp\{-1 + \lambda_1 + \lambda_2 x_k\}$$

כך שהתקבל כי אמנם ההסתברויות חיוביות. את כופלי לגרנז' קובעים ע"י קיום האילוצים (פעולה שצריכה להיעשות נומרית בשל היותן של משוואות אלו סתומות).

13.3 תנאים הכרחיים ומספיקים מסדר שני - אילוצי שיויון

בדומה לנעשה בפרק 2, רצוננו לקבל תנאים על הנגזרות מסדר שני של הפונקציות המעורבות, מהן נוכל גם לדלות תנאים הכרחיים (שהרי עד כה התקבלו תנאים מספיקים בלבד). המשפט אשר יובא להלן מהווה למעשה המשך ישיר של משפט כופלי לגרנז' אשר הוצג בסעיף הקודם.

משפט 13.2: (משפט כופלי לגרנז' - המשך). תהי X^* נקודת מינימום של $f(x)$ המקיימת את האילוץ $h(x)=0$. נניח כי המטריצה $\nabla h(X^*)$ בעלת דרגה מלאה. אזי קיים וקטור כופלי לגרנז' אחד ויחיד λ^* כך שמתקיים:

$$\nabla f(X^*)+\nabla h(X^*)\cdot\lambda^*=0$$

אם בנוסף מתקיים כי - $f(x),h(x)\in C^2$, מתקבל כי:

$$\forall y\in V(X^*)\qquad y^T\left[\nabla^2 f(X^*)+\sum_{k=1}^{m}\lambda_k^*\nabla^2 h_k(X^*)\right]y\geq 0$$

$$\text{where:}\qquad V(X^*)=\left\{y\middle|\ \nabla h^T(X^*)y=0\right\}$$

הוכחה: נשתמש בהוכחה של המשפט הקודם ונמשיך ממנה ברצף. כזכור, קיבלנו סידרת בעיות ללא אילוצים אשר פתרונותיה מתכנסים לפתרון הבעיה הנתונה כאן. בהוכחה הקודמת השתמשנו בתנאים על הנגזרות הראשונות - כעת נעשה שימוש בתנאים על הנגזרות השניות, כפי שהן ניתנו בפרק 2. ההסיין של סידרת הבעיות נתון ע"י:

$$\nabla^2 F_j\left(x_j\right)=\nabla^2 f\left(x_j\right)+j\cdot\nabla h\left(x_j\right)\cdot\nabla h^T\left(x_j\right)+j\sum_{k=1}^{m}h_k\left(x_j\right)\nabla^2 h_k\left(x_j\right)+\alpha I$$

וידוע כי מטריצה זו חיובית חצי מוגדרת לכל j ולכל $\alpha>0$. נבחר וקטור y מתוך $V(X^*)$ - כלומר מתקיים כי $\nabla h^T(X^*)y=0$. נבנה מתוך וקטור זה את הוקטור y_j ע"י:

$$y_j=y-\nabla h\left(x_j\right)\left[\nabla h^T\left(x_j\right)\nabla h\left(x_j\right)\right]^{-1}\nabla h^T\left(x_j\right)y$$

קל לראות כי וקטור זה מקיים כי:

$$\nabla h^T\left(x_j\right)y_j=\nabla h^T\left(x_j\right)y-\nabla h^T\left(x_j\right)\nabla h\left(x_j\right)\left[\nabla h^T\left(x_j\right)\nabla h\left(x_j\right)\right]^{-1}\nabla h^T\left(x_j\right)y=0$$

כלומר - הוקטור אשר בנינו ניצב לכל עמודותיה של המטריצה $\nabla h\left(x_j\right)$. בשל היות ההסיין שנכתב קודם מטריצה חיובית חצי מוגדרת מתקבל:

$$0 \le y_j^T \nabla^2 F_j(x_j) y_j = y_j^T \nabla^2 f(x_j) y_j + j \cdot y_j^T \nabla h(x_j) \cdot \nabla h^T(x_j) y_j +$$
$$+ j \sum_{k=1}^{m} h_k(x_j) y_j^T \nabla^2 h_k(x_j) y_j + \alpha y_j^T I y_j =$$
$$= y_j^T \left[\nabla^2 f(x_j) + j \sum_{k=1}^{m} h_k(x_j) \nabla^2 h_k(x_j) \right] y_j + \alpha \left\| y_j \right\|^2$$

לפי תוצאת ההוכחה של משפט 13.1 קיבלנו כי - $\lim_{j \to \infty} \left\{ j \cdot h(x_j) \right\} \overset{\Delta}{=} \lambda^*$. השאפת j לאינסוף ולקיחת $\alpha > 0$ לאפס תיתן כי $y \in V(X^*)$ וכן:

$$0 \le y^T \left[\nabla^2 f(X^*) + \sum_{k=1}^{m} \lambda_k^* \nabla^2 h_k(X^*) \right] y$$

כטענת המשפט. ■

כך קיבלנו כי כמו שבתנאים ההכרחיים מסדר ראשון מעורבות נגזרות ראשונות של האילוצים, כך בתנאים הכרחיים מסדר שני מעורבות במידה דומה נגזרות שניות (מטריצות הסיין) של פונקציות האילוצים. באנלוגיה למקרה נטול האילוצים - קיבלנו כי בנקודת מינימום לוקלית של הפונקציה תחת האילוצים הגרדיאנט המוכלל (זה הכולל את גרדיאנטי האילוצים) מתאפס, וההסיין המוכלל (זה הכולל את הסייני האילוצים) הינו חיובי חצי מוגדר.

מסגרת מעניינת אשר התקבלה כאן היא שהחיוביות החצי מוגדרת של ההסיין המוכלל אינה ביחס לכל המרחב אלא רק לוקטורים אשר ניצבים לעמודות המטריצה $\nabla h(X^*)$. לדרישה זו משמעות אינטואיטיבית - וקטורים ניצבים לעמודות אלה הם הוקטורים שהליכה לאורכם (לפחות מקומית) שומרת על קיום האילוצים (הליכה לכיוון הגרדיאנט משנה הכי מהר את ערך הפונקציה - הליכה בכיוון האנך לגרדיאנט נותנת תנועה על קו גובה קבוע). לכן - משמעות דרישה זו הינה כי החיוביות החצי מוגדרת נדרשת רק עבור וקטורים המקיימים את האילוצים.

דוגמה 13.4 : (המשך של דוגמה 13.1) עבור הבעיה:

$$\begin{aligned} \text{Max} \quad & x_1 x_2 + x_1 x_3 + x_2 x_3 \\ \text{S.T.} \quad & x_1 + x_2 + x_3 = 3 \end{aligned}$$

הגדרנו את הלנגרז'יאן - $L(x_1, x_2, x_3, \lambda) = x_1 x_2 + x_1 x_3 + x_2 x_3 + \lambda [x_1 + x_2 + x_3 - 3]$. גזירתו כפי שתואר קודם נתנה מערכת משוואות ליניארית שפתרונה קל. נחשב את מטריצת ההסיין המוכללת:

$$\nabla^2 f(X^*) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad \nabla^2 h(X^*) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\Rightarrow \nabla^2 f(X^*) + \lambda^* \nabla^2 h(X^*) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

מטריצה זו בבירור אינה חיובית מוגדרת. עם זאת, רצוננו למצוא את התנהגותה עבור וקטורים המקיימים:

$$\nabla h^T(X^*)y = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} y = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = 0$$

נגדיר אם כך את הוקטור הכי כללי המקיים זאת ע"י - $y^T = \begin{bmatrix} y_1 & y_2 & -(y_1+y_2) \end{bmatrix}$. שימוש בוקטור 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 y אלו מתקבל כי ההסיין המוכלל הינו שלילי מוגדר - עובדה המאפשרת כי מדובר בנקודת מקסימום לוקלית.

בדומה לנאמר בסעיף הקודם, נוכל להתיחס להסיין המוכלל כהסיין של פונקציית הלנגרז'יאן. כיוון שפונקציה זו מוגדרת כ - $L(x,\lambda) = f(x) + \lambda h(x)$, קל לראות כי גזירה פעמיים לפי הוקטור x מניבה:

$$\nabla_{xx}^2 L(x,\lambda) = \nabla_{xx}^2 f(x) + \lambda \nabla_{xx}^2 h(x) = \nabla_{xx}^2 f(x) + \sum_{k=1}^{m} \lambda_k \nabla_{xx}^2 h_k(x)$$

עד כה עסקנו בתנאים הכרחיים מסדר ראשון ושני. כעת נדון בתנאים מספיקים (מסדר שני) לקבלתה של נקודת מינימום לוקלית. גם כאן נגלה אנלוגיה פשוטה וישירה למקרה נטול האילוצים.

משפט 13.3: בהינתן $f(x),h(x)\in C^2$, ובהינתן נקודה X^* ווקטור λ^* עבורם מתקיים כי:

$$1.\ \nabla_x L(X^*,\lambda^*)=0 \quad ; \quad \nabla_\lambda L(X^*,\lambda^*)=0$$
$$2.\ \forall y\neq 0 \ \mid\ \nabla h(X^*)y=0 \quad \Rightarrow \quad y^T\nabla^2_{xx}L(X^*,\lambda^*)y>0$$

אזי X^* היא נקודת מינימום לוקלית של $f(x)$ תחת האילוץ $h(x)=0$.

הוכחה: ברור כי X^* היא נקודה המקיימת את האילוץ $h(X^*)=0$ לפי נתוני המשפט. נניח כי משפט זה אינו נכון - דהיינו - קיימת סידרת נקודות x_k המתכנסת ל- X^*, אשר כל נקודותיה מקיימות את הקשרים הבאים:

$$\forall k:\quad 1.\ h(x_k)=0$$
$$2.\ f(x_k)<f(X^*)$$

דהיינו - כל אחת מנקודות סידרה זו מקיימת את האילוצים, וערך הפונקציה בנקודות אלה נמוך מערכה בנקודה X^*. נייצג את אוסף נקודות אלה ע"י הביטוי הבא:

$$x_k=X^*+\delta_k s_k \quad \text{where}\ \ s_k=\frac{x_k-X^*}{\|x_k-X^*\|} \quad \delta_k=\|x_k-X^*\|$$

לכן ברור כי - $\lim_{k\to\infty}x_k=X^* \ \Rightarrow\ \lim_{k\to\infty}\delta_k=0$, ולסידרת הוקטורים המנורמלים s_k (אשר הינה חסומה) קיימת תת-סידרה המתכנסת לוקטור כלשהו S^*. במקום להציע סימון חדש לסידרה זו - נניח ללא כל אובדן כלליות כי כל הסידרה s_k במלואה מתכנסת ל- S^*. קיים הקשר הבא:

$$h(x_k)=h(X^*)=0$$
$$\Rightarrow\quad \frac{h(x_k)-h(X^*)}{\delta_k}=\frac{h(X^*+\delta_k s_k)-h(X^*)}{\delta_k}\xrightarrow[k\to\infty]{}\nabla h(X^*)S^*=0$$

לפי נוסחת טיילור נוכל לכתוב את שני הקשרים הבאים - הראשון עבור האילוץ ה- j:

$$0=h_j(x_k)=h_j(X^*)+\delta_k\nabla h_j(X^*)s_k+\frac{\delta_k^2}{2}s_k^T\nabla^2 h_j\big(\theta_1X^*+(1-\theta_1)x_k\big)s_k$$

והשני בהתיחס לפונקציה f:

$$0 \ge f(x_k) - f(X^*) = \delta_k \nabla f(X^*) s_k + \frac{\delta_k^2}{2} s_k^T \nabla^2 f\left(\theta_2 X^* + (1-\theta_2) x_k\right) s_k$$

כאשר $0 \le \theta_1, \theta_2 \le 1$. הכפלת המשוואה הראשונה בוקטור λ^* וחיבורה לאי-השיויון השני, והשאפת k לאינסוף נותנת:

$$0 \ge \delta_k \left[\nabla f(X^*) + \nabla h_j(X^*)\lambda^*\right] s_k + \frac{\delta_k^2}{2} s_k^T \left[\nabla^2 f(X^*) + \sum_{j=1}^{m} \lambda_j^* \nabla^2 h_j(X^*)\right] s_k$$

מהנחות המשפט מתקבל כי הגורם הראשון מתאפס לכל k. וכך מתקבל כי הגורם השני הינו מטריצה שלילית חצי מוגדרת לוקטור S^* אשר מקיים - $\nabla h(X^*)S^* = 0$ - פירוש הדבר סתירה להנחות המשפט, ולכן ברור כי המשפט נכון. ■

דוגמה 13.5 : נפתור את בעיית האופטימיזציה הבאה:

$$\begin{aligned} &\text{Min.} \quad x + y^2 + yz + 2z^2 \\ &\text{S.T.} \quad x^2 + y^2 + z^2 = 1 \end{aligned}$$

פונקציית הלנגרז'יאן נתונה ע"י:

$$L(x,y,z,\lambda) = x + y^2 + yz + 2z^2 + \lambda\left[x^2 + y^2 + z^2 - 1\right]$$

ולכן:

$$\begin{bmatrix} \frac{\partial L(x,y,z,\lambda)}{\partial x} \\ \frac{\partial L(x,y,z,\lambda)}{\partial y} \\ \frac{\partial L(x,y,z,\lambda)}{\partial z} \\ \frac{\partial L(x,y,z,\lambda)}{\partial \gamma} \end{bmatrix} = \begin{bmatrix} 1 + 2\lambda x \\ 2y + z + 2\lambda y \\ y + 4z + 2\lambda z \\ x^2 + y^2 + z^2 - 1 \end{bmatrix} = 0$$

ושני פתרונות אפשריים הם:

$$\begin{bmatrix} x \\ y \\ z \\ \lambda \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -0.5 \end{bmatrix} \qquad \text{or} \qquad \begin{bmatrix} x \\ y \\ z \\ \lambda \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0.5 \end{bmatrix}$$

ניקח את הפתרון המוצע הראשון. נתבונן במטריצת ההסיין המוכללת עבורו:

$$\nabla^2 f(1,0,0) - 0.5\nabla^2 h(1,0,0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 4 \end{bmatrix} - 0.5\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix}$$

בבירור מטריצה זו אינה חיובית מוגדרת. עם זאת, עלינו לשפוט את תכונותיה תחת תת מרחב הוקטורים המקיימים:

$$\nabla h(1,0,0)y = \begin{bmatrix} 2x & 2y & 2z \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = 2y_1 = 0$$

לכן תחת תת מרחב זה מתקבל:

$$\begin{bmatrix} 0 & y_2 & y_3 \end{bmatrix}\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 & y_2 & y_3 \end{bmatrix}\begin{bmatrix} 0 \\ y_2 + y_3 \\ y_2 + 3y_3 \end{bmatrix} =$$
$$= y_2^2 + 2y_2y_3 + 3y_3^2 = \left[y_2 + y_3\right]^2 + 2y_3^2 > 0$$

כך שבודאות מלאה נוכל לקבוע כי נקודה זו הינה נקודת מינימום מקומית. נתיחס כעת לפתרון האפשרי השני כפי שהתקבל קודם. עבור מתקבל כי מטריצת ההסיין המוכללת הינה:

$$\nabla^2 f(-1,0,0) - 0.5\nabla^2 h(-1,0,0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 4 \end{bmatrix} + 0.5\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 5 \end{bmatrix}$$

ומטריצה זו חיובית מוגדרת. תת מרחב הוקטורים המעניין אותנו הפעם הוא:

$$\nabla h(1,0,0)y = \begin{bmatrix} 2x & 2y & 2z \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -2 & 0 & 0 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = -2y_1 = 0$$

לכן תחת תת מרחב זה מתקבל כמקודם כי ההסיין המוכלל תחת התניית תת-המרחב הינה חיובית מוגדרת ממש - דהיינו - גם הפתרון השני הינו נקודת מינימום לוקלית. תוצאה זו צפויה כיוון שבחירת x=1 או x=-1 אינה משפיעה על ערך הפונקציה עליה נעשית פעולת המינימיזציה וערכה זהה.

13.4 רגישות והקשר לכופלי לגרנז'

מהי משמעותם של כופלי לגרנז' ? מה משמעות ערכיהם? מה ערכים אלו מעידים על הבעיה? בשאלות אלו נעסוק במסגרת סעיף זה. נתחיל דרך דוגמאות אשר ימחישו את המשמעות באופן אינטואיטיבי, ולאחר מכן ניגש לתיאור מסודר ותאורתי יותר.

דוגמה 13.6 : על-פי דוגמה 13.1 קיבלנו כי פתרון בעיית המקסימיזציה:

$$\begin{aligned} &\text{Min} \quad x_1x_2 + x_1x_3 + x_2x_3 \\ &\text{S.T.} \quad x_1 + x_2 + x_3 = 3 \end{aligned}$$

נתון ע"י:

$$\begin{bmatrix} L_{x_1}(x_1,x_2,x_3,\lambda) \\ L_{x_2}(x_1,x_2,x_3,\lambda) \\ L_{x_3}(x_1,x_2,x_3,\lambda) \\ L_{x_4}(x_1,x_2,x_3,\lambda) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 3 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -2 \end{bmatrix}$$

כאשר פונקציית הלנגרז'יאן נתונה ע"י :

$$L(x_1,x_2,x_3,\lambda) = x_1x_2 + x_1x_3 + x_2x_3 + \lambda[x_1 + x_2 + x_3 - 3]$$

כאשר פונקציית המחיר f(x) גדלה פי α, הלנגרז'יאן יהיה:

$$L(x_1,x_2,x_3,\lambda) = \alpha[x_1x_2 + x_1x_3 + x_2x_3] + \lambda[x_1 + x_2 + x_3 - 3]$$

והפתרון לבעיה יהיה:

$$\begin{bmatrix} 0 & \alpha & \alpha & 1 \\ \alpha & 0 & \alpha & 1 \\ \alpha & \alpha & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 3 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix}=\begin{bmatrix} 1 \\ 1 \\ 1 \\ -2\alpha \end{bmatrix}$$

כלומר, ערכו של כופל הלגרנז׳ גדל באותו יחס, אך הפתרון נותר כמקודם. אנו יכולים להתיחס לפונקציית הלגרנז׳יאן כפונקציית מחיר בה יש מרכיב אחד של מחיר לגורם שאותו רצוננו להקצין, ומרכיב שני אשר מתמחר את אי-קיום האילוצים. כאשר גדל מחיר אחד, גדל גם השני באותו יחס כדי לשמור על היחס האופטימלי ביניהם - יחס זה הינו כופל הלגרנז׳. לכן - משמעות אשר נוכל ליחס לכופלי לגרנז׳ היא יחס אופטימלי בין מחיר הפונקציה ומחיר האילוצים.

דוגמה 13.7: עבור האילוץ $x_1+x_2+x_3=3+\delta$ בבעיה הנ״ל, כאשר $\delta \to 0$, $\delta > 0$, פונקציית הלגרנז׳יאן החדשה תהיה:

$$L(x_1,x_2,x_3,\lambda)=x_1x_2+x_1x_3+x_2x_3+\lambda[x_1+x_2+x_3-3-\delta]$$

ופתרון הבעיה החדשה נתון ע״י:

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 0 \\ 3+\delta \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \lambda \end{bmatrix}=\begin{bmatrix} 1+\delta/3 \\ 1+\delta/3 \\ 1+\delta/3 \\ -2(1+\delta/3) \end{bmatrix}$$

הפעם קיבלנו כי כל הפתרון משתנה וכן גם כופל הלגרנז׳. נשים לב כי ערך הפונקציה אשר עוברת מקסימום בבעיה השתנה מ- 3 עבור הבעיה המקורית, ל:

$$3\left(1+\frac{\delta}{3}\right)^2=3\left(1+\frac{2\delta}{3}+\frac{\delta^2}{9}\right)\cong 3+2\delta$$

דהיינו - שינוי של רמת האילוץ ב- δ גרמה לערך הפונקציה שעוברת תהליך קיצון לגדול ב- 2δ. אנו נראה מיד כי למעשה הגידול הצפוי הוא מכפלה של $-\lambda^*=2$ - בגודל השינוי האילוץ δ. דהיינו - כופל לגרנז׳ משמש בהקשר זה כמידת רגישותה של פונקציית המחיר לשינויים באילוצים.

משפט 13.4: (משפט הרגישות). בהינתן הבעיה המאולצת - $\text{Min. } f(x) \text{ S.T. } h(x)=0$ כאשר $h,f \in C^2$, נניח כי X^*,λ^* מגדירים את פתרון חוקי (כלומר - עמודותיה של $\nabla h(X^*)$ בלתי תלויים) של בעיה זו, ופתרון זה מקיים את התנאי המספיק מסדר שני (דהיינו - ההסיין המוכלל המותנה חיובי מוגדר ממש). נגדיר משפחה חדשה של בעיות פרמטריות מהצורה:

$$\text{Min. } f(x) \text{ S.T. } h(x)=c$$

כאשר c וקטור פרמטרים במרחב R^m. אזי, לכל תחום של c הכולל את האפס יש סידרה של פתרונות $X(c)$ וכן $\lambda(c)$ לבעיות הנ"ל, כך ש- $X(0)=X^*$, ומתקיים:

$$\nabla_c f(X(c))\Big|_{c=0} = -\lambda^T(c)\Big|_{c=0}$$

הוכחה: מערכת המשוואות הבאה מייצגת מערכת של n+m משוואות ב- n+m נעלמים (אשר פתרונה הוא X^*,λ^* עבור c=0):

$$\nabla f(x)+\nabla h(x)\lambda=0 \;\; ; \;\; h(x)=c$$

מטריצת היעקוביין של מערכת משוואות זו סביב הנקודה X^*,λ^* נתונה ע"י:

$$J=\begin{bmatrix} \nabla^2 f(X^*)+\sum_{k=1}^{m}\lambda_k^*\nabla^2 h(X^*) & \nabla h^T(X^*) \\ \nabla h^T(X^*) & 0 \end{bmatrix}$$

מטריצה זו אינה סינגולרית - נראה זאת. נניח כי הכפלת מטריצה זו בוקטור $[y,z]$ שווה אפס. לכן:

$$\begin{bmatrix} \nabla^2 f(X^*)+\sum_{k=1}^{m}\lambda_k^*\nabla^2 h(X^*) & \nabla h(X^*) \\ \nabla h^T(X^*) & 0 \end{bmatrix}\begin{bmatrix} y \\ z \end{bmatrix}=$$

$$=\begin{bmatrix} \left[\nabla^2 f(X^*)+\sum_{k=1}^{m}\lambda_k^*\nabla^2 h(X^*)\right]y+\nabla h(X^*)z \\ \nabla h^T(X^*)y \end{bmatrix}=0$$

הכפלת חלקו העליון של וקטור זה ב- y^T ושימוש בעובדה שלפי הנ"ל - $\nabla h^T(X^*)y=0$, מתקבל:

$$y^T\left[\nabla^2 f(X^*)+\sum_{k=1}^{m}\lambda_k^*\nabla^2 h(X^*)\right]y=0$$

והדבר עומד בסתירה להנחות המשפט - לכן ברור כי האפשרות היחידה היא ש- y=0. לכן מתקבל כי $\nabla h(X^*)z=0$, ומכיוון שעמודותיה של המטריצה $\nabla h(X^*)$ אינן תלויות לפי הנחת המשפט - מתקבל כי גם z=0. כך הראינו כי היעקוביין הנ"ל אינו סינגולרי.

בשל אי-הסינגולריות של J ורציפותן של הפונקציות המעורבות (צריך לומר כאן יותר - טענה זו מתבססת על משפט ה- Implicit Function - אך אנו נשאיר זאת כך), קיים תחום S המכיל סביב האפס וקיימים $\lambda(c)$, $X(c)$ כך ש:

$$\nabla f(X(c))+\nabla h(X(c))\lambda(c)=0 \;\; ; \;\; h(X(c))=c$$

ועבור c קרוב דיו לאפס ברור כי פתרון זה גם מקיים כי ההסיין המותנה חיובי מוגדר. נכפול המשוואות לעיל בגורם $\nabla X(c)$ ונקבל:

$$\nabla X(c)\cdot\nabla f(X(c))+\nabla X(c)\cdot\nabla h(X(c))\cdot\lambda(c)=0$$

נשתמש בכלל השרשרת:

$$h(X(c))=c \;\Rightarrow\; \nabla_c h(X(c))=I=\nabla X(c)\cdot\nabla h(X(c))$$

ולכן:

$$0=\nabla X(c)\cdot\nabla f(X(c))+\nabla X(c)\cdot\nabla h(X(c))\cdot\lambda(c)=\nabla_c f(X(c))+\lambda(c)$$

אשר מוכיח את המשפט. ■

אם נחזור לדוגמה האחרונה - קיבלנו כי שינוי של רמת האילוץ ב- δ גרמה לערך הפונקציה שעוברת תהליך קיצון לגדול ב- 2δ. מתקיימת המשוואה:

$$\nabla_c f(X(c))=\frac{f(X(\delta))-f(X(0))}{\delta}=\frac{f(X(\delta))-f(X^*)}{\delta}=-\lambda^T(0)=2$$

כטענת המשפט.

13.5 אילוצי אי- שיויון - הכללת התוצאות הקודמות

כעת נדון בבעיית האופטימיזציה הכללית ביותר אשר תכלול אילוצי אי-שיויון. אנו נניח כי הבעיה הינה בעלת המבנה הכללי הבא:

$$\text{Min. } f(x) \quad \text{S.T.} \quad h_1(x)=0 \;\; ..... \; h_m(x)=0 \quad \{h(x)=0\}$$
$$g_1(x)\le 0 \;\; ..... \; g_p(x)\le 0 \quad \{g(x)\le 0\}$$

כאשר $f, h_k, g_j \in C^1$

הגדרה 13.2: אילוץ אי-שיויון ייקרא אקטיבי אם הוא מתקבל בפתרון בשיויון.

הגדרה 13.3: נקודה X^* המקיימת את האילוצים $h(X^*)=0\ ,g(X^*)\le 0$ תיקרא רגולרית אם אוסף הוקטורים $\{\nabla h_k(X^*)\}_{k=1}^{m},\{\nabla g_j(X^*)\}_{j\in A(X^*)}$ הינם בלתי תלויים ליניארית, כאשר $A(X^*)$ הינו אוסף כל האינדקסים בתחום $j\in[1,p]$ עבורם מתקיים כי - $g_j(X^*)=0$ - האילוצים האקטיביים.

להגדרה זו חשיבות רבה - כפי שאנו נראה מיד, בבעיות עם אילוצי אי-שיויון מתקבל כי אילוצי אי-שיויון נחלקים לשתי קטגוריות - חלקם הופכים לאילוצי שיויון בפועל, וחלקם אינם באים כלל לידי ביטוי ואין להם השפעה על הפתרון. המשפט הבא מרכז תוצאה זו ומגדיר את התנאים ההכרחיים של על נקודת מינימום לוקלית של בעיה מאולצת כללית לקיים.

משפט 13.5: (תנאים הכרחיים עפ"י Kuhn-Tucker - שנת 1961). תהי X^* נקודת מינימום לוקלית של $f(x)$ המקיימת $h(X^*)=0$, וכן - $g(X^*)\le 0$, ונניח כי הנקודה X^* הינה רגולרית. אזי קיימים שני וקטורי כופלי לגרנז' יחידים $\lambda^*=[\lambda_1 \;\; \cdots \;\; \lambda_m],\ \mu^*=[\mu_1 \;\; \cdots \;\; \mu_p]$ המקיימים כי עבור הגדרת הלנגרז'יאן הבאה:

$$L(x,\lambda,\mu)=f(x)+\sum_{k=1}^{m}\lambda_k h_k(x)+\sum_{j=1}^{p}\mu_j g_j(x)$$

את הקשרים הבאים:

$$\nabla_x L(X^*,\lambda^*,\mu^*)=0$$
$$\forall j\in A(X^*)\ \ \mu_j\ge 0$$
$$\forall j\notin A(X^*)\ \ \mu_j=0$$

ובנוסף:

$$\forall y\in V(X^*)\quad y^T\nabla^2_{xx}L(X^*,\lambda^*,\mu^*)y\ge 0$$

$$\text{where:}\quad V(X^*)=\left\{y\ \middle|\ \begin{array}{c}\text{for } k=1..m\ \ \nabla h_k^T(X^*)y=0\\ \&\\ \text{for } j\in A(X^*)\ \ \nabla g_j^T(X^*)y=0\end{array}\right\}$$

הוכחה: למעשה ההוכחה של משפט זה דומה מאוד להוכחות של משפטים 13.1 ו- 13.2, ולכן לא תובא במלואה. נגדיר את הפונקציות:

$$j=1,\ \dots\ ,p\qquad g_j^+(x)=\max\{0,g_j(x)\}$$

ונבנה את משפחת בעיות האופטימיזציה נטולות האילוצים הבאות התלויות בפרמטר r:

$$F_r(x)=f(x)+\frac{r}{2}\|h(x)\|^2+\frac{r}{2}\sum_{j\in A(X^*)}\left[g_j^+(x)\right]^2+\frac{1}{2}\|x-X^*\|^2$$

כאשר $S_\varepsilon\overset{\Delta}{=}\{x|\ \|x-X^*\|\le\varepsilon\}$ ו- ε נקבע להיות קטן דיו כדי לקיים כי $\forall x\in S_\varepsilon\ \ f(x)\ge f(X^*)$.

נשים לב כי הגרדיאנט של הפונקציה $\left[g_j^+(x)\right]^2$ היא $2g_j^+(x)\nabla g_j(x)$. ניתן להראות כי אם הסידרה $\{X_r\}_{r=1}^{\infty}$ מהווה את אוסף הפתרונות למשפחת הבעיות הנ"ל אזי מתקיים כי סידרה זו מתכנסת ל- X^*. יתרה מזו - במקרה זה יתקבל כי כופלי לגרנז' מתקבלים כגבולות הבאים:

$$\lambda_k^*=\lim_{r\to\infty}r\cdot h_k(X_r)\quad;\quad \mu_j^*=\lim_{r\to\infty}r\cdot g_j^+(X_r)$$

ומאחר וברור כי $g_j^+(X_r)\ge 0$ מתקבל כי:

$$\nabla_x L(X^*,\lambda^*,\mu^*)=0$$
$$\forall j\in A(X^*)\ \ \mu_j\ge 0$$
$$\forall j\notin A(X^*)\ \ \mu_j=0$$

ובאופן דומה ניתן להראות כי מתקיים כי:

$$\forall y\in V(X^*)\ \ \ y^T\left\{\nabla^2 f(X^*)+\sum_{k=1}^{m}\lambda_k^*\nabla^2 h_k(X^*)+\sum_{j\in A(X^*)}\mu_j^*\nabla^2 g_j(X^*)\right\}y\ge 0$$

והמשפט הוכח. ■

משפט זה מציג את האבחנה בין אילוצים אקטיביים לכאלה שאינם אקטיביים בקרב קבוצת אילוצי האי-שיויון. אילוצי האי-שיויון שאינם אקטיביים נופלים מהניתוח ואינם משפיעים על דבר בפתרון.

באשר להיותם של כופלי הלגרנז' עבור אילוצי האי-שיויון האקטיביים חיוביים - תוצאה זו צפויה ואינטואיטיבית. אם נזכור את משמעותם של כופלים אלה בהקשר לסעיף הקודם, הרי שקביעת אילוץ האי-שיויון להיות $g_j(x)\le\delta\ \ [\delta>0]$ מקלה את האילוצים ולפיכך צפויה להקטין את $f(x)$ המינימלי. לכן - לפי הגדרת הרגישות יתקבל כי $\nabla_\delta f(X)=-\mu_j\le 0$ ולכן ערכים אלו חיוביים.

המשפט הבא הינו משפט מקביל למשפט 13.3 ובו תנאים הכרחיים ומספיקים לקבלת נקודת מינימום לוקלית של הבעיה המאולצת עם אילוצי שיויון ואי-שיויון. משפט זה יובא ללא הוכחה בשל היותה דומה להוכחת משפט 13.3.

משפט 13.6: (תנאי הכרחי ומספיק מסדר שני). נניח כי X^*, λ^*, μ^* מקיימים את הקשרים הבאים:

1. מתקיים: $g(X^*)\le 0\qquad h(X^*)=0\qquad \nabla_x L(X^*,\lambda^*,\mu^*)=0$.
2. מתקיים: $\forall j\notin A(X^*)\ \ \mu_j=0\ \ ,\ \ \forall j\in A(X^*)\ \ \mu_j>0$ (הדרישה כאן היה לחיוביות ממש של כופלי הלגרנז' האקטיביים).
3. מתקיים: $\forall y\in V(X^*)\ \ \ y^T\nabla_{xx}^2 L(X^*,\lambda^*,\mu^*)y>0$

אזי X^* מובטחת להיות נקודת מינימום לוקלית של הבעיה המקורית עם האילוצים.

דוגמה 13.8 : נפתור את הבעיה הבאה:

$$\begin{aligned} &\text{Min.} \quad (x+y)^2 - 10(x+y) \\ &\text{S.T.} \quad x^2 + y^2 \le 5 \quad ; \quad 3x + y \le 6 \end{aligned}$$

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:

$$\begin{aligned} 0 &= 2(x+y) - 10 + 2x\mu_1 + 3\mu_2 \\ 0 &= 2(x+y) - 10 + 2y\mu_1 + \mu_2 \end{aligned}$$

נניח כי שני האילוצים אינם אקטיביים. במקרה כזה נקבל כי $\mu_1 = \mu_2 = 0$ ולכן:

$$\begin{bmatrix} 10 \\ 10 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad \Rightarrow \quad \forall\alpha, \quad \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha \\ 5-\alpha \end{bmatrix}$$

הצבת תוצאה זו באילוצים נותנת:

$$\begin{aligned} x^2 + y^2 &= \alpha^2 + (5-\alpha)^2 = 25 - 10\alpha \le 5 \quad &\Rightarrow \quad \alpha \ge 2 \\ 3x + y &= 3\alpha + 5 - \alpha = 5 + 2\alpha \le 6 \quad &\Rightarrow \quad \alpha \le 0.5 \end{aligned}$$

וקבוצת פתרונות זו ריקה - משמע - ההנחה כי שני האילוצים אינם אקטיביים אינה נכונה. נניח כי האילוץ הראשון בלבד אקטיבי. נפתור את מערכת המשוואות הבאה:

$$\left.\begin{aligned} 5 &= x + y + x\mu_1 \\ 5 &= x + y + y\mu_1 \\ x^2 + y^2 &= 5 \end{aligned}\right\} \; x = y \quad \Rightarrow \quad x = y = \sqrt{\frac{5}{2}} \quad \Rightarrow \quad \mu_1 = \frac{5 - 2\sqrt{\frac{5}{2}}}{\sqrt{\frac{5}{2}}} = \frac{5\sqrt{2} - 2\sqrt{5}}{\sqrt{5}} > 0$$

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:

$$\left.\begin{array}{r}10=2(x+y)+3\mu_2\\10=2(x+y)+\mu_2\\3x+y=6\end{array}\right\}\ \mu_2=0\quad\Rightarrow\quad\left.\begin{array}{r}x+y=5\\3x+y=6\end{array}\right\}x=0.5\ \ y=4.5$$

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:

$$\left.\begin{array}{r}x^2+y^2=5\\3x+y=6\end{array}\right\}\ \begin{array}{l}x^2+36-36x+9x^2=5\ \Rightarrow\ 10x^2-36x+31=0\\x_1=2.17\quad y_1=-0.53\ \text{ or }\ x_2=1.425\quad y_2=1.723\end{array}$$

הצבת פתרונות אלו במשוואות:

$$\begin{aligned}0&=2(x+y)-10+2x\mu_1+3\mu_2\\0&=2(x+y)-10+2y\mu_1+\mu_2\end{aligned}$$

נותנת לשני הפתרונות האפשריים:

$$\left.\begin{array}{r}6.72=4.34\mu_1+3\mu_2\\6.72=-1.06\mu_1+\mu_2\end{array}\right\}\ \mu_1=-1.78\quad\mu_2=4.825$$

$$\left.\begin{array}{r}10=7.36+2.32\mu_1+3\mu_2\\10=7.36+5.04\mu_1+\mu_2\end{array}\right\}\ \mu_1=0.9882\quad\mu_2=0.2945$$

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.

13.6 סיכום

בפרק זה גיבשנו תנאים הכרחיים ומספיקים לנקודות מינימום (ומקסימום) לבעיות אופטימיזציה הכוללות אילוצי שיוויון ואי-שיוויון. במסגרת זו נפגשנו באופן מסודר עם המושגים פונקציית הלנגרז'איין, כופלי לגרנז', מטריצת ההסיין המוכללת ועוד. מוטיב אשר חזר על עצמו במספר הוכחות היה החלפת הבעיה המאולצת במשפחת בעיות שקולה ללא אילוצים. דבר זה נותן כיוון אפשרי באשר לפתרון נומרי של בעיות עם אילוצים.

פרק 14 - בעיות אופטימיזציה עם אילוצים - גישות לפתרון

14.1 כללי

בפרק קודם הצגנו תוצאות אנליטיות בהקשר לבעיות אופטימיזציה עם אילוצי שיויון ואי-שיויון. מרכיבים מרכזיים בתוצאות אלו היו הלגרז'יאן וכופלי לגרנז'. עם זאת, פרק קודם התמקד בדרכים אנליטיות ללא דיון בהמרת תוצאות אלו לעולם הנומרי. השאלה עליה נענה בפרק זה היא - כיצד ניתן לרתום את תוצאות הפרק הקודם לשם בניית שיטות נומריות לפתרון בעיות אופטימיזציה כלליות הכוללות אילוצים. נרצה כמובן להכליל את האלגוריתמים אותם פגשנו בפרקים קודמים ואשר התיחסו לבעיות ללא אילוצים, ולהרחיבם לטיפול בבעיות החדשות.

פרק זה יציג את דרכי הטיפול המקובלות לטיפול בבעיות אופטימיזציה עם אילוצים באופן שיטחי ובמבנה של סקירה רחבה. נושא זה ראוי שיוקדש לו קורס נפרד בהיותו עשיר למדי בתוצאות ובשיטות. במסגרת פרק זה רצוננו להציג את האינטואיציה מאחורי השיטות הנומריות המקובלות בתחום זה, וכך נכיר שלוש גישות ראשיות לטיפול בבעיה בת n נעלמים עם m אילוצים:

1. גישות מחיר ומחסום (Penalty & Berrier Methods) אשר מציעות כאלטרנטיבה בעיה חדשה ללא אילוצים בה יש n נעלמים למציאה.
2. גישות הטלה למרחב האילוצים, אשר פועלות בתת-מרחב שמימדו n-m. גישה זו שקולה לאמירה כי בשל קיומם של m אילוצים, ניתן להיפטר באופן זה או אחר מ- m נעלמים מוקטור הנעלמים המקורי.
3. גישות מבוססות לנגרז'יאן אשר יוצרות בעיה חדשה בת n+m נעלמים בשל צירופם של כופלי לגרנז' לבעיה המקורית.

14.2 גישת המחיר - The Penalty Method

אנו נתיחס לבעיה הכללית - $\text{Min. } f(x) \;\; \text{S.T. } x \in S$, כאשר האילוץ $x \in S$ מתיחס לאוסף אילוצים פונקציונליים אשר מגדירים מרחב פתרונות חוקיים S. כתחליף לבעיה הנ"ל, נגדיר פונקציית מחיר $P(x)$ בעלת התכונות הבאות:

1. P תהיה פונקציה רציפה.
2. אי-שליליות - $\forall x \;\; P(x) \geq 0$.
3. $P(x)=0$ רק לנקודות $x \in S$.

בהינתן פונקציית מחיר כזו, נגדיר משפחה של בעיות אופטימיזציה פרמטריות נטולות אילוצים:

$$\text{Min.}\ \ f(x)+c_k P(x)$$

אזי ניתן להראות (דורש הוכחה) 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 מתואר בציור 14.1. 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 כאשר c_k גבוה, מטריצת ההסיין של הפונקציה עם המחיר בעלת פיזור ערכים עצמיים גבוה ההולך לאינסוף כאשר c_k הולך לאינסוף. לכן - טיפול ישיר בבעיה עם c_k גבוה יניב תהליך איטרטיבי איטי ביותר. כתחליף, מוצע תהליך בו c_k גדל במקביל לחיפוש רציף של הפתרון.

1. התחל בערך נמוך c_1 ופתור את הבעיה $\text{Min.}\ \ f(x)+c_1P(x)$ בגישה כלשהי המתאימה לבעיה הנתונה. קבל את X_1 - פתרון הבעיה.
2. האיטרציה ה- k בחר $c_k > c_{k-1}$ ופתור את הבעיה החדשה, כשהפתרון הקודם X_{k-1} משמש כאיתחול. קבל את הפתרון החדש X_k.
3. עבור c_k גדול דיו ניתן לעצור את התהליך האיטרטיבי. כלל עצירה יכול להיות מידת השינוי של הפתרון, דהיינו - $\|X_k - X_{k-1}\|/\|X_k\| \le \varepsilon$.

<u>ציור 14.1 - שיטת המחיר לפתרון בעיית אופטימיזציה עם אילוצים</u>

עד כה לא הראינו כיצד צריכה להיבנות פונקציית המחיר. נניח אם כך כי לפנינו הבעיה הכללית הבאה:

$$\begin{aligned}\text{Min.}\ \ f(x)\quad \text{S.T.}\ \ & h_k(x)=0 \quad 1\le k\le m\\ & g_j(x)\le 0 \quad 1\le j\le p\end{aligned}$$

אזי, נוכל להגדיר באופן הבא את פונקציית המחיר:

$$P(x)=\sum_{j=1}^{p}\rho_g\left\{\max\left[0,g_j(x)\right]\right\}+\sum_{k=1}^{m}\rho_h\left\{h_k(x)\right\}$$

כאשר הפונקציות $\rho_g(\alpha),\rho_h(\alpha)$ חיוביות לכל α, שוות לאפס רק עבור $\alpha=0$ ומונוטוניות עולות כשמתרחקים מהאפס (לשני הצדדים). פונקציית מחיר זו עונה על כל הדרישות אותן הזכרנו קודם (דורש הוכחה).

דוגמה 14.1 : 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:

$$\text{Min. } q(x,y,c)=x^2+xy+y^2-2y+c(x+y-2)^2$$

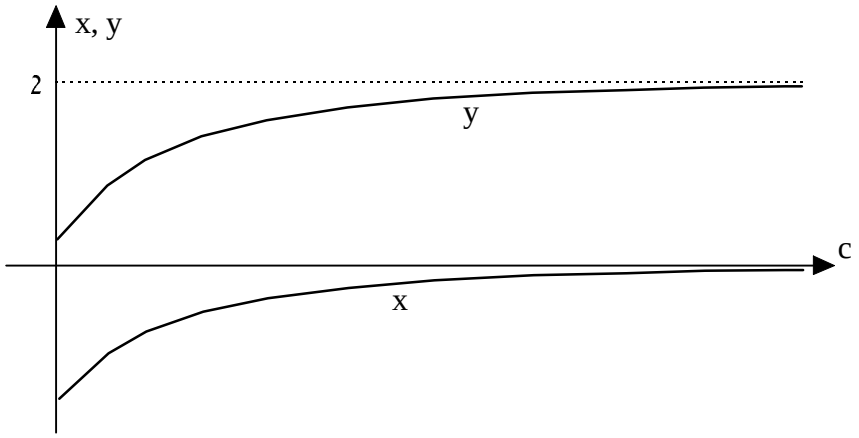
גזירה לפי הנעלמים x ו- y מניבה:

$$\frac{\partial q(x,y,c)}{\partial x}=2x+y+2c(x+y-2)=(2+2c)x+(1+2c)y-4c=0$$
$$\frac{\partial q(x,y,c)}{\partial y}=x+2y-2+2c(x+y-2)=(1+2c)x+(2+2c)y-2-4c=0$$

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:



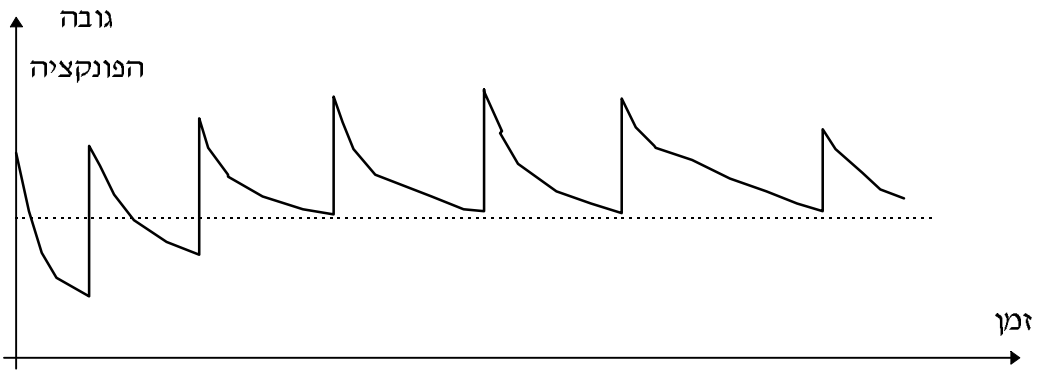
ציור 14.2 - פתרונות הבעיה בדוגמה 14.1 כפונקציה של c

עבור הבעיה הנ"ל קל גם לקבל פתרון ישיר, ע"י הצבת האילוץ בפונקציית המחיר, דהיינו - $x = y - 2$. הבעיה תהיה שקולה לבעיה הבאה, נטולת האילוצים:

$$\text{Min.} \quad (2-y)^2 + (2-y)y + y^2 - 2y = y^2 - 4y + 4 = (y-2)^2$$

ולכן ברור כי הפתרון נתון ע"י $y = 2, x = 0$, כפי שקיבלנו משיטת המחיר.

נסתכל על גובה הפונקציה $f(x) + cP(x)$ כפונקציה של הזמן. כשציר הזמן מתקדם אנו מבצעים עוד ועוד איטרציות, ומפעם לפעם אנו מגדילים את הערך c. גובה הפונקציה לפיכך יראה כשן משור. למספר איטרציות הבאות בזו אחר זו נקבל ירידה ואז שוב תגבה הפונקציה בשל הגדלת c. ציור 14.3 ממחיש עיקרון זה.



ציור 14.3 - גובה הפונקציה (עם המחיר כפונקציה של הזמן בתהליך הפתרון

בהנחה שקבוצת הפתרונות החוקיים אינה ריקה, נגיע לערכי c בהם המחיר 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.

עד כה הגדרנו סידרת בעיות פרמטריות נטולות אילוצים כתחליף לבעיה אחת עם אילוצים. מסתבר כי ניתן לקבוע כתחליף לבעיה המקורית בעיה יחידה אשר תייצגה נאמנה, ופתרונה יתלכד עם פתרון הבעיה המקורית. גישה זו קרויה פונקציית המחיר המדויקת, וזו נתונה ע"י הקשר:

$$P(x) = \sum_{j=1}^{p} \max\left[0, g_j(x)\right] + \sum_{j=1}^{p} \left|h_k(x)\right|$$

(כלומר, הפונקציות $\rho_{g,h}$ הן פונקציות ערך מוחלט). עבור כל בחירה של c המקיימת:

$$c > \max\left\{|\lambda_k|\ ,1 \le k \le m\ ;\ \mu_j\ ,1 \le j \le p\right\}$$

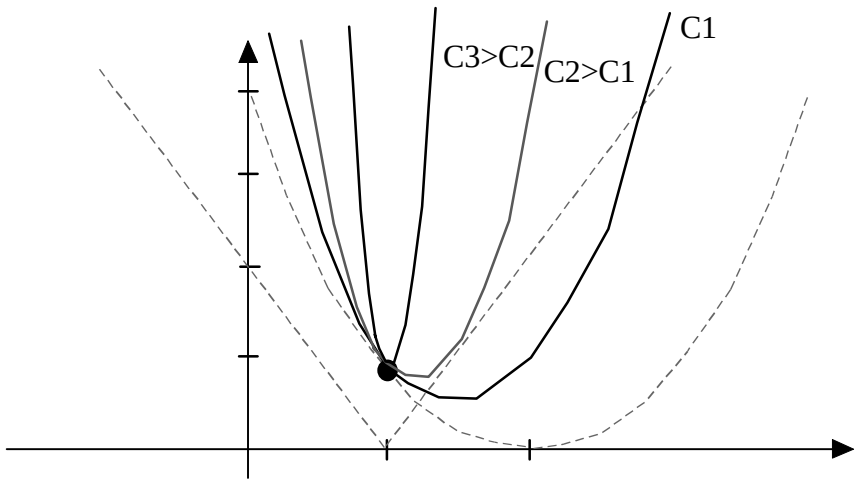
מתקבל כי פתרון בעיה זו זהה לפתרון הבעיה המקורית עם אילוצים (דורש הוכחה).

דוגמה 14.2: נציג את הבעיה הפשוטה הבאה אשר נועדה להמחיש אינטואיטיבית את משמעות התוצאה הנ"ל. נתיחס לבעיית האופטימיזציה הבאה אשר פתרונה פשוט:

$$\text{Min.}\ \ f(x) = (x-2)^2 \quad \text{S.T.}\ \ x = 1$$

על פי גישת המחיר המדויקת נבצע מינימיזציה של הפונקציה - $q(x,c) = (x-2)^2 + c|x-1|$ עבור c גדול מערך מסויים. גרפית תיראה פונקציה זו כבציור 14.4. ואנו 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:

$$2(x-2) + \lambda = 0,\ \ x-1 = 0 \quad \Rightarrow \quad x = 1,\ \lambda = 2$$



ציור 14.4 - תיאור הפונקציה הכוללת מחיר בדוגמה 14.2

כאשר c>2 מתקבל כי מימין ומשמאל לנקודה x=1 בגרף הנ"ל תחול עליה כיוון שמתקיים:

$$\nabla f\left(1^+\right) + 2\nabla h\left(1^+\right) = 0 \quad \Rightarrow \quad \nabla f\left(1^+\right) < c \cdot \nabla h\left(1^+\right)$$

(העליה משמאל ברורה לחלוטין כיוון שחיברנו שתי פונקציות עולות). לכן, ברור כי בחירה זו של c תיתן כי פתרון המינימום מתלכד עם x=1, כדרוש.

דוגמה 14.3 : כהמשך לדוגמה 14.1, נמצא את כופל הלגרנז' אשר אמור לשמש לבעיה.

$$\begin{array}{l} \text{Min. } f(x,y)=x^2+xy+y^2-2y \\ \text{S.T. } h(x,y)=x+y-2=0 \end{array} \Rightarrow L(x,y,\lambda)=x^2+xy+y^2-2y+\lambda(x+y-2)$$

ומתקבל:

$$\begin{array}{l} L_x=2x+y+\lambda=0 \\ L_y=x+2y-2+\lambda=0 \\ L_\lambda=x+y-2=0 \end{array} \Rightarrow \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \\ \lambda \end{bmatrix}=\begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \\ \lambda \end{bmatrix}=\begin{bmatrix} 0 \\ 2 \\ -2 \end{bmatrix}$$

לכן, לכל c חיובי (גדול מ- 2-) יתקבל כי מינימיזציה של הפונקציה:

$$q(x,y,c)=x^2+xy+y^2-2y+c|x+y-2|$$

תיתן פתרון זהה השווה לפתרון הבעיה המקורית.

14.3 גישת המחסום - The Berrier Method

גישת המחסום דומה לגישת המחיר כפי שנראה מיד, אך מתאימה לבעיות בהן יש אילוצי אי-שיוויון בלבד. אנו נתיחס לבעיה הכללית - $\text{Min. } f(x) \text{ S.T. } x\in S$, כאשר האילוץ $x\in S$, כמקודם, מתיחס לאוסף אילוצים פונקציונליים אשר מגדירים מרחב פתרונות חוקיים S. הרעיון בגישת המחסום הוא להציע בעיה תחליפית בה נבנית פונקציית מחסום B(x) אשר מונעת מתהליך החיפוש לצאת את תחום הקבוצה S. האלגוריתם יאותחל בנקודה המקיימת את האילוצים ומצויה **בתוך S**, ויישום תהליך חיפוש כלשהו על הפונקציה המשולבת הכוללת את פונקצית המחסום. הפונקציה B(x) בעלת התכונות הבאות:

1. B תהיה פונקציה רציפה.
2. אי-שליליות - $\forall x \;\; B(x)\ge 0$.
3. עבור נקודות השפה של S מתקיים - $B(x)\to\infty$

בהינתן פונקציית מחסום כזו, נגדיר משפחה של בעיות אופטימיזציה פרמטריות נטולות אילוצים:

$$\text{Min.}\quad q(x,c_k)=f(x)+\frac{1}{c_k}B(x)$$

אזי 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 מתואר בציור 14.5. 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 פונקציית המחסום, נניח אם כך כי לפנינו הבעיה הכללית הבאה (אילוצי אי-שיוויון בלבד):

$$\text{Min.}\quad f(x)\quad \text{S.T.}\quad g_j(x)\le 0\quad 1\le j\le p$$

אזי, נוכל להגדיר באופן הבא את פונקציית המחסום, כאשר הפונקציות $\rho(\alpha)$ חיובית לכל α, שווה לאפס רק עבור $\alpha=0$, מונוטונית עולה, וכן - $\rho(\alpha)\xrightarrow[\alpha\to\infty]{}\infty$:

$$B(x)=\sum_{j=1}^{p}\rho\left\{\frac{-1}{g_j(x)}\right\}$$

דרך אחרת לקביעת פונקציית המחסום היא:

$$B(x)=-\sum_{j=1}^{p}\log\left\{-g_j(x)\right\}$$

בחירה זו לא מקיימת את התכונה $\forall x\ \ B(x)\ge 0$, 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.

דוגמה 14.4: עבור הבעיה הפשוטה הבאה נמחיש את משמעות פונקציית המחסום, לעומת פונקציית המחיר. נניח כי לפנינו הבעיה הבאה:

$$\text{Min.}\quad f(x)=(x-3)^4\quad \text{S.T.}\quad g_1(x)=x-4<0\ \ \&\ \ g_2(x)=2-x<0$$

שיטת המחיר היתה מציעה כאפשרות את הפונקציה הבאה:

$$P(x)=\max\{0,x-4\}^2+\max\{0,2-x\}^2=\begin{cases}(x-2)^2 & x<2\\ 0 & 2\le x\le 4\\ (x-4)^2 & 4<x\end{cases}$$

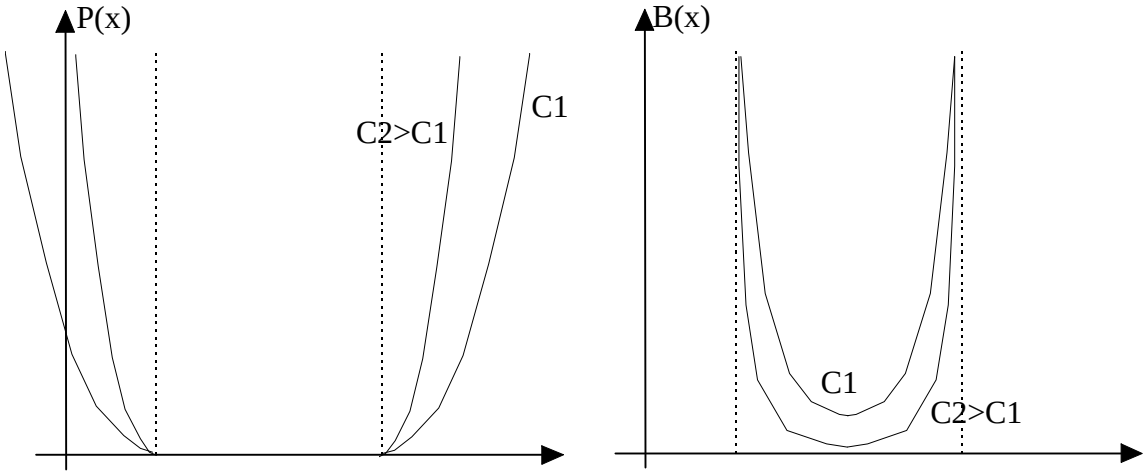
שיטת המחסום היתה מציעה כאפשרות את הפונקציה - $B(x)=\left(\frac{1}{4-x}\right)^2+\left(\frac{1}{x-2}\right)^2$.

1. מצא נקודה X_0 התחלתית אשר מקיימת את האילוצים, ומצויה בתוך התחום S.
2. התחל בערך נמוך c_1 ופתור את הבעיה $\text{Min. } f(x)+c_1^{-1}B(x)$ בגישה כלשהי המתאימה לבעיה הנתונה. קבל את X_1 - פתרון הבעיה.
2. האיטרציה ה- k בחר $c_k>c_{k-1}$ ופתור את הבעיה החדשה, כשהפתרון הקודם X_{k-1} משמש כאיתחול. קבל את הפתרון החדש X_k.
3. עבור c_k גדול דיו ניתן לעצור את התהליך האיטרטיבי. כלל עצירה יכול להיות מידת השינוי של הפתרון, דהיינו - $\|X_k-X_{k-1}\|/\|X_k\|\le\varepsilon$.

<u>ציור 14.5 - שיטת המחסום לפתרון בעיית אופטימיזציה עם אילוצים</u>

גרפית היו שתי גישות אלה נראות כבציור 14.6. וכפי שאנו רואים - בעוד שפונקציית 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 פונקציית 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 המובאית למינימום. פונקציה זו היא $f(x)+c^{-1}B(x)$, ובניגוד לגישת המחיר, בכל פעם שמגדילים את הקבוע c נקבל כי הפונקציה מונמכת, והתהליך האיטרטיבי מקטין אף הוא את גובה הפונקציה.



ציור 14.6 - תיאור פונקציית המחיר והמחסום לבעיה נתונה

גישת המחסום לא יכולה לטפל בבעיות שיויון בשל היות הקבוצה S הנובעת מהן קבוצה ללא איזור פנימי - לכל נקודה המקיימת את האילוצים קיימת נקודה קרובה לה כרצוננו אשר אינה מקיימת את האילוצים, ולכן $\text{Interior}\{S\} = \phi$.

נראה זאת לאילוץ שיויון ליניארי - נניח כי האילוץ הוא AX=B - סט של m משוואת אילוץ. נניח כי מצאנו פתרון אחד X_0 (מיני-רבים) המקיים אילוץ זה. נניח כי יש לנו וקטור Y שאינו בגרעין של A, דהיינו - $AY \neq 0$ (יש אינסוף וקטורים כאלה כיוון שמימד תת-המרחב המשלים לגרעין הוא m). אזי, לכל $\varepsilon > 0$ כלשהו קטן כרצוננו, הוקטור $X + \varepsilon Y$ לא יקיים את האילוצים:

$$A(X + \varepsilon Y) = AX + \varepsilon AY = B + \varepsilon AY \neq B$$

וכך קיבלנו כי לכל נקודה בקבוצת הפתרונות החוקיים נוכל להתקרב כרצוננו ע"י נקודות מחוץ לקבוצה זו - לכן אין לה פנים.

14.4 שיטת ההטלה למרחב האילוצים

הרעיון המרכזי של גישה זו 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.

נתיחס לבעיה הבאה הכוללת n נעלמים, m אילוצי שיויון, ו- p אילוצי אי-שיויון:

$$\text{Min.}\quad f(x)\quad \text{S.T.}\quad h(x)=0,\ \ g(x)\le 0$$

שיטת ה- SD מציעה את משוואת האיטרציה - $X_{k+1}=X_k-\mu\nabla f(X_k)$. נניח כי הפתרון הנתון באיטרציה ה- k - X_k - מקיים את האילוצים. נניח כי בנקודה זו q אילוצים פעילים (כל אילוצי השיויון, וחלק מאילוצי האי-שיויון), והגרדיאנט עבורם ידוע לנו ונתון כמטריצה $A(X_k)$ בגודל $[q\times n]$. הסימון $h(X)$ יכלול את כל המשוואות הללו יחדיו, ו- $A(X_k)$ היא מטריצת היעקוביאן של סט משוואות אלה. סביב הנקודה X_k באופן מקומי נוכל לקבוע את הקשר הבא המהווה קירוב טיילור מסדר ראשון של משוואת אלה:

$$\begin{aligned}&h(X)=h(X_k)+A(X_k)\cdot[X-X_k]\\&\Rightarrow A(X_k)X=A(X_k)X_k-h(X_k)\quad\Rightarrow\quad AX=B\end{aligned}$$

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:

$$P(X_k)=I-A^T(X_k)\left[A(X_k)A^T(X_k)\right]^{-1}A(X_k)$$

ונציע את האיטרציה הבאה:

$$X_{k+1}=P(X_k)\left[X_k-\mu\nabla f(X_k)\right]=X_k-\mu P(X_k)\nabla f(X_k)$$

הטלת הנקודה X_k למרחב אילוצים אינה משנה אותה כיוון שהיא מקיימת אותם לפי הנחתנו.

נחזור להתחלה: נראה תחילה את מקורו של אופרטור הטלה זה, כשאנו עושים שימוש בקירוב הליניארי של האילוצים כמתואר לעיל. בפועל, נתונה נקודה X שאינה מקיימת את האילוצים, ורצוננו למצוא את הנקודה הקרובה ביותר אליה Y אשר תקיימם. בניסוח מתמטי, זה נראה כך (אנו עושים שימוש בקירוב הליניארי של האילוצים):

$$\text{Min.}\ \frac{1}{2}\|Y-X\|^2\quad \text{S.T.}\quad AY=B$$

פתרון בעיה זו יכול להיעשות ע"י q כופלי לגרנז' (אחד כנגד כל אילוץ שיויון):

$$L(Y,\lambda)=\frac{1}{2}\|Y-X\|^2+\lambda^T[AY-B]$$
$$\frac{\partial L(Y,\lambda)}{\partial Y}=Y-X+A^T\lambda=0 \quad \Rightarrow Y=X-A^T\lambda$$

הצבת הפתרון באילוץ נותנת:

$$A\left[X-A^T\lambda\right]=B \;\Rightarrow\; \lambda=\left[AA^T\right]^{-1}[AX-B]$$
$$\Rightarrow\; Y=X-A^T\left[AA^T\right]^{-1}AX+A^T\left[AA^T\right]^{-1}B$$

נסמן את מטריצת ההטלה כפי שהצענו קודם - $P=I-A^T\left[AA^T\right]^{-1}A$, ובעצם התוצאה הנ"ל אומרת כי על מנת לקבל את הנקודה הקרובה ביותר ל- X ואשר משמרת את האילוצים, עלינו לבצע את החישוב:

$$Y=PX+A^T\left[AA^T\right]^{-1}B=PX+\tilde{B}$$

אם הנקודה הנתונה X מקיימת את האילוצים נקבל כי AX=B. לכן:

$$PX+\tilde{B}=X+A^T\left[AA^T\right]^{-1}[-AX+B]=X$$

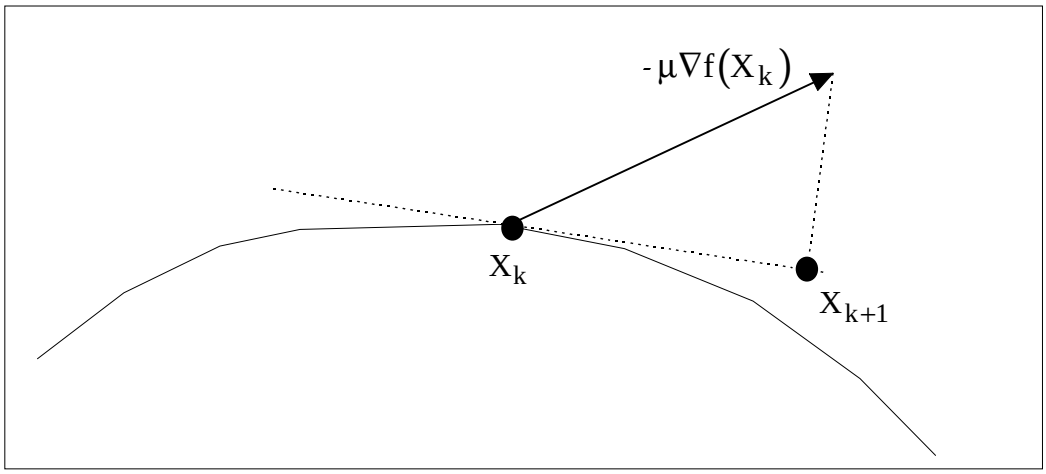
נחזור כעת לאלגוריתם המוצע: לכאורה, יש לנו טעות כיוון שהצענו לקחת את תוצאת ה- SD לכופלה ב- P, תוך התעלמות מהוקטור הנוסף $\tilde{B}$. נראה כעת כי אין שגיאה כאן. נתחיל בביטוי המדויק אותו יש להפעיל:

$$X_{k+1}=P(X_k)\left[X_k-\mu\nabla f(X_k)\right]+\tilde{B}(X_k)=P(X_k)X_k+\tilde{B}(X_k)-\mu P(X_k)\nabla f(X_k)$$

כזכור, הנקודה X_k מקיימת את האילוצים ולכן - $X_k=P(X_k)X_k+\tilde{B}(X_k)$, וכך אנו מקבלים את המשוואה שהוצעה קודם.

נקודה חשובה נוספת היא כיוון החיפוש המיושם בפועל - נשים לב כי כיוון העדכון החדש הוא הוקטור $P(X_k)\nabla f(X_k)$ - זהו כיוון ירידה כיוון שאופרטור ההטלה מהווה מטריצה חיובית חצי מוגדרת (נובע מהגדרתו).

אופרטור ההטלה אשר הצענו הינו אופרטור הטלה מדויק 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 14.7.



ציור 14.7 - המחשת הסטיה הצפויה בהטלה ליניארית על אילוצים לא ליניאריים

רעיון מעניין לפתרון הבעיה נ״ל הוא הבא: נניח כי הפעלת האיטרציה כולל ההטלה הביאו לידינו את הנקודה Y. נרצה כעת לבצע תיקון על Y לשם הבאתה לקיים את האילוצים, ולקבל Y* אשר מקיימת h(Y*)=0 (הפונקציה הוקטורית h כוללת גם את אילוצי אי-השיוויון הפעילים). נציע עדכון מהצורה:

$$Y^* = Y + A^T(X_k)\alpha$$

אשר מוסיף ל- Y מרכיב שאינו בתת-המרחב אליו הטלנו קודם. נדרוש:

$$h(Y^*) = 0 = h\left(Y + A^T(X_k)\alpha\right) \cong h(Y) + A(X_k)A^T(X_k)\alpha$$

$$\Rightarrow \begin{aligned} \alpha &= -\left[A(X_k)A^T(X_k)\right]^{-1} h(Y) \\ Y^* &= Y - A^T(X_k)\left[A(X_k)A^T(X_k)\right]^{-1} h(Y) \end{aligned}$$

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 Y אינה רחוקה מהנקודה X_k, כיוון שאנו עושים שימוש בגרדיאנטים מנקודה זו לשם חישוב העדכונים השונים. תיאור מסודר של האלגוריתם המתקבל עבור גירסת ה- SD מובא בציור 14.8.

1. התחל את האלגוריתם בנקודה המקיימת את האילוצים - X_0.
2. באיטרציה ה- k - נניח כי בידינו הנקודה X_k אשר מקיימת את האילוצים, והמטריצה $A(X_k)$ - גרדיאנטי האילוצים הפעילים בנקודה זו.
3. בנה את $P(X_k)=I-A^T(X_k)\left[A(X_k)A^T(X_k)\right]^{-1}A(X_k)$.
4. חשב את הנקודה Y לפי - $Y=X_k-\mu P(X_k)\nabla f(X_k)$ (- צעד SD קלאסי).
5. עדכן את Y לפי $Y=Y-A^T(X_k)\left[A(X_k)A^T(X_k)\right]^{-1}h(Y)$ מספר פעמים עד לקבלת Y אשר מקיים את האילוצים. נקודת הסיום של תהליך איטרטיבי פנימי זה תהיה X_{k+1}.
6. עצור את האלגוריתם כאשר $\left\|P(X_k)\nabla f(X_k)\right\|\le\varepsilon$.

<u>ציור 14.8 - אלגוריתם הטלת הגרדיאנט</u>

מסתבר כי אלגוריתם זה בעל תכונות התכנסות דומות לאלו של ה- SD, כאשר הערכים העצמיים אשר קובעים את קצב ההתכנסות הם אלה של ההסיין המוכלל של הבעיה.

14.5 שיטת ה- Interior Point לפתרון בעיות תכנות ליניארי

בעיות תכנות ליניאריות (Linear Programming - LP) הוצגו בפרק 12 בהרחבה, שם גם הצגנו את הדרך הקלאסית לפתרונן - אלגוריתם ה- Simplex. אלגוריתם 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 רוסי 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:

$$\text{Min.} \quad C^T X \quad \text{S.T.} \quad AX = B \quad X \geq 0$$

יש לנו כאן שני סוגי אילוצים - אילוצים לינאריים הקלים לטיפול ע"י הטלה (נזכור כי בשיטת ההטלה, התוצאה המתקבלת מדוייקת ואינה כרוכה בקירוב), ואילוצי אי-שיוויון פשוטים הניתנים לטיפול ע"י שיטת המחסום. נציע את פונקציית המחסום הבאה:

$$G(X) = -\sum_{k=1}^{n} \log\{x_k\}$$

וכעת נעבור לפתרון הבעיה השקולה הבאה:

$$\text{Min.} \quad C^T X - \frac{1}{\text{Const}} \cdot \sum_{k=1}^{n} \log\{x_k\} \quad \text{S.T.} \quad AX = B$$

ואת בעיה זו נפתור ע"י איטרציה של אלגוריתם כמו SD, NSD, DNSD, או אפילו ניוטון, כשלאחריה אנו מטילים על אילוצי השיוויון את התוצאה לקבלת נקודה המקיימת את האילוצים.

נקודה שניה חשובה כאן היא מבנה הפונקציה המובאית למינימום. זו פונקציה שחישוב נגזרותיה ואיחסונם קל מאוד. זאת כיוון ש:

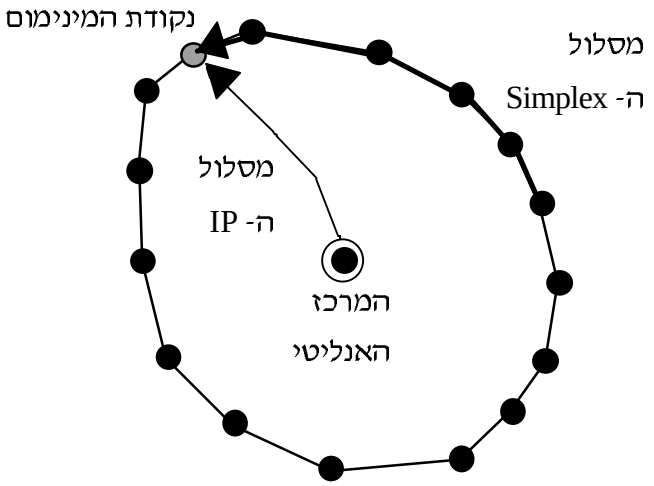
$$\nabla f(X) = \begin{bmatrix} C_1 - \dfrac{1}{\text{Const} \cdot x_1} \\ C_2 - \dfrac{1}{\text{Const} \cdot x_2} \\ \vdots \\ C_n - \dfrac{1}{\text{Const} \cdot x_n} \end{bmatrix} \qquad \nabla^2 f(X) = \frac{1}{\text{Const}} \begin{bmatrix} \dfrac{1}{x_1^2} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{x_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{1}{x_n^2} \end{bmatrix}$$

וההסיין הינו מטריצה אלכסונית - איחסונו קל והיפוכו טריויאלי. לכן ניתן להשתמש באלגוריתם מסדר 2, ולהאיץ את ההתכנסות. כיוון שהפונקציה המשולבת (יחד עם המחסום) היא קמורה ממש אין גם כל חשש להינעלות על מינימה לוקלית.

האיתחול לאלגוריתם נקבע ע"י קביעת הקבוע Const להיות כמעט אפס, ולפתור את הבעיה:

$$\text{Min.} \; -\frac{1}{\text{Const}} \cdot \sum_{k=1}^{n} \log\{x_k\} \quad \text{S.T.} \quad AX = B$$

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) אנו מתכנסים לנקודה 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. ציור 14.9 ממחיש הבדל זה.



ציור 14.9 - שיטת ה- Interior Point (IP) מול שיטת ה- Simplex

14.6 שיטות מבוססות לגרנז'

עבור בעיית אופטימיזציה עם אילוצי שיוויון - $\{\text{Min.} \; f(x) \; \text{S.T.} \; h(x)=0\}$, התנאים ההכרחיים אשר צריכים להתקבל בנקודת המינימום הם:

$$\nabla f(x) + \lambda^T \nabla h(x) = 0 \quad ; \quad h(x) = 0$$

במסגרת השיטות מבוססות הלגרנז' מוצע לפתור מערכת משוואות זו או פונקציות הקשורות אליה ולמצוא בדרך זו את הצמד - X^*, λ^*. 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:

$$\text{Min.}\quad M(x,\lambda)=\frac{1}{2}\left|\nabla f(x)+\lambda^T\nabla h(x)\right|^2+\frac{1}{2}\left|h(x)\right|^2=\frac{1}{2}\left|\nabla L(x,\lambda)\right|^2+\frac{1}{2}\left|h(x)\right|^2$$

תחת תנאים מסויימים (דרגה מלאה של המטריצה $\nabla h(X^*)$, היות מטריצת ההסיין המוכללת חיובית מוגדרת וכו') ניתן להראות כי לפונקציה המשולבת $M(x,\lambda)$ מינימום גלובלי בנקודה (X^*,λ^*). יתרה מזו - 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 במינימה לוקלית, ויש לאתחל את האלגוריתם בנקודה שונה.

נציע את האלגוריתם הבא למינימיזציה של הפונקציה $M(x,\lambda)$. יש לשים לב לעובדה שזה אינו אלגוריתם ה- SD אלא גירסה פשוטה אף ממנו:

$$\begin{bmatrix}X_{k+1}\\ \lambda_{k+1}\end{bmatrix}=\begin{bmatrix}X_k\\ \lambda_k\end{bmatrix}+\alpha_k\begin{bmatrix}-\nabla f(X_k)-\lambda_k^T\nabla h(X_k)\\ h(X_k)\end{bmatrix}=\begin{bmatrix}X_k\\ \lambda_k\end{bmatrix}+\alpha_k\begin{bmatrix}-\nabla L(X_k,\lambda_k)\\ h(X_k)\end{bmatrix}$$

נראה כי אלגוריתם זה מגשים ירידה בערך הפונקציה $M(x,\lambda)$. לצורך זה נייצג את פונקציה זו כטור טיילור:

$$\begin{aligned}
&M(X_{k+1},\lambda_{k+1})-M(X_k,\lambda_k)=\\
&\quad=\alpha_k\left[-\frac{\partial M(X_k,\lambda_k)}{\partial X_k}\nabla L^T(X_k,\lambda_k)+\frac{\partial M(X_k,\lambda_k)}{\partial \lambda_k}h^T(X_k)\right]=\\
&\quad=-\alpha_k\left[\nabla L^T(X_k,\lambda_k)H(X_k,\lambda_k)+h^T(X_k)\nabla h(X_k)\right]\nabla L(X_k,\lambda_k)+\\
&\qquad+\alpha_k h^T(X_k)\nabla h(X_k)\nabla L(X_k,\lambda_k)=-\alpha_k\nabla L^T(X_k,\lambda_k)H(X_k,\lambda_k)\nabla L(X_k,\lambda_k)\le 0
\end{aligned}$$

כך שהאיטרציות המוצעות מבטיחות ירידה בערך הפונקציה המשולבת. המטריצה H הינה ההסיין המוכלל של פונקציית הלגרנז'יאן. סיכום של אלגוריתם זה מובא בציור 14.10.

גישה מסדר שני לפתרון מערכת המשוואות - $\nabla L(x,\lambda)=0\ ;\ h(x)=0$ היא אלגוריתם ניוטון-רפסון הרב-מימדי. בגישה זו מבוצעת ליניאריזציה של משוואות אלה סביב הנקודה הנוכחית, ונמצאת הנקודה החדשה אשר פותרת את המערכת הליניארית. ליניאריזציה מניבה את המערכת:

$$\begin{gathered}
\nabla L(X_k,\lambda_k)+H(X_k,\lambda_k)(X_{k+1}-X_k)+\nabla h(X_k)(\lambda_{k+1}-\lambda_k)=0\\
h(X_k)+\nabla^T h(X_k)(X_{k+1}-X_k)=0
\end{gathered}$$

1. אתחל את האלגוריתם ב- $\{X_0,\lambda_0\}$ כלשהם.
2. באיטרציה ה- k חשב את ערך פונקציות האילוצים - $h(X_k)$, ואת ערך הפונקציה $\nabla L(X_k,\lambda_k)$.
3. בצע עדכון לפתרון על-פי המשוואה:

$$\begin{bmatrix} X_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} X_k \\ \lambda_k \end{bmatrix} + \alpha_k \begin{bmatrix} -\nabla L(X_k,\lambda_k) \\ h(X_k) \end{bmatrix}$$

כאשר α_k נקבע לפי חיפוש על הישר המוגדר כאן על הפונקציה $M(x,\lambda)$

4. אם מתקבל כי $\|X_{k+1}-X_k\|^2+\|\lambda_{k+1}-\lambda_k\|^2>\varepsilon_0$ - חזור ל- 2 ובצע איטרציה נוספת.
5. אם מתקבל כי $\|X_{k+1}-X_k\|^2+\|\lambda_{k+1}-\lambda_k\|^2\le\varepsilon_0$ אך $M(x,\lambda)>\varepsilon$ חזור להתחלה עם איתחול שונה.
6. אם מתקבל כי $\|X_{k+1}-X_k\|^2+\|\lambda_{k+1}-\lambda_k\|^2\le\varepsilon_0$, וכן - $M(x,\lambda)\le\varepsilon$ - עצור את האלגוריתם.

<u>ציור 14.10 - אלגוריתם מסדר ראשון לפתרון בעיית אופטימיזציה עם אילוצי שיויון בגישת לגרנז'</u>

ומכאן נציע את המשוואה האיטרטיבית:

$$\begin{bmatrix} X_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} X_k \\ \lambda_k \end{bmatrix} + \alpha_k \begin{bmatrix} H(X_k,\lambda_k) & \nabla h(X_k) \\ \nabla^T h(X_k) & 0 \end{bmatrix}^{-1} \begin{bmatrix} -\nabla L(X_k,\lambda_k) \\ -h(X_k) \end{bmatrix}$$

הוספנו כאן מכפלה בסקלר אשר אות יש למצוא בגישות חיפוש על ישר לשם מציאת מינימום על הישר המוצע עבור הפונקציה $M(x,\lambda)$. באופן זה נבטיח כי אי-דיוקים של הליניאריזציה לא יביאו לסטיה חמורה ולהתבדרות. ניתן להראות כי הכיוון המוצע באיטרציה הנ"ל הוא כיוון ירידה, בגישה דומה לגישה שננקטה עבור השיטה הקודמת, וכי אלגוריתם זה מתכנס. הוצעו עבור אלגוריתם זה גישות ניוטוניות למחצה אשר מחליפות את H בקירוב שלו הנבנה מנגזרות מסדר ראשון - לא נראה

גישה זו במסגרת זו. ציור 14.10 מתאר את גישה זו באותה מידה, כשיש להחליף את משוואת האיטרציה.

14.7 סיכום

בפרק זה ראינו גישות מעשיות לפתרון בעיות אופטימיזציה הכוללות אילוצים. ראינו את גישות המחיר והמחסום אשר בוחרות להמיר את הבעיה לסידרת בעיות פרמטריות חדשות ללא אילוצים, בהן הנעלם בעל מימד n. גישת ההטלה לתת-מרחב האילוצים מציעה את אלגוריתם ה-SD עם הטלה אשר מבטיחה את קיום האילוצים בכל איטרציה. לבסוף ראינו גישות המציעות מינימיזציה של פונקציה משולבת הכוללת את כופלי לגרנז' כחלק מנעלמיה.

למרות מראיהן השונה של הגישות הנ"ל, יש קוים רבים של דימיון ביניהן. דימיון זה בא לידי ביטוי בשיאו בסוגיית קצב ההתכנסות של אלגוריתמים אלו. מסתבר כי אין גישה אחת טובה מהאחרות באופן שיהפוך אותה לכדאית ביותר - כל השיטות מתכנסות באופן שתלוי במטריצת ההסיין המוכללת בדרך זו או אחרת.

כפי שנאמר בפתיחת פרק זה, ישנן בתחום זה של בעיות עם אילוצים תוצאות רבות ומעניינות אשר תצדקנה קורס נפרד. פרק זה מהווה סקירה כללית ומתן תחושה ולא יותר בהתיחס לדיון בבעיית אופטימיזציה עם אילוצים.