מבוא לאופטימיזציה

236330

הרצאות ווידיאו, אביב 2010

פרופסור חבר מיכאל ציבולבסקי

נכתב ע"י רמי נודלמן

אביב 2015

תוכן עניינים

הרצאה 1a – אלגוריתמים של אופטימיזציה לא-לינארית

הגדרות:

אופטימיזציה ללא אילוצים:

אופטימיזציה של פונקציה מסויימת $f:\mathbb{R}^n \to \mathbb{R}$ ללא אילוצים, זה למעשה מציאת הנקודה $\overline{x}$ אשר תביא את ערך הפונקציה $f\left(\overline{x}\right)$ למינימום שלה, כלומר:

$$\underset{\overline{x}\in\mathbb{R}^n}{\arg\min}\ f\left(\overline{x}\right)$$

כאשר אין אילוצים על התחום של הנקודות $\overline{x}$ (בתנאי שהן נמצאות בתחום ההגדרה של הפונקציה). בשלב זה אנחנו מניחים כי $f\left(\overline{x}\right)$ היא פונקציה רציפה ו"חלקה" (כלומר גזירה).

אופטימיזציה עם אילוצים:

באופטימיזציה עם אילוצים אנחנו מנסים למצוא את הנקודה $\overline{x}$ שתביא את הפונקציה $f:\mathbb{R}^n \to \mathbb{R}$ לערך המינימום, כלומר:

$$\underset{\overline{x}\in\mathbb{R}^n}{\arg\min}\ f\left(\overline{x}\right)$$

אך כעת אנחנו נותנים הגבלות על הנקודות האפשריות (מחפשים נקודות רק בתחומים מוגדרים מסויימים), כלומר מגבילים את התחום ע"י אילוצים כלשהם. ואנחנו מניחים כי $f\left(x\right)$ היא פונקציה רציפה ו"חלקה" (גזירה ברציפות) וכן האילוצים הם למשל:

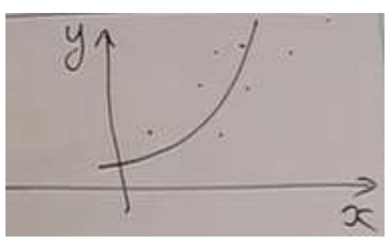
$$g_i\left(\overline{x}\right)\ge 0\ ,\ i=1,\dots,m$$
$$h_j\left(\overline{x}\right)=0\ ,\ j=1,\dots,l$$

האילוץ הראשון הוא אילוץ שנוצר ע"י אי-שיוויון (Inequality constrain) והשני נוצר על ידי שיוויון (Equality Constrain).

הקדמה ודוגמאות:

בקורס אנחנו נעסוק בחיפוש התנאים האופטימליים, כלומר חיפוש התנאים על הפונקציה שעבורם נקבל את הפתרון האופטימלי וכן נעסוק באלגוריתמים איטרטיביים (שמתבססים על חישובים נומריים) שישיגו את הפתרון האופטימלי (בין אם האלגוריתם ייתן לנו את הפתרון המדוייק או רק נוכל להגיע אליו בצורה "גבולית", בצורה ששואפת לפתרון האופטימלי).

דוגמה: (דוגמה אופטימיזציה ללא אילוצים) (Parametric Regression)



נניח ויש לנו מידע אמפירי, שבו x הוא הארגומנט ואילו y אלו המדידות , כלומר יש לנו נקודות על מערכת הצירים. כעת, אנחנו מתבוננים בנקודות הללו ומניחים שהקשר בין x לבין y הוא ריבועי, כלומר $y=f\left(x\right)=x^2$, אך אנחנו רואים כי יש רעש/סטייה בין מה שאנחנו רוצים לראות במדוייק לבין תוצאות הניסוי.

ננסח את הבעיה בצורה פורמלית:

בהינתן קבוצה של זוגות $\left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{M}$ אנחנו רוצים למצוא פונקציה ריבועית $f\left(x\right)$ עם ווקטור מקדמים $\bar{w}$:

$$f\left(x,\bar{w}\right)=w_0+w_1x+w_2x^2$$

כעת אנחנו רוצים לייצג את הבעיה כבעית אופטימיזציה והאופטימיזציה היא למעשה מציאת הוקטור $\bar{w}$ אשר יביא לנו שגיאה מינימלית – מהי השגיאה? הביטוי לשגיאה הוא הביטוי שאנחנו נגדיר עבור כל בעיה לגופה. נניח ואנחנו רוצים להתבונן בשגיאה הריבועית, כלומר סכום ריבועי ההפרשים בין הפונקציה שאנחנו מציעים לבין המדידות בניסוי:

$$\min_{\bar{w}}\sum_{i=1}^{M}\left(f\left(x^{(i)},\bar{w}\right)-y^{(i)}\right)^2$$

כאשר $y^{(i)}$ זו המדידה ה- i-ית.

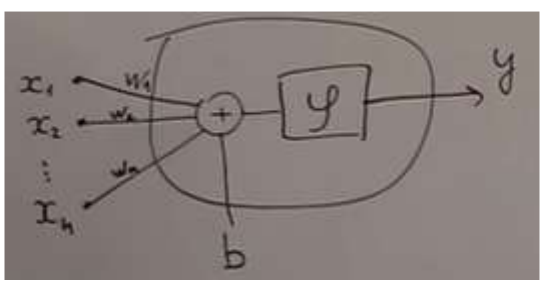
במקרה הכללי יותר, במקום להניח כי לפונקציה $f\left(x\right)$ יש מקדמים קבועים, ניתן להניח כי המקדמים הם בעצמם פונקציות של הארגומנט x, למשל:

$$f\left(x,\bar{w}\right)=\sum_{k}w_kg_k\left(x\right)$$

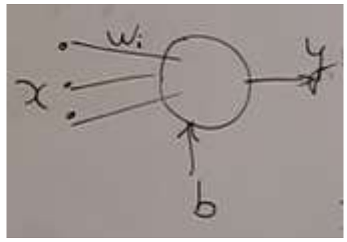
כאשר הפונקציות $\left\{g_k\right\}$ ידועות. במקרה זה, הבעיה נקראת Linear Regression Problem. אמנם הפונקציות $\left\{g_k\right\}$ אינן בהכרח לינאריות אך הפונקציה $f\left(x\right)$ היא לינארית ביחס לווקטור $\bar{w}$.

לעיתים נתקלים בבעיות בהן הפונקציה $f\left(x\right)$ היא לא לינארית ביחס לווקטור $\bar{w}$, כלומר כל אחד מהרכיבים של הווקטור $\bar{w}$ הוא לא מקדם קבוע אלא פונקציה בעצמו, כלומר $\bar{w}=\bar{w}\left(x\right)$.

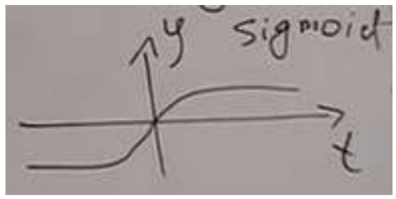
<u>דוגמה:</u> (Non Linear Regression) – למשל Neural Networks (מערכות עצבים)



נניח ואנחנו רוצים למצוא אופטימזציה של פונקציה בעלת מספר משתנים, $\bar{x}=\left\{x_1,x_2,...,x_n\right\}$ כאשר b הוא קבוע. נתייחס לרכיבים של ווקטור $\bar{x}$ כאל "כניסות למערכת", כאשר המערכת הוא נוירון בודד והפלט של נוירון בודד הוא פונקציה של הכניסות שלו. נניח כי הפלט הוא סכום של כניסות מסויימות כאשר לכל כניסה יש משקל שונה לאחר מכן, הסכום הזה נכנס לפונקציה φ ואנחנו מקבלים פלט y. ניתן לכתוב את המערכת הנ"ל בצורה הבאה:



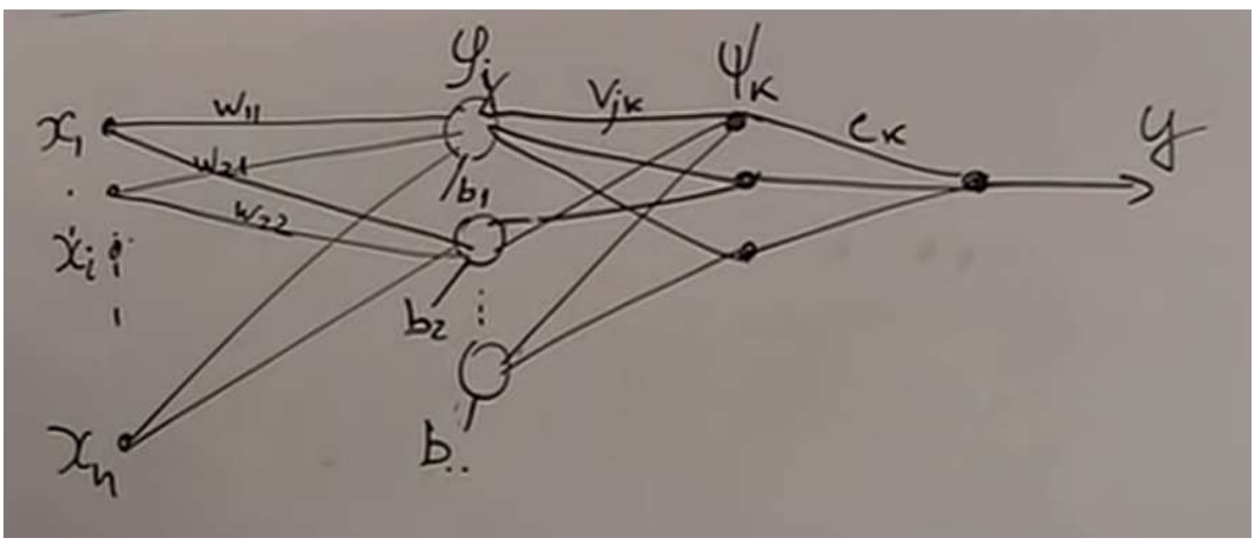
פונקציה זו, שהנוירון פולט נקראת sigmoid והיא נראת כמו $\arctan(x)$:



ובצורה פורמלית ניתן לתאר את הפלט של הנוירון בצורה הבאה:

$$y=\varphi\left(\bar{w}^T\cdot\bar{x}+b\right)$$

נתבונן כעת ברשת של נוירונים. נניח כי יש לנו "כניסות" $\bar{x}=\{x_1,x_2,...,x_n\}$ וכן הנוירון ה- i מיוצג ע"י הפונקציה φ_i. כלומר נוירון מקבל את כל הכניסות אבל במשקלים שונים. לאחר הרמה הראשונה של הנוירונים, ייתכן ויש עוד נוירונים אך אנחנו נפשט את המודל ונניח כי יש לנו רק רמה אחת של נוירונים ולאחר מכן יש לנו במקום נוירונים, יש לנו מערכות לינאריות, ψ_k ונקבל:



כלומר פלט הרשת כולה הוא:

$$y(\bar{x})=\sum_k\psi_k\left(\sum_j\varphi_j\left(\left(\sum_i x_i w_{ij}\right)+b_j\right)\right)$$

נסמן את הרכיב של ווקטור $\bar{u}$ להיות : $u_j = \left(\sum_i x_i w_{ij}\right) + b_j$ ובכתיב מטריצי נקבל : $u_j = W^T \cdot \bar{x} + b_j$, וכן נקבל:

$$y(\bar{x}) = \sum_k \psi_k \left(\sum_j \varphi_j \left(u_j(\bar{x}) \right) \right)$$

נוכל לכתוב זאת גם בצורה הבאה:

$$\varphi(u) \triangleq \begin{pmatrix} \varphi_1\left(u_1(\bar{x})\right) \\ \varphi_1\left(u_2(\bar{x})\right) \\ \vdots \end{pmatrix}$$

ולכן נקבל:

$$y = f\left(\bar{x}, W, \bar{b}, V, \bar{c}\right) = c^T \psi_k \left(V^T \varphi_j \left(W^T \cdot \bar{x} + b \right) \right)$$

(כאשר אותיות גדולות הן מטריצות)

כלומר קיבלנו מערכת שאנחנו רוצים לבצע עליה אופטימיזציה ביחס למשתנים: $W, \bar{b}, V, \bar{c}$.

כעת נניח ויש לנו שוב זוגות של קואורדינטות ונתונים של ניסוי אך הפעם הארגומנט הוא ווקטור ותוצאת הניסוי היא סקלר כלומר:

$$\left\{ \bar{x}^{(i)}, y^{(i)} \right\}_{i=1}^{M}$$

ואנחנו נרצה למצוא פונקציה (רשת נוירונים) כזו שעבור סט נתונים מסויים, הפונקציה תשערך את המדידות $y^{(i)}$ באופן הקרוב ביותר. ובאופן פורמלי, נרצה למצוא למשל את סכום השגיאות הריבועיות המנימלי:

$$\min_{W,\bar{b},V,\bar{c}} \sum_{i=1}^{M} \left[f\left(\bar{x}, W, \bar{b}, V, \bar{c}\right) - y^{(i)} \right]^2$$

זו כמובן בעית אופטימיזציה לא לינארית ביחס לפרמטרים $W, \bar{b}, V, \bar{c}$, פרט ביחס לווקטור $\bar{c}$, שביחס אליו הבעיה היא דווקא כן לינארית שכן נזכור כי הווקטור $\bar{c}$ נמצא מחוץ לפונקציה ψ. בקורס זה נלמד, איך למצוא אופטימיזציה גם לסוג בעיות כאלה.

<u>דוגמה</u>: בעית אופטימיזציה עם אילוצים (Resource Assignment)

למשל, יש לנו מספר תחנות כוח ומספר לקוחות ויש מגבלות על כמות הכוח שכל תחנה מסוגלת לספק וכן לכל לקוח יש דרישות שיש לקיים וגם יש עלויות שונות עבור ההובלה/שינוע של האנרגיה מתחנה מסויימת ללקוח. אנחנו נרצה לבצע אופטימיזציה על העלות של הבעיה, כלומר למצוא פתרון אופטימלי של חלוקת האנרגיה בין תחנות הכוח השונות לבין הלקוחות השונים במחיר הזול ביותר.

נתונים:

יש לנו M תחנות כוח. s_m זה הכוח המקסימלי שתחנת כוח m מסוגלת לייצר.

יש לנו N לקוחות. c_n זו דרישת הכוח של הלקוח ה- n -י.

p_{mn} זו העלות של שינוע יחידת אנרגיה מתחנת כוח m ללקוח n.

המטרה:

למצוא את מטריצת הנעלמים X שבה $x_{m,n}$ מייצג את כמות יחידות האנרגיה שתחנה m מספקת ללקוח n.

באופן פורמלי הבעיה היא:

$$\min_X \sum p_{mn} \cdot x_{mn}$$

האילוצים הם:

- כל תחנה מספקת לכל היותר את כל הכוח שהיא מסוגלת: $\sum_{n=1}^{N} x_{mn} \le s_m \ : \ \forall 1 \le m \le M$.
- כל לקוח מקבל את כל האנרגיה שהוא דרש: $\sum_{m=1}^{M} s_{mn} = c_n \ : \ \forall 1 \le n \le N$.
- כל תחנה מספקת אנרגיה ללקוח (ולקוח לא מספק אנרגיה לתחנת הכוח) כלומר מתקיים: $x_{mn} \ge 0$

בבעיה זו אנחנו רואים כי יש אילוצים מסוג אי-שיוויון וגם שיוויון.

כל האילוצים הנ"ל הם לינאריים ביחס ל- X ולכן זו בעית אופטימיזציה לינארית, נקראת: Linear Programming Problem.

הרצאה 1b – חזרה על אלגברה לינארית

$\mathbb{R}$ - מרחב אוקלידי ממימד 1 (הציר הממשי).

$\mathbb{R}^n$ - מרחב אוקלידי ממימד n.

וקטורים במרחב $\mathbb{R}^n$ הם מהצורה (וקטור עמודה): $\bar{x}=\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\in\mathbb{R}^n$.

<u>מרחב וקטורי (לינארי):</u>

יהי S מרחב וקטורי וכן $\bar{x},\bar{y}\in S$ אז מתקיים: $\alpha\bar{x}+\beta\bar{y}\in S$ לכל $\alpha,\beta\in\mathbb{R}$.

<u>מרחב וקטורי אפיני:</u>

נאמר כי Q מרחב אפיני וכן $\bar{b}\in S$ אם מתקיים:

$$Q=S+\bar{b}=\left\{\bar{y}=\bar{b}+\bar{x}\ \mid\ \bar{x}\in S\right\}$$

אם למשל המרחב S הוא ממימד 2 אז אפשר לומר כי Q הוא מרחב אפיני אם הוא שווה למרחב S אך מוזז ע"י הווקטור $\bar{b}$.

מרחב אפיני מקיים את תכונת הלינאריות, כלומר אם $x,y\in Q$ אז:

$$\alpha\bar{x}+\beta\bar{y}\in Q$$

כאשר האילוץ על α,β הוא: $\alpha+\beta=1$ לכל α,β.

<u>הגדרה:</u> (נורמה וקטורית)

נניח כי $\bar{x}\in\mathbb{R}^n$. הנורמה של הווקטור $\bar{x}$ היא פונקציה שמעבירה אלמנטים ממרחב $\mathbb{R}^n$ למרחב $\mathbb{R}$. הנורמה של $\bar{x}\in\mathbb{R}^n$ מסומנת להיות $\|\bar{x}\|$.

<u>התכונות של כל נורמה:</u>

1) לכל $\bar{x}\in\mathbb{R}^n$ מתקיים: $\|\bar{x}\|\geq 0$.
2) לכל $\bar{x}\in\mathbb{R}^n,\alpha\in\mathbb{R}$ מתקיים: $\|\alpha\bar{x}\|=|\alpha|\cdot\|\bar{x}\|$
3) $\|\bar{x}\|=0\Leftrightarrow\bar{x}=\bar{0}$.
4) אי-שיוויון המשולש: $\|\bar{x}+\bar{y}\|\leq\|\bar{x}\|+\|\bar{y}\|$.

<u>דוגמאות לנורמות:</u>

- נורמה אוקלידית ($\mathbb{R}^n$):

$$\forall\bar{x}\in\mathbb{R}^n:\ \|\bar{x}\|_2=\sqrt{\bar{x}^T\cdot\bar{x}}=\sqrt{\sum_{i=1}^{n}|x_i|^2}$$

- ההכללה של נורמה אוקלידית למרחב $\mathbb{R}^p$ ($l_p \ norm$) :

$$\forall \bar{x} \in \mathbb{R}^p: \ \ \|\bar{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

מקרה פרטי חשוב של $l_p \ norm$ הוא l_1 וזה למעשה:

$$\forall \bar{x} \in \mathbb{R}^1 = \mathbb{R}: \ \ \|\bar{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

ומקרה פרטי נוסף של $l_p \ norm$ הוא l_∞ וזה למעשה:

$$\forall \bar{x} \in \mathbb{R}^1 = \mathbb{R}: \ \ \|\bar{x}\|_\infty = \lim_{x\to\infty} \|\bar{x}\|_p = \max_i |x_i|$$

<u>נורמה של מטריצות:</u>

יש כמה סוגים של נורמות של מטריצות. אחד מהם הוא דומה לנורמה אוקלידית, שמחשיב למעשה את המטריצה כווקטור אחד ארוך. סוג שני נקרא הנורמה המטריצית לפי פרוביני (Frobeinus Norm) :

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \ , \ \ \|A\|_F \triangleq \sqrt{\sum_{i,j} \left(a_{ij}\right)^2}$$

ניתן לראות את ההגדרה האחרונה גם בצורה אחרת וקומפקטית יותר:

$$Trace\left(A^T A\right)$$

(כאשר $Trace(A) = \sum_{i=1}^{n} a_{ii}$)

ניתן גם להגדיר סוג אחר של נורמה על מטריצות אם נחשוב על מטריצות כעל אופרטורים לינארים. מטריצות הן למעשה אופרטורים שמעבירים וקטורים ממרחב אחד לאחר ובין היתר משנים את הנורמה של הווקטור עליו הן פועלות. נגדיר את הנורמה של מטריצה להיות הנורמה המקסימלית שווקטור באורך יחידה שעליו פועלת המטריצה מקבל (יוסבר בהמשך). נורמה כזו נקראת Induced matrix norm או במילים אחרות Operator matrix norm. בצורה פורמלית:

$$\bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \ , \ \ \|A\|_{p,q} \triangleq \max_{\|x\|_p = 1} \|Ax\|_q$$

כלומר הנורמה $\|A\|_{p,q}$ היא הנורמה המקסימלית של הווקטור $A\bar{x}$ תחת נורמה q כאשר $\bar{x}$ הוא כל ווקטורי היחידה האפשריים.

<u>מכפלה פנימית של מטריצות:</u>

נניח A, B מטריצות ריבועיות מאותו סדר, אז המכפלה הפנימית של A, B :

$$\langle A, B \rangle \triangleq \sum_{i,j} a_{ij} b_{ij} = Trace\left(A^T B\right) = Trace\left(BA^T\right)$$

(הערה: המכפלה הפנימית הנ"ל מוגדרת בדומה לאיך שמוגדרת מכפלה פנימית בין שני ווקטורים ולמעשה מכפלה פנימית זו מתייחסת למטריצות כאל ווקטורים "ארוכים")

הערה: מה שמאפשר את שני המעברים האחרונים היא תכונת ההזזה הציקלית תת אופרטור ה-Trace:

$$Trace(VW) = Trace(WV)$$

<u>ערכים עצמיים של מטריצות:</u>

תהיי A מטריצה מסדר $n \times n$ וכן $\bar{x} \in \mathbb{R}^n$ וגם $\lambda \in \mathbb{R}$. אם מתקיים:

$$A\bar{x} = \lambda\bar{x}$$

נאמר כי $\bar{x}$ הוא ווקטור עצמי (ו"ע) וכן λ הוא ערך עצמי (ע"ע). המשמעות היא שכאשר מטריצה A פועלת על ווקטור $\bar{x}$ היא לא משנה את כיוונו אלא רק משפיעה על מגמתו ו/או ערכו.

למטריצה מסדר $n \times n$ יש לכל היותר n ו"ע-ים בלתי תלויים ו- n ע"ע.

נרשום את הו"ע-ים ואת הע"ע-ים בצורה קומפקטיבית וכן נניח כי כל הו"ע הם בלתי תלויים. נרשום את הו"ע-ים של מטריצה S בתור עמודות המטריצה:

$$s_i = \begin{pmatrix} s_{1i} \\ \vdots \\ s_{ni} \end{pmatrix} \qquad \Rightarrow \qquad S = \begin{pmatrix} \bar{s}_1 & \cdots & \bar{s}_n \end{pmatrix}$$

וכן נרשום:

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

ולכן נוכל לרשום:

$$AS = S\Lambda$$

ואם הו"ע-ים הם בלתי תלויים אז המטריצה S מקיימת: $\det(S) \neq 0$ ולכן יש למטריצה S מטריצה הופכית ולכן נוכל לכפול במטריצה ההופכית ונקבל:

$$AS = S\Lambda \qquad \Rightarrow \qquad A = S\Lambda S^{-1}$$

בנוסף, אם המטריצה A היא סימטרית אז כל הו"ע-ים של A הם אורתוגונליים ולכן נקבל:

$$S \cdot S^T = I$$

(כאשר מנרמלים את הו"ע העצמיים להיות בגודל של ווקטורי יחידה)

מהמשוואה האחרונה אנחנו רואים כי: $S^T = S^{-1}$ ולכן נוכל במקרה שמטריצה A היא סימטרית לרשום:

$$A = S\Lambda S^{-1} = S\Lambda S^T$$

דוגמה:

באמצעות ע"ע נוח לבצע חישוב מסויימים על מטריצות – למשל חישוב מטריצה הופכית.

אם נניח כי $A = S\Lambda S^{-1}$ אז זה אומר כי $A^{-1} = S\Lambda^{-1}S^{-1}$ כי ניתן לראות כי מתקיים:

$$AA^{-1} = \left(S\Lambda S^{-1}\right)\left(S\Lambda^{-1}S^{-1}\right) = S\Lambda\left(S^{-1}S\right)\Lambda^{-1}S^{-1} = S\left(\Lambda\Lambda^{-1}\right)S^{-1} = SS^{-1} = I$$

בנוסף אפשר לחשב את החזקות של מטריצה. למשל:

$$A^2 = A \cdot A = S\Lambda S^{-1} \cdot S\Lambda S^{-1} = S\Lambda^2 S^{-1}$$

וכיוון שמטריצה Λ היא אלכסונית זה מאד קל כיוון שפשוט מעלים בחזקה כל אחד מהרכיבים שלה:

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \quad \Rightarrow \quad \Lambda^p = \begin{pmatrix} (\lambda_1)^p & & 0 \\ & \ddots & \\ 0 & & (\lambda_n)^p \end{pmatrix}$$

פונקציות מטריציות: (בהקבלה לפונקציות סקלריות)

אנחנו רגילים שפונקציות פועלות על סקלרים או לכל היותר ווקטורים אך אין מניעה שתהיינה פונקציות שפועלות על מטריצות.

נניח ויש לנו פונקציה סקלרית כלומר $\varphi: \mathbb{R} \to \mathbb{R}$ שמוגדרת באופן הבא:

$$\varphi(t) = \sum_i c_i t^i \quad , \qquad c_i \in \mathbb{R},\ 0 \le i \le \infty$$

אז אפשר להגדיר באופן דומה פונקציה $\varphi_A : n \times n \to n \times n$:

$$\varphi(A) = \sum_i c_i A^i, \qquad c_i \in \mathbb{R},\ 0 \le i \le \infty$$

כיוון שראינו שאפשר לחשב חזקות של מטריצה באמצעות הע"ע של המטריצה ובאמצעות הו"ע-ים של המטריצה אפשר לרשום:

$$\varphi(A) = \sum_i c_i A^i = \sum_i c_i S\Lambda^i S^{-1} = S\left(\sum_i c_i \Lambda^i\right)S^{-1} = S\begin{pmatrix} \sum_i c_i(\lambda_1)^i & & 0 \\ & \ddots & \\ 0 & & \sum_i c_i(\lambda_n)^i \end{pmatrix}S^{-1}$$

$$= S\begin{pmatrix} \varphi(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & \varphi(\lambda_n) \end{pmatrix}S^{-1} = S\varphi(\Lambda)S^{-1}$$

הגדרה: חיוביות ואי-שליליות של מטריצות סימטריות

נאמר כי המטריצה A חיובית, $A \succ 0$ (Positive definite), או אי-שלילית, $A \succeq 0$ (Positive semi-defenite) אמ"מ

$$\forall \bar{x} \neq \bar{0}: \quad \bar{x}^T A \bar{x} > 0 \quad \left(\text{Positive defenite} \right)$$
$$\forall \bar{x} \neq \bar{0}: \quad \bar{x}^T A \bar{x} \geq 0 \quad \left(\text{Positive (semi)defenite} \right)$$

<u>משפט</u>:

מטריצה היא חיובית (אי-שלילית) אמ"מ כל הערכים העצמיים של המטריצה A הם חיוביים (אי-שליליים).

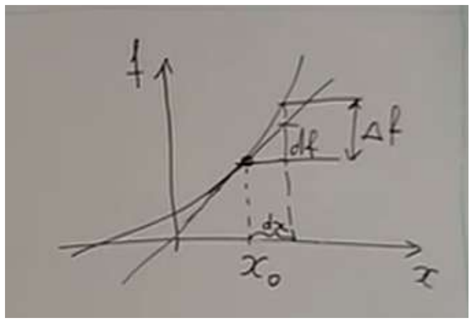
<u>הוכחה</u>:

$$\forall \bar{x} \neq \bar{0}: \quad \bar{x}^T A \bar{x} = \underbrace{\bar{x}^T S}_{\equiv \bar{y}^T} \Lambda \underbrace{S^T \bar{x}}_{\equiv \bar{y}} = \bar{y}^T \Lambda \bar{y} = \sum_i \lambda_i \left(y_i \right)^2 \geq 0 \qquad \Leftrightarrow \qquad \forall \lambda_i \geq 0$$

הרצאה 02-03 – אלגוריתמים של אופטימיזציה לא לינארית. נגזרות של פונקציות מרובות משתנים: גרדיאנט והסיאן

תזכורת: פונקציה של משתנה יחיד

נניח כי $f\left(x\in\mathbb{R}\right):\mathbb{R}\to\mathbb{R}$.



כאשר נשאיף את dx לאפס, כלומר $dx\to 0$ אז נוכל לומר כי מתקיים:

$$\Delta f = df + o\left(dx\right)$$

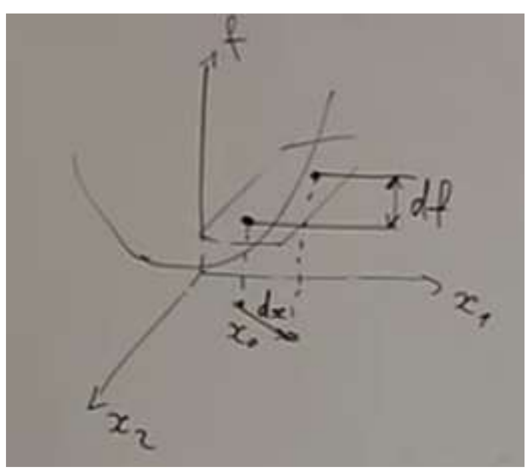
ואנחנו קוראים לביטוי df דיפירנציאל ולכן אפשר לומר כי הדיפרנציאל של פונקציה f הוא החלק הלינארי של השינוי בפונקציה f .

כמו כן, אנחנו יודעים מחדו"א שהמשיק מקיים:

$$df = f'\left(x_0\right)dx$$

פונקציה של מספר משתנים:

נניח כי $f\left(x_1,x_2\right):\mathbb{R}^2\to\mathbb{R}$ ונניח שאנחנו מסתכלים על המישור המשיק לנקודה $\left(x_1,x_2\right)$ ואנחנו מתקדמים מנקודה $\left(x_1,x_2\right)$ ע"י וקטור $d\overline{x}=\left(dx_1,dx_2\right)$:



מחדו"א 2 ידוע כי הדיפרנציאל של פונקציה דיפרנציאבילית מרובת משתנים, $f\left(x_1, x_2, \dots, x_n\right)$, הוא הסכום:

$$df\left(\overline{x}\right) = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \dots + \frac{\partial f}{\partial x_n} dx_n = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i$$

וניתן לרישום בצורה קומפקטית יותר כמכפלה פנימית:

$$df\left(\overline{x}\right) = \left\langle \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}, \begin{pmatrix} dx_1 \\ \vdots \\ dx_n \end{pmatrix} \right\rangle = \left\langle \overline{\left(\frac{\partial f}{\partial x}\right)}, d\overline{x} \right\rangle$$

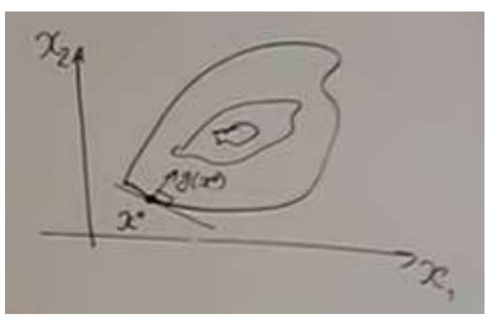
ואפשר להבחין למעשה מתקיים כי ה<u>וקטור</u> $\overline{\left(\frac{\partial f}{\partial x}\right)}$ הוא למעשה $\nabla_{\overline{x}} f\left(\overline{x}\right)$ שנסמנו $\overline{g}\left(\overline{x}\right)$ וזה למעשה <u>הגרדיאנט של הפונקציה</u> $f\left(\overline{x}\right)$. לכן ניתן לרשום:

$$\boxed{df\left(\overline{x}\right) = \left\langle \overline{g}\left(\overline{x}\right), d\overline{x} \right\rangle}$$

הגדרה זו מאד שימושית כי היא תעזור לנו לבנות גרדיאנטים של פונקציות מסובכות שניתקל בהן במהלך הקורס, כי לבנות את הגרדיאנט לפי הנגזרות החלקיות יהיה מאד מסובך.

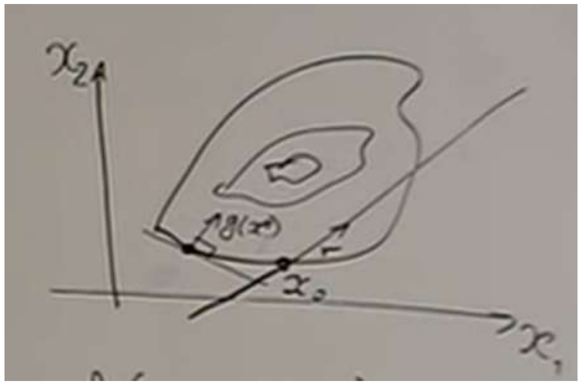
<u>נגזרת כיוונית בכיוון כלשהו:</u>

נתבונן שוב בפונקציה של שני משתנים: $f\left(x_1, x_2\right): \mathbb{R}^2 \to \mathbb{R}$ ונצייר אותה בצורה של מפת טופוגרפית (הקווים בשרטוט הם קווי שווי ערך/גובה):



אנחנו נראה כי הגרדיאנט (שהוא בעצמו <u>וקטור</u>) מצביע על כיוון העליה החדה ביותר בערכי הפונקציה, וכן הוא מאונך למישור המשיק לנקודה.

אנחנו בעיקר נתעניין בהתנהגות הפונקציה כאשר אנחנו מתחילים בנקודה מסויימת $\overline{x}_0$ בתחום ומתקדמים לאורך וקטור $\overline{r}$.



כלומר אנחנו רוצים להעריך את הפונקציה f בנקודות $\bar{x}_0+\alpha\bar{r}$ (כאשר $\alpha\in\mathbb{R}$) ונוכל למעשה להגדיר פונקציה $\varphi(\alpha):\mathbb{R}\to\mathbb{R}$ בצורה הבאה:

$$\varphi(\alpha)\triangleq f\left(\bar{x}_0+\alpha\bar{r}\right)$$

ואנחנו נרצה למצוא את הנגזרת של הפונקציה $\varphi(\alpha)$ שהגדרנו בנקודה $\bar{x}_0$ (כלומר בנקודה ההתחלתית שבה אנחנו מתבוננים בפונקציה $f\left(\bar{x}\right)$), כלומר למצוא את:

$$\varphi'(\alpha)\Big|_{\alpha=0}$$

הביטוי $\varphi'(\alpha)$ נקרא <u>הנגזרת הכיוונית</u> של $f\left(\bar{x}_0\right)$ בכיוון $\bar{r}$ ונסמנה להיות:

$$\boxed{\varphi'(\alpha)\Big|_{\alpha=0}=\frac{\partial}{\partial\alpha}f\left(\bar{x}_0+\alpha\bar{r}\right)\triangleq f_{\bar{r}}\left(\bar{x}_0\right)}$$

זוהי תוצאה מאד חשובה בקורס שנשתמש בה רבות.

<u>דיפרנציאל וגרדיאנט של פונקציה מרובת משתנים:</u>

השאלה שתעסיק אותנו כעת זה האם בהינתן הגרדיאנט בנקודה מסויימת $\bar{g}\left(\bar{x}_0\right)\equiv\nabla f\left(\bar{x}_0\right)$ האם אנחנו יכולים לדעת מהי הנגזרת הכיוונית בכיוון כלשהו $\bar{r}$, כלומר האם ע"י הגרדיאנט ניתן לדעת מה היא $f_{\bar{r}}\left(\bar{x}_0\right)$?

ניתן למצוא את התשובה באמצעות הצבה ישרה למשוואות שמצאנו עד כה אך גם ניתן להשתמש בהגדרה של הדיפרנציאל של הפונקציה $f\left(\bar{x}\right)$.

לפי מה שראינו עבור משתנה יחיד:

$$df\left(\bar{x}\right)=f'\left(x_0\right)d\bar{x}$$

מתקיים באופן דומה עבור פונקציה של מספר משתנים:

$$\boxed{df\left(\bar{x}\right)=\underbrace{\bar{g}^T\left(\bar{x}_0\right)\cdot d\bar{x}}_{scalar}\underset{\substack{if\ c\in\mathbb{R}\\ \Rightarrow c=c^T}}{=}\left(d\bar{x}\right)^T\cdot\bar{g}\left(\bar{x}_0\right)}$$

כעת, כיוון שהגדרנו $\varphi(\alpha) \triangleq f\left(\overline{x}_0 + \alpha\overline{r}\right)$, אזי φ היא פונקציה של משתנה יחיד וע"י החלפת משתנים,

כאשר נחליף $\overline{x} \triangleq \overline{x}_0 + \alpha\overline{r}$ ולכן נקבל $d\overline{x} = d\alpha\overline{r}$ (כי $\overline{x}_0$ קבוע), כלומר:

$$d\varphi(\alpha) = g^T\left(\overline{x}_0\right)\cdot \underbrace{d\overline{x}}_{d\alpha\overline{r}} = g^T\left(\overline{x}_0\right)\cdot d\alpha\overline{r} = d\alpha\cdot g^T\left(\overline{x}_0\right)\cdot\overline{r}$$

$$\Rightarrow \frac{d\varphi}{d\alpha} = g^T\left(\overline{x}_0\right)\cdot\overline{r}$$

ולכן לפי הגדרת הנגזרת לפי הדיפרנציאל נקבל:

$$\boxed{f'_{\overline{r}}\left(\overline{x}_0\right) \triangleq \varphi'(\alpha)\Big|_{\alpha=0} = \frac{d\varphi(\alpha)}{d\alpha}\Bigg|_{\alpha=0} = g^T\left(\overline{x}_0\right)\cdot\overline{r}}$$

(למעשה ניזכר כי ראינו תוצאה זו גם בחדו"א 2, כאשר אמרנו כי הנגזרת הכיוונית היא ההיטל של הגרדיאנט בכיוון שאנחנו מחפשים, בחדו"א 2 ביצענו גם נרמול לאורך הגרדיאנט ופה לא)

<u>מטריצת ההסיאן (Hessian)</u>:

תזכורת: פונקציה של משתנה יחיד

הדיפרנציאל של פונקציה בעלת משתנה יחיד מקיים:

$$df = f'(x)dx$$

והנגזרת של דיפרנציאל של פונקציה של משתנה יחיד היא:

$$df' = f''(x)dx$$

כעת ננסה למצוא את הנגזרת של הדיפרנציאל של פונקציה עם מספר משתנים. ראינו כבר כי מתקיים עבר פונקציה עם מספר משתנים:

$$df\left(\overline{x}\right) = \overline{g}^T\left(\overline{x}\right)\cdot d\overline{x}$$

כעת נמצא את ההקבלה למשוואה של הנגזרת השניה של פונקציה עם משתנה יחיד שכאמור היא $df' = f''(x)dx$. אנחנו למעשה מחפשים את הדיפרנציאל של הגרדיאנט (כפי שעברנו מהביטוי df לביטוי df'), אך <u>הגרדיאנט הוא וקטור</u> ולכן גם הדיפרנציאל שלו הוא וקטור, ולכן אנחנו מצפים שבמקום וקטור שמוכפל ב- $d\overline{x}$ אנחנו צריכים מטריצה (משיקולי מימדים), וזהו בדיוק ההסיאן:

$$\boxed{d\overline{g}\left(\overline{x}\right) = H\left(\overline{x}\right)\cdot d\overline{x}}$$

ונהוג לסמן: $H\left(\overline{x}\right) \triangleq \nabla^2 f\left(\overline{x}\right) \equiv \nabla^2_{\overline{x}} f\left(\overline{x}\right)$. כאשר $\nabla^2 = \left(\frac{\partial^2}{\partial x_1^2}, \quad \cdots \quad, \frac{\partial^2}{\partial x_n^2}\right)$.

<u>הוכחה</u>:

לפי ההגדרה הפורמלית של מטריצת ההסיאן שלומדים בחדו"א 2, מטריצת ההסיאן היא:

$$H\left(\bar{x}\right)=\begin{pmatrix}\frac{\partial^2 f}{\partial x_1\partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1\partial x_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 f}{\partial x_n\partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n\partial x_n}\end{pmatrix}$$

ונבחין במקרה בו הפונקציה $f\left(x\right)$ דיפרנציאבלית, אז לפי משפט שוורץ אפשר להחליף את סדר הגזירה ולכן <u>מטריצת ההסיאן היא סימטרית</u>.

כמו כן, ניתן לקבל מתוך ההגדרה שאנחנו פיתחנו כי מתקיים:

$$H\left(\bar{x}\right)=\begin{pmatrix}\frac{\partial g_1\left(\bar{x}\right)}{\partial x_1} & \cdots & \frac{\partial g_1\left(\bar{x}\right)}{\partial x_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial g_n\left(\bar{x}\right)}{\partial x_1} & \cdots & \frac{\partial g_n\left(\bar{x}\right)}{\partial x_n}\end{pmatrix}=\begin{pmatrix}\underbrace{\nabla}_{\substack{Row\\Vector}}\cdot\overbrace{g_1^T\left(\bar{x}\right)}^{\substack{Row\\Vector}}\\ \vdots\\ \nabla\cdot g_n^T\left(\bar{x}\right)\end{pmatrix}$$

(כאשר $\nabla=\left(\frac{\partial}{\partial x_1},\quad\cdots\quad,\frac{\partial}{\partial x_n}\right)$ ולכן $\underset{\substack{Row\\Vector}}{\nabla}\cdot\overbrace{g_1^T\left(\bar{x}\right)}^{\substack{Row\\Vector}}=Row\ Vector$ (והמכפלה היא איבר-איבר)).

והביטוי האחרון למטריצת ההסיאן נכון כי ראינו שמתקיים: $g\left(\bar{x}\right)=\left(g_1\left(\bar{x}\right)\quad\cdots\quad g_n\left(\bar{x}\right)\right)^T=\left(\frac{\partial f}{\partial x_1}\quad\cdots\quad\frac{\partial f}{\partial x_n}\right)^T$,

כלומר $g\left(\bar{x}\right)$ הוא <u>פונקציה וקטורית</u>, וכל רכיב של הווקטור הוא למעשה פונקציה סקלרית של מספר משתנים (כמספר הרכיבים של ווקטור $\bar{x}$). בנוסף, ואפשר לראות בקלות כי מתקיים:

$$\frac{\partial^2 f\left(\bar{x}\right)}{\partial x_i\partial x_j}=\frac{\partial}{\partial x_i}\left(\frac{\partial f\left(\bar{x}\right)}{\partial x_j}\right)=\frac{\partial}{\partial x_i}\left(g_j\left(\bar{x}\right)\right)=\frac{\partial g_j\left(\bar{x}\right)}{\partial x_i}$$

ולכן:

$$H\left(\bar{x}\right)=\begin{pmatrix}\frac{\partial^2 f\left(\bar{x}\right)}{\partial x_1\partial x_1} & \cdots & \frac{\partial^2 f\left(\bar{x}\right)}{\partial x_1\partial x_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 f\left(\bar{x}\right)}{\partial x_n\partial x_1} & \cdots & \frac{\partial^2 f\left(\bar{x}\right)}{\partial x_n\partial x_n}\end{pmatrix}=\begin{pmatrix}\frac{\partial g_1\left(\bar{x}\right)}{\partial x_1} & \cdots & \frac{\partial g_1\left(\bar{x}\right)}{\partial x_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial g_n\left(\bar{x}\right)}{\partial x_1} & \cdots & \frac{\partial g_n\left(\bar{x}\right)}{\partial x_n}\end{pmatrix}=\begin{pmatrix}\underbrace{\nabla}_{\substack{Row\\Vector}}\cdot\overbrace{g_1^T\left(\bar{x}\right)}^{\substack{Row\\Vector}}\\ \vdots\\ \nabla\cdot g_n^T\left(\bar{x}\right)\end{pmatrix}$$

כמו כן, נפשט את הביטוי $H\left(\bar{x}\right)\cdot d\bar{x}$ ונקבל:

$$H(\bar{x})\cdot d\bar{x} = H(\bar{x})\begin{pmatrix} dx_1 \\ \vdots \\ dx_n \end{pmatrix} = \begin{pmatrix} \overbrace{\nabla g_1^T(\bar{x})}^{\text{Row Vector}}\cdot \overset{\text{Column Vector}}{d\bar{x}} \\ \vdots \\ \nabla g_n^T(\bar{x})\cdot d\bar{x} \end{pmatrix} = \begin{pmatrix} \overbrace{dg_1(\bar{x})}^{\text{Scalar}} \\ \vdots \\ dg_n(\bar{x}) \end{pmatrix} = d\bar{g}(\bar{x})$$

מש"ל.

<u>הנגזרת של הנגזרת הכיוונית בכיוון כלשהו:</u> (Second Directional Derivative)

ראינו כי אם נגדיר $\varphi(\alpha) = f(\bar{x}_0 + \alpha\cdot\bar{r})$ אז הנגזרת הכיוונית של f בנקודה $\bar{x}_0$ בכיוון $\bar{r}$ מסומנת להיות $f_{\bar{r}}(\bar{x}_0)$ והיא שווה לביטוי:

$$f_{\bar{r}}(\bar{x}_0) = \varphi'(\alpha)\big|_{\alpha=0} = g^T(\bar{x}_0)\cdot\bar{r}$$

(כאשר $g^T(\bar{x}_0)$ זה הגרדיאנט (שהוא וקטור) של הפונקציה $f(\bar{x})$ בנקודה $\bar{x}_0$.

הנגזרת הכיוונית השניה מוגדרת באופן דומה ונקבל:

$$\varphi''(\alpha)\big|_{\alpha=0} = f''_{\bar{r}\bar{r}}(\bar{x}_0) = \left(\frac{\partial^2}{\partial\alpha^2} f(\bar{x}_0 + \alpha\bar{r})\right)\Bigg|_{\alpha=0} = \left(\frac{\partial}{\partial\alpha} f'_r\left(\underbrace{\bar{x}_0 + \alpha\bar{r}}_{\triangleq\bar{x}}\right)\right)\Bigg|_{\alpha=0}$$

אנחנו יודעים כי מתקיים:

$$f_{\bar{r}}(\bar{x}_0) = g^T(\bar{x}_0)\cdot\bar{r} = \bar{r}^T\cdot g(\bar{x}_0)$$

ולכן הדיפרנציאל של הנגזרת הכיוונית הוא (נזכור כי למעשה $g(\bar{x}_0) \equiv \bar{g}(\bar{x}_0)$ הוא גרדיאנט ולכן הוא וקטור!) :

$$df'_{\bar{r}}(\bar{x}_0) = \bar{r}^T\cdot d\bar{g}(\bar{x}_0) = \bar{r}^T\cdot H(\bar{x}_0)\cdot d\bar{x} = \bar{r}^T\cdot H(\bar{x}_0)\cdot d\alpha\bar{r} = \bar{r}^T\cdot H(\bar{x}_0)\cdot\bar{r}\cdot d\alpha$$

ולכן לפי הגדרת הנגזרת באמצעות הדיפרנציאל נקבל כי מתקיים:

$$\boxed{f''_{\bar{r}\bar{r}}(\bar{x}_0) = \frac{df'_{\bar{r}}(\bar{x}_0)}{d\alpha} = \bar{r}^T H(\bar{x}_0)\bar{r}}$$

וזו תבנית ריבועית שכבר ראינו (בהקשרים של ערכים עצמיים ווקטורים עצמיים)

<u>דיפרנציאל של אופרטור לינארי:</u>

נניח כי $\bar{y} = A\bar{x}$, כאשר A זו מטריצה. הדיפרנציאל של אופרטור לינארי הוא:

$$d\bar{y} \triangleq A(\bar{x} + d\bar{x}) - A\bar{x} = Ad\bar{x}$$

<u>גרדיאנט של פונקציה לינארית:</u>

נניח כי $\bar{b}$ הוא וקטור עמודה. אם נניח כי הפונקציה הסקלרית היא $f(\bar{x}) = \bar{b}^T\cdot\bar{x}$ אז הדיפרנציאל שלה הוא:

$$\forall d\bar{x}:\ \ df\left(\bar{x}\right)=\bar{b}^T d\bar{x}$$

ואם נשווה את הביטוי האחרון להגדרה שראינו $df\left(\bar{x}\right)=\bar{g}^T\left(\bar{x}\right)\cdot d\bar{x}$ נקבל:

$$\bar{g}\left(\bar{x}\right)=\nabla f\left(\bar{x}\right)=\bar{b}$$

<u>דוגמה</u>: (מציאת גרדיאנט של פונקציה בעלת תבנית ריבועית)

נניח כי נתונה לנו הפונקציה הסקלרית (התוצאה שלה היא סקלר) של מספר משתנים: $f\left(\bar{x}\right)=\bar{x}^T A\bar{x}$, כאשר $\bar{x}$ הוא ווקטור עמודה מימד n והמטריצה A היא ריבועית מימד $n\times n$. אפשר לרשום בצורה מפורשת את הפונקציה בצורה הבאה:

$$f\left(\bar{x}\right)=\sum_{i,j}x_i a_{i,j}x_j$$

אנחנו רוצים למצוא את הגרדיאנט של הפונקציה הנ"ל בנקודה כלשהי. ניתן למצוא את הגרדיאנט לפי ההגדרה, לגזור את הפונקציה כל פעם לפי אחד המשתנים ובסופו של דבר לקבל את הגרדיאנט אך אנחנו נמצא אותו באמצעות הגדרת הדיפרנציאל שלמדנו. לפי שיטות של מציאת דיפרנציאל, אנחנו צריכים לסכום מספר איברים כמספר ההופעות של $\bar{x}$ בפונקציה $f\left(\bar{x}\right)$ ובכל פעם להתייחס למופע אחד של $\bar{x}$ בתור משתנה, ואל השאר בתור קבועים (דומה לנגזרת של מכפלה). הדיפרנציאל של פונקציה $f\left(\bar{x}\right)$ הוא:

$$df\left(\bar{x}\right)=\left(d\bar{x}^T\right)\cdot\underset{\text{Constant}}{A\bar{x}}+\underbrace{\underset{\text{Constant}}{\bar{x}^T A}\cdot\left(d\bar{x}\right)}_{Scalar}$$

נבחין כי הביטוי $\bar{x}^T A\cdot\left(d\bar{x}\right)$ הוא סקלר ולכן ניתן לעשות עליו את פעולת ה-Transpose מבלי שהוא ישנה משהו (כיוון שמתקיים $\forall c\in\mathbb{R}: c^T=c$) ונקבל:

$$df\left(\bar{x}\right)=\left(d\bar{x}^T\right)\cdot A\bar{x}+\underbrace{\bar{x}^T A\cdot\left(d\bar{x}\right)}_{Scalar}=\left(d\bar{x}^T\right)\cdot A\bar{x}+\left(d\bar{x}^T\right)A^T\cdot\bar{x}=d\bar{x}^T\left(A+A^T\right)\bar{x}$$

כעת אפשר לזהות מתוך הביטוי האחרון את הגרדיאנט, שכן ראינו לפי ההגדרה כי הגרדיאנט הוא פונקציה שתלויה ב- $\bar{x}$ ומוכפלת בדיפרנציאל כפי שראינו בהגדרה: $df\left(\bar{x}\right)=\bar{g}^T\left(\bar{x}\right)\cdot d\bar{x}=d\bar{x}^T\cdot\bar{g}\left(\bar{x}\right)$ ולכן קיבלנו כי מתקיים עבור הפונקציה $f\left(\bar{x}\right)=\bar{x}^T A\bar{x}$ כי הגרדיאנט שלה הוא:

$$\bar{g}\left(\bar{x}\right)=\left(A+A^T\right)\bar{x}$$

וכן במקרה בו A היא סימטרית, אז מתקיים $A=A^T$ ונקבל:

$$\bar{g}\left(\bar{x}\right)=2A\bar{x}$$

(נבחין כי אם הפונקציה $f\left(\bar{x}\right)=\bar{x}^T A\bar{x}$ הייתה של משתנה יחיד אז היה מתקיים $f\left(x\right)=Ax^2$ וברור כי $f'\left(x\right)=2Ax$)

כעת נמצא את מטריצת ההסיאן של הפונקציה $f\left(\bar{x}\right)=\bar{x}^T A\bar{x}$. לפי מה שלמדנו מתקיים כי $d\bar{g}\left(\bar{x}\right)=H\left(\bar{x}\right)\cdot d\bar{x}$ ולכן אנחנו רוצים למצוא את הדיפרנציאל של $\bar{g}\left(\bar{x}\right)=\left(A+A^T\right)\bar{x}$:

$$d\bar{g}(\bar{x}) = \left(A + A^T\right) d\bar{x}$$

ולכן אנחנו מקבלים (ע"י השוואה לביטוי $d\bar{g}(\bar{x}) = H(\bar{x}) \cdot d\bar{x}$):

$$H(\bar{x}) = A + A^T$$

וכן במקרה בו A היא סימטרית, אז מתקיים $A = A^T$ ונקבל:

$$H(\bar{x}) = 2A$$

מדוגמה זו אנחנו למדים כי השיטה של מציאת הגרדיאנט ומטריצת ההסיאן של פונקציות מרובות משתנים באמצעות הגדרת הגרדיאנט והדיפרנציאל כפי שראינו אותה קודם נוחה ופשוטה יחסית. בעתיד נוכל להשתמש בשיטה זו גם עבור פונקציות מסובכות הרבה יותר, למשל פונקציות המתארות רשתות עצביות (Neural Networks).

<u>פיתוח טור טיילור של פונקציה מרובת משתנים והקשר לגרדיאנט ולמטריצת ההסיאן:</u>

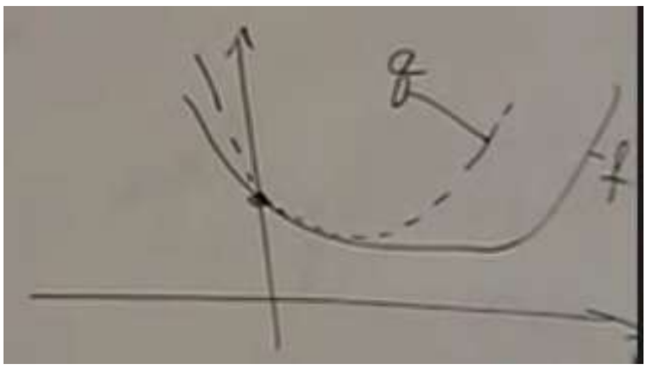
האיברים הראשונים (עד סדר שני) של פיתוח טיילור של פונקציה של מספר משתנים ניתן לרישום בצורה הבאה:

$$f(\bar{x}_0 + \bar{r}) = \underbrace{f(\bar{x}_0) + \bar{g}^T(\bar{x}_0) \cdot \bar{r} + \frac{1}{2}\bar{r}^T H(\bar{x}_0)\bar{r}}_{q(\bar{r})} + \ldots$$

נסמן את האיברים הראשונים הללו להיות:

$$q(\bar{r}) = f(\bar{x}_0) + \bar{g}^T(\bar{x}_0) \cdot \bar{r} + \frac{1}{2}\bar{r}^T H(\bar{x}_0)\bar{r}$$

נתבונן על הפונקציה $f(\bar{x}_0 + \bar{r})$ (שהיא פונקציה של $\bar{r}$, כי $\bar{x}_0$ הוא קבוע), אזי זה אומר כי ציר x של הגרף הבא הוא $\bar{r}$:



נבחין כי עבור $\bar{r} = 0$ נקבל בדיוק את הנקודה $f(\bar{x}_0)$. כמו כן, נבחין כי $q(\bar{r})$ הוא פיתוח טיילור עד סדר שני, ולכן הוא מנסה לקרב את פונקציה $f(\bar{x})$ בסביבת הנקודה $\bar{x}_0$ ולפונקציה $q(\bar{r})$ יש את אותה הנגזרת הראשונה וגם השניה בנקודה $\bar{x}_0$ בדיוק כמו לפונקציה $f(\bar{x})$, אך כיוון שאנחנו עוסקים בפונקציות של מספר משתנים הנגזרת הראשונה היא הגרדיאנט(!), נוכל לסכם זאת כך:

$$\bar{r}=\bar{0} \quad \Rightarrow \qquad \begin{aligned} & f\left(\bar{x}_0\right)=q\left(\bar{r}=\bar{0}\right) \\ & \nabla f\left(\bar{x}_0\right)=\bar{g}\left(\bar{x}_0\right)=\nabla q\left(\bar{r}\right) \\ & \nabla^2 f\left(\bar{x}_0\right)=\nabla^2 q\left(\bar{r}\right) \end{aligned}$$

ניתן בקלות להוכיח את שלושת המשוואות לעיל ע"י הצבה ישירה של $\bar{r}=\bar{0}$ בהגדרה של $q\left(\bar{r}\right)$ ולגזור פעם ופעמיים עבור שתי המשוואות האחרונות ולהיווכח שזה נכון, נבצע זאת:

המשוואה הראשונה היא הצבה ישירה.

המשוואה השניה (נזכור כי מטריצת ההסיאן היא סימטרית, כיוון שאנחנו מניחים שהפונקציה $f\left(\bar{x}\right)$ היא דיפרנציאבילית ולכן מתקיים משפט שוורץ שמאפשר החלפת סדר גזירה):

$$\nabla q\left(\bar{r}\right)\Big|_{\bar{r}=0} \underset{f\left(\bar{x}_0\right)\text{ is a constant}}{=} \left(\bar{g}\left(\bar{x}_0\right)+\frac{1}{2}\cdot 2H\left(\bar{x}_0\right)\bar{r}\right)\Bigg|_{\bar{r}=0} = \bar{g}\left(\bar{x}_0\right)=\nabla f\left(\bar{x}_0\right)$$

(הערה: נבחין כי מתקיים: $\frac{\partial}{\partial \bar{r}}\bar{g}^T\left(\bar{x}_0\right)\cdot\bar{r}=\bar{g}\left(\bar{x}_0\right)$ ובאופן כללי: $\frac{\partial}{\partial \bar{x}}\bar{a}^T\cdot\bar{x}=\bar{a}$)

המשוואה השלישית:

$$\nabla^2 q\left(\bar{r}\right)\Big|_{\bar{r}=0} \underset{\bar{g}^T(x)\text{ is a constant}}{=} \frac{1}{2}\cdot 2H\left(\bar{x}_0\right)=H\left(\bar{x}_0\right)$$

והוכחנו את הנדרש.

בקורס זה לא נשתמש בסדרים גבוהים יותר של טור טיילור אך בפיתוח עד סדר שני נשתמש רבות באלגוריתמים לאופטימיזציה.

<u>פונקציות של מטריצות:</u>

נגדיר פונקציה סקלרית $f: n\times n \rightarrow \mathbb{R}$, כלומר אם X היא מטריצה מסדר $n\times n$ אז $f\left(X\right)$ הוא סקלר.

באופן דומה לאיך שהגדרנו את הגרדיאנט של פונקציה של מספר משתנים (ווקטור של משתנים), בתור ווקטור של הנגזרות הראשונות של הפונקציה ביחס לכל אחד מהמשתנים, ניתן להגדיר את הגרדיאנט של פונקציה של מטריצה באמצעות מטריצת הנגזרת הראשונה של הפונקציה ביחס לכל אחד מהמשתנים באופן הבא:

$$Grad\ \ f=G\left(X\right)=\begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{n1}} & \cdots & \frac{\partial f}{\partial x_{nn}} \end{pmatrix}$$

<u>מכפלה פנימית/סקלרית של מטריצות:</u>

עד היום הכרנו את ההגדרה של מכפלה סקלרית בין ווקטורים. מכפלה סקלרית/פנימית בין שתי מטריצות מוגדרת באופן הבא:

$$\left\langle A,B\right\rangle=\sum_{i,j}a_{ij}b_{ij}=Trace\left(A^T B\right)$$

<u>הדיפרנציאל של פונקציה סקלרית של מטריצות:</u>

נניח כי $f : n\times n \to \mathbb{R}$ אז מתקיים כי דיפרנציאל של מטריצה f הוא:

$$df = \sum_{i,j} \frac{\partial f}{\partial x_{ij}} \cdot dx_{ij} = Trace\left(G^T\left(X\right) \cdot dX\right)$$

כאשר נגדיר:

$$dX = \begin{pmatrix} dx_{11} & \cdots & dx_{1n} \\ \vdots & \ddots & \vdots \\ dx_{n1} & \cdots & dx_{nn} \end{pmatrix}$$

אך נבחין כי לפי ההגדרה של מכפלה פנימית בין מטריצות נקבל כי מתקיים:

$$df = \sum_{i,j} \frac{\partial f}{\partial x_{ij}} \cdot dx_{ij} = Trace\left(G^T\left(X\right) \cdot dX\right) = \left\langle G\left(X\right), dX \right\rangle$$

אך זה לא דבר חדש לנו, שכן ראינו כי בדיוק אותו הדבר מתקיים עבור דיפרנציאל של פונקציה של מספר משתנים.

<u>דוגמה:</u> (גרדיאנט של רשת עצבים)

נזכיר כי רשת עצבים (בעלת שכבה אחת) פשוטה ניתנת לתיאור ע"י הפונקציה הסקלרית (מרובת המשתנים) הבאה:

$$f\left(\bar{x}, W, \bar{b}, \bar{v}\right) = v^T \varphi\left(\underbrace{W^T \bar{x} + \bar{b}}_{\triangleq \bar{u}}\right)$$

כאשר $\bar{\varphi}$ היא פונקציה ווקטורית, $\bar{\varphi} : \mathbb{R}^n \to \mathbb{R}^n$, כלומר:

$$\bar{\varphi}\left(\bar{u}\right) = \begin{pmatrix} \varphi\left(u_1\right) \\ \vdots \\ \varphi\left(u_n\right) \end{pmatrix} \qquad \Rightarrow \qquad d\bar{\varphi} = \begin{pmatrix} \varphi'\left(u_1\right) du_1 \\ \vdots \\ \varphi'\left(u_n\right) du_n \end{pmatrix} = \begin{pmatrix} \varphi'\left(u_1\right) & & 0 \\ & \ddots & \\ 0 & & \varphi'\left(u_n\right) \end{pmatrix} \begin{pmatrix} du_1 \\ \vdots \\ du_n \end{pmatrix}$$

האופטימיזציה שאנחנו מחפשים היא למעשה לחפש את המקדמים של מטריצה W , וכן את המקדמים של הווקטורים $\bar{b}, \bar{v}$. בכדי למצוא את המקדמים, אנחנו צריכים את הנגזרות ביחס לכל אחד מהמשתנים $W, \bar{b}, \bar{v}$ אך בדוגמה זו נמצא רק את הגרדיאנט של f ביחס למטריצה W.

בשביל למצוא את הגרדיאנט של f ביחס למטריצה W נרצה למצוא את הדיפרנציאל של פונקציה f ביחס לדיפרנציאל (שינוי) באחד המשתנים, ובמקרה שלנו, אנחנו מחפשים שינוי במטריצה W (כאשר כל שאר המשתנים הם קבועים) ולכן נסמן את הדיפרנציאל של f ביחס למטריצה W להיות $d_W f$.

תחילה נסמן:

$$\vartheta' \triangleq \begin{pmatrix} \varphi'(u_1) & & 0 \\ & \ddots & \\ 0 & & \varphi'(u_n) \end{pmatrix}$$

ולכן, לפי הסימון נקבל:

$$d\bar{\varphi} = \vartheta' \cdot d\bar{u}$$

כעת אנחנו נמצא את הדיפרנציאל של f ביחס למטריצה W. אנחנו רואים לפי הסימון של $\bar{u}$, שהוא פונקציה של W והדיפרנציאל של $\bar{u}$ ביחס למטריצה W הוא:

$$d_W \bar{u} = \left(dW^T\right)\bar{x}$$

ולכן:

$$d_W f = \bar{v}^T d\bar{\varphi} = \bar{v}^T \vartheta' d\bar{u} = \bar{v}^T \vartheta' \left(dW^T\right)\bar{x}$$

אנחנו יודעים כי לפי הגדרת הגרדיאנט, אנחנו נרצה כעת כי הביטוי dW^T יהיה מוכפל בביטוי אחר משמאל (שיהווה את הגרדיאנט של פונקציה f), אך כעת אנחנו רואים כי הוא במרכז ולכן נרצה לבודד אותו. נבחין כי הביטוי $\bar{v}^T \vartheta' \left(dW^T\right)\bar{x}$ הוא סקלר וכן אנחנו יודעים כי לכל סקלר מתקיים $c = Trace(c)$ ולכן נקבל:

$$d_W f = \bar{v}^T \vartheta' \left(dW^T\right)\bar{x} = Trace\left(\bar{v}^T \vartheta' \left(dW^T\right)\bar{x}\right)$$

אך תחת $Trace$ יש לנו את תכונת הקומוטטיביות הציקלית (כלומר $Trace(A \cdot B) = Trace(B \cdot A)$ (גם עבור מטריצות וגם עבור ווקטורים וגם עבור סקלרים)) ולכן נוכל לרשום:

$$d_W f = \bar{v}^T \vartheta' \left(dW^T\right)\bar{x} = Trace\left(\underset{A}{\bar{v}^T}\ \underbrace{\vartheta' \left(dW^T\right)\bar{x}}_{B}\right) = Trace\left(\underbrace{\vartheta' \left(dW^T\right)}_{C}\ \underset{D}{\bar{x} \cdot \bar{v}^T}\right) = Trace\left(\bar{x} \cdot \bar{v}^T \vartheta' \left(dW^T\right)\right)$$

נשתמש בתכונת הקומוטטיביות פעם נוספת ונקבל:

$$d_W f = \bar{v}^T \vartheta' \left(dW^T\right)\bar{x} = Trace\left(\underbrace{\bar{x} \cdot \bar{v}^T \vartheta'}_{A}\underbrace{\left(dW^T\right)}_{B}\right) = Trace\left(\underbrace{\left(dW^T\right)}_{C}\underbrace{\bar{x} \cdot \bar{v}^T \vartheta'}_{D}\right) = \left\langle dW, \bar{x} \cdot \bar{v}^T \vartheta' \right\rangle$$

כמו כן, ראינו כי מתקיים:

$$df = \sum_{i,j} \frac{\partial f}{\partial x_{ij}} \cdot dx_{ij} = Trace\left(G^T(X) \cdot dX\right) = \left\langle G(X), dX \right\rangle \qquad \Rightarrow \qquad df = \left\langle G(X), dX \right\rangle$$

ולכן נוכל להשוות בין הביטויים האחרונים ונקבל:

$$G_W(X) = \bar{x} \cdot \bar{v}^T \cdot \vartheta'$$

כמו כן ניתן באופן דומה לחשב את הגרדיאנטים ביחס למשתנים האחרים $\bar{v},\bar{b}$, כלומר לחשב את הביטויים $\nabla_{\bar{v}} f, \nabla_{\bar{b}} f$.

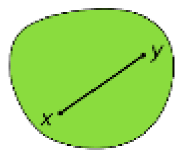
הרצאה Convex Sets and Functions – 04-05, קבוצות קמורות ופונקציות קמורות

<u>הגדרה</u>: (קבוצה קמורה)

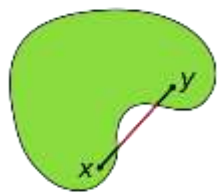
קבוצה אשר הקו הישר בין כל שתי נקודות בקבוצה נמצא כולו בקבוצה. ובאופן פורמלי, קבוצה C תקרא קבוצה קמורה, אם לכל שתי נקודות $x, y \in C$ מתקיים:

$$\left(\alpha x+(1-\alpha) y\right) \in C \quad , \qquad \forall \alpha \in [0,1]$$

<u>דוגמה לקבוצה קמורה</u>:



<u>דוגמה לקבוצה לא קמורה</u>:



<u>הגדרה</u>: (פונקציה קמורה)

פונקציה $f(x)$ (של משתנה יחיד) תקרא פונקציה קמורה, אם לכל שתי נקודות x_1, x_2 , הקו הלינארי שמחבר בין $f(x_1), f(x_2)$ יהיה מעל (גדול או שווה) הפונקציה $f(x)$ בתחום $[x_1, x_2]$ ובצורה פורמלית:

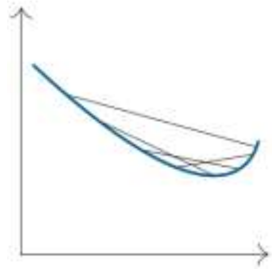
הפונקציה $f(x)$ היא פונקציה קמורה היא לכל x_1, x_2 מתקיים:

$$f\left(\alpha x_1+(1-\alpha) x_2\right) \leq \alpha f(x_1)+(1-\alpha) f(x_2) \ , \qquad \forall \alpha \in [0,1]$$

(הערה: פונקציה תקרא <u>קמורה ממש</u>, תדרוש אי שיוויון חזק בהגדרה, במקום אי-שיוויון חלש)

(הערה: פונקציה קמורה יכולה להיות גם פונקציה של מספר משתנים, וההגדרה המילולית משתנה להיות במקום קו ישר, אלא הישר שעובר על המרחק בין ערכי הפונקציה וכן צריך לשפר את ההגדרה הפורמלית)

<u>דוגמה לפונקציה קמורה</u>:



<u>הגדרה</u>: (פונקציה קעורה)

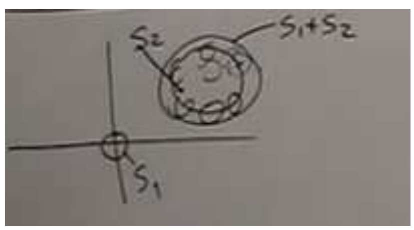
פונקציה $f(x)$ (של משתנה יחיד) תקרא פונקציה קעורה אם $(-f(x))$ היא פונקציה קמורה.

(הערה: פונקציה תקרא <u>קעורה ממש</u>, אם $(-f(x))$ היא קמורה ממש)

<u>תכונות של קבוצות קמורות</u>: (להוכיח כשיעורי בית את התכונות)

נניח כי הקבוצות C_i הן קבוצות קמורות. מתקיימות התכונות הבאות:

1) הקבוצה $\bigcap_i C_i$ היא קבוצה קמורה.
2) הקבוצה $C_i + C_j = \left\{x_p + x_k \mid x_p \in C_1, x_k \in C_2\right\}$ היא קבוצה קמורה.



(אפשר להבין זאת אם נחשוב על נקודות כעל וקטורים, ולכן הקבוצה $C_i + C_j$ היא הקבוצה של כל הנקודות מתוך C_i כאשר אנחנו מוסיפים לכל נקודה כזו את כל הנקודות מתוך C_j, ואפשר להתייחס לכל הנקודות בתור וקטורים, ולכן למעשה קיבלנו כי סביב כל נקודה בקבוצה C_i קיבלנו את הקבוצה C_j).

3) הקבוצה $\left\{A\overline{x} \mid \overline{x} \in C_i, A: n \times n\right\}$ הקבוצה שנוצרת ע"י העתקה (מטריצה מייצגת) היא קבוצה קמורה.
4) קבוצות שמייצגות את כל התחום הפנימי של ערך זהה של פונקציה קמורה (פונקצית של מספר משתנים), ובצורה פורמלית: $\left\{\overline{x} \in C_i \mid f(\overline{x}) : C \to \mathbb{R}, f(\overline{x}) \le \alpha\right\}$.

<u>הוכחה לתכונה (1)</u>:

נתבונן בשתי נקודות כלשהן $x, y \in \bigcap_i C_i$. אם $x, y \in \bigcap_i C_i$ אז לפי ההגדרה של חיתוך, שתי הנקודות שייכות לכל קבוצה קמורה C_i ולכן הקו הישר העובר בין שתי הנקודות שייך לכל קבוצה C_i, אך זה אומר, לפי ההגדרה, שכל קו כזה שייך גם

לחיתוך $\bigcap_i C_i$ ולכן הוכחנו כי לכל שתי נקודות $x,y\in\bigcap_i C_i$ הקו הישר המחבר בין הנקודות נמצא בקבוצה $\bigcap_i C_i$ ולכן $\bigcap_i C_i$ הוא קבוצה קמורה לפי ההגדרה. מש"ל.

<u>תכונות של פונקציות קמורות</u>: (להוכיח כשיעורי בית)

נניח כי $f(\overline{x}),g(\overline{x})$ פונקציות קמורות.

1) נגדיר את הפונקציה $\varphi:\mathbb{R}\to\mathbb{R}$ בצורה הבאה: $\varphi(\alpha)=f(\overline{x}+\alpha\overline{r})$ לכל $\alpha\in\mathbb{R}$ ולכל $\overline{x},\overline{r}\in\mathbb{R}^n$, אזי הפונקציה $\varphi(\alpha)$ היא פונקציה קמורה.
2) (הרחבה לתכונה 1) אם לכל $\alpha\in\mathbb{R}$ ולכל $\overline{x},\overline{r}\in\mathbb{R}^n$ ולכל פונקציה $\varphi:\mathbb{R}\to\mathbb{R}$ שמוגדרת בצורה הבאה: $\varphi(\alpha)=f(\overline{x}+\alpha\overline{r})$, הפונקציה $\varphi(\alpha)$ היא פונקציה קמורה אז גם הפונקציה $f(\overline{x})$ היא פונקציה קמורה.
3) ("לינאריות") הפונקציה $v(\overline{x})=\alpha f(\overline{x})+\beta g(\overline{x})$ היא פונקציה קמורה לכל $\alpha,\beta>0$.
4) הפונרציה $m(\overline{x})=\max\{f(\overline{x}),g(\overline{x})\}$ היא פונקציה קמורה.
5) נניח כי $h:\mathbb{R}\to\mathbb{R}$ פונקציה קמורה ומונוטונית עולה אז הפונקציה $h(f(\overline{x}))$ היא פונקציה קמורה.

<u>פונקציה קמורה מוכללת (Extended value convex function)</u>:

נניח כי $C\in\mathbb{R}^n$ קבוצה קמורה וכן $f(\overline{x}):\mathbb{R}^n\to\mathbb{R}$ היא פונקציה קמורה. הפונקציה הקמורה המוכללת מוגדרת להיות פונקציה $\overline{f}(\overline{x}):\mathbb{R}^n\to\mathbb{R}$ ומוגדרת באופן הבא:

$$\overline{f}(\overline{x})=\begin{cases} f(\overline{x}) & , \quad x\in C \\ \infty & , \quad x\notin C\end{cases}$$

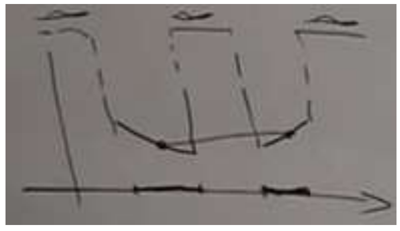
וההגדרה היא:

$$\overline{f}(\alpha\overline{x}_1+(1-\alpha)\overline{x}_2)\le\alpha f(\overline{x}_1)+(1-\alpha)f(\overline{x}_2)$$

או במובן שקול (ע"י אלגברה פשוטה):

$$\overline{f}(\overline{x}_2+\alpha(\overline{x}_1-\overline{x}_2))\le f(\overline{x}_2)+\alpha\left(f(\overline{x}_1)-f(\overline{x}_2)\right)$$

<u>הערה</u>: נבחין כי גם אם הפונקציה המקורית $f(\overline{x})$ היא לא פונקציה קמורה (כי למשל היא מורכבת רק משני תחומים שבהם היא קמורה), אז לפי ההגדרה לעיל, עדיין אנחנו נקבל כי הפונקציה המוכללת היא קמורה למרות שניתן לחשוב בטעות כי הקו הנראה בשרטוט (למטה) הוא הקו הלינארי, אך זה לא נכון כי הקו הלינארי הוא בעל ערך אינסופי לפי ההגדרה לעיל:



<u>הערה</u>: הגדרה זו נוחה כאשר אנחנו רוצים להתייחס לפונקציה שקמורה רק בתחום מסויים שלה (אך לא בכל בתחום) ולכן נוח להתייחס לפונקציה הקמורה המוכללת המוגדרת על ידה, כי התחום של הפונקציה המוכללת הקמורה שלה היא כל $\mathbb{R}^n$).

<u>הגדרה</u>: (אפיגרף (epigraph))

אפיגרף של פונקציה $f\left(\bar{x}\right)$, הוא קבוצת הנקודות שנמצאות על או מעל הגרף $f\left(\bar{x}\right)$. ובאופן פורמלי: נניח כי $C\in\mathbb{R}^n$ היא קבוצה קמורה, וכן $f\left(\bar{x}\right):C\to\mathbb{R}$ אז מתקיים:

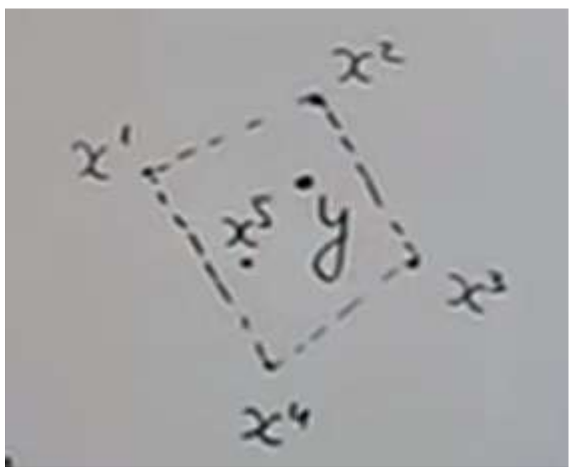
$$epi(f)=\left\{\left(\bar{x},y\right)\mid x\in C, y\in\mathbb{R}, y\geq f\left(\bar{x}\right)\right\}$$

<u>משפט:</u>

$f\left(\bar{x}\right)$ היא פונקציה קמורה אמ"מ $epi\left(f\right)$ היא קבוצה קמורה. (הוכחה: שיעורי בית)

<u>הגדרה</u>: (קומבינציה קמורה convex combination)

תהי קבוצה $C\in\mathbb{R}^n$ ויהיו $\bar{x}_i\in\mathbb{R}^n$ נקודות וכן $\alpha_i\in\mathbb{R}$ לכל $1\leq i\leq\infty$ כאשר $\forall i:\alpha_i\geq 0$ וכן $\sum_i\alpha_i=1$ אז אם הנקודה $y=\sum_{i=1}^m\alpha_i\bar{x}_i$ (כאשר $1\leq m\leq\infty$) נמצאת בתוך הקבוצה C, לכל הנקודות $\bar{x}_i\in\mathbb{R}^n$ ולכל $\alpha_i\in\mathbb{R}$ אז הקבוצה C היא קומבינציה קמורה.



(הנקודה $y=\sum_{i=1}^m\alpha_i\bar{x}_i$ נמצאת בתוך התחום שנוצר ע"י הנקודות $\bar{x}_i\in\mathbb{R}^n$).

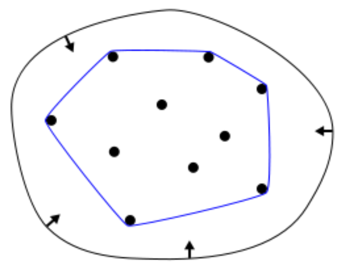
<u>הגדרה</u>: (קמור, convex hull)

הקמור של גוף או אוסף של גופים הוא הגוף הקמור המינימלי המכיל אותם. הקמור של נקודות במישור, הוא סוג של גומיה שנמתחת מעל כל הנקודות המישור ונמתחה כך שתכיל את כל הנקודות ולאחר מכן שוחררה. סימון:

$$conv\left\{\bar{x}_i\right\}_{i=1}^m$$

(הסבר לסימון: הקבוצה הקמורה המינימלית שמכילה את כל הנקודות $\left\{\bar{x}_i\right\}_{i=1}^m$)

<u>דוגמה לקמור</u>:



<u>הגדרה שקולה לקמור</u>:

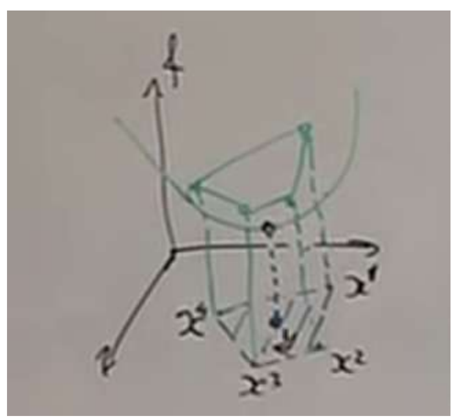
הקבוצה הקמורה המינימלית (בשטח/נפח וכו') שמכילה את כל הנקודות $\left\{\overline{x}_i\right\}_{i=1}^{m}$.

<u>אי-שיוויון ינסן</u>: (נלמד בקורס הסתברות ח')

נניח כי $f\left(\overline{x}\right)$ פונקציה קמורה (של n משתנים), וכן $\overline{x}_i \in \mathbb{R}^n$ נקודות בתחום, אז לכל $1 \le i \le n$, $\alpha_i \ge 0$ וכן $\sum_{i=1}^{n} \alpha_i = 1$ מתקיים:

$$f\left(\sum_{i=1}^{n} \alpha_i \overline{x}_i\right) \le \sum_{i=1}^{n} \alpha_i f\left(\overline{x}_i\right)$$

ולשם המחשה:



ערך הפונקציה בנקודה $\sum_{i=1}^{n} \alpha_i \overline{x}_i$ הוא בהכרח קטן או שווה לקומבינציה הלינארית של הערכים של הפונקציה בנקודות $\overline{x}_i$.

<u>דוגמה</u>: (דוגמה לשימוש באי-שיוויון ינסן)

נוכיח כי הממוצע החשבוני גדול או שווה לממוצע הגיאומטרי של קבוצת נקודות (חלק ממשפט אי-שיוויון הממוצעים), כלומר נוכיח כי:

$$\frac{1}{m}\sum_{i=1}^{m} x_i \geq \left(\prod_{i=1}^{m} x_i\right)^{1/m}$$

<u>הוכחה:</u>

נזכור כי הפונקציה $\log(\cdot)$ היא פונקציה קעורה, ולכן ע"י הפעלת פונקצית $(-\log(\cdot))$ (מינוס לוג שהיא פונקציה קמורה לפי ההגדרה) על שני האגפים ופישוט אי-השיוויון נוכל להשתמש באי-שיוויון ינסן ולקבל את מה שנדרשנו להוכיח.

<u>נגזרות מסדר ראשון ושני (גרדיאנטים) של פונקציות קמורות:</u>

<u>אי-שיוויון הגרדיאנט:</u>

הפונקציה $f(\bar{x}):\mathbb{R}^n \to \mathbb{R}$ היא פונקציה קמורה ו"חלקה" אמ"מ לכל $\bar{d} \in \mathbb{R}^n$ מתקיים

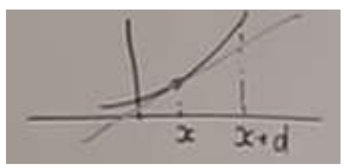
$$f(\bar{x}+\bar{d}) \geq f(\bar{x}) + \bar{d}^T \nabla f(\bar{x})$$

או באופן שקול:

$$f(\bar{x}+\bar{d}) - f(\bar{x}) \geq \bar{d}^T \nabla f(\bar{x})$$

(הערה: $f(\bar{x}) + \bar{d}^T \nabla f(\bar{x})$ הוא פיתוח טיילור מסדר ראשון של $f(\bar{x})$)

<u>המחשה של אי-שיוויון הגרדיאנט עבור פונקציה של משתנה יחיד:</u>



אנחנו רואים כי המשיק (קו משיק/משטח משיק) נמצא על או מתחת לגרף הפונקציה בנקודה אליה הוא משיק.

<u>נגזרת שניה של פונקציה קמורה (במקרה של פונקציה של משתנה יחיד):</u>

אם $f(x)$ היא פונקציה קמורה אז לכל $x_1 < x_2$ מתקיים $f'(x_1) \leq f'(x_2)$ וכן לכל x מתקיים $f''(x) > 0$.

<u>נגזרת שניה של פונקציה קמורה (במקרה של פונקציה של כמה משתנים):</u>

אם $f(\bar{x})$ היא פונקציה קמורה אז מטריצת ההסיאן של פונקציה $f(\bar{x})$ מקיימת $H(\bar{x}) \succeq 0$ (positive semi-definite).

<u>הוכחה:</u>

נתון כי $f(\bar{x})$ היא פונקציה קמורה, ולכן לכל פונקציה $\varphi:\mathbb{R}^n \to \mathbb{R}$ שנגדיר באופן הבא: $\varphi(\alpha) = f(\bar{x}+\alpha\bar{r})$ (לכל $\bar{x},\bar{r} \in \mathbb{R}^n$ וכן לכל $\alpha \in \mathbb{R}$) הפונקציה φ תהיה קמורה ולכן מתקיים לכל α מתקיים: $\varphi''(\alpha) \geq 0$. כמו כן, ראינו כי הנגזרת הכיוונית, מסדר שני, של הפונקציה $f(\bar{x})$, אשר מוגדרת להיות $f''_{\bar{r}\bar{r}}(\bar{x})$ (והיא גם שווה לפי ההגדרה ל- $\varphi''(\alpha)\big|_{\alpha=0}$) ולכן קיבלנו:

$$\forall \bar{r} \in \mathbb{R}^n: \quad \varphi''(\alpha)\big|_{\alpha=0} = f''_{\bar{r}\bar{r}}(\bar{x}) = \bar{r}^T \cdot \nabla^2 f(\bar{x}) \cdot \bar{r} = \bar{r}^T H(\bar{x}) \bar{r} \geq 0$$

ולכן, לפי ההגדרה של מטריצה אי-שלילית (ההגדרה אומרת כי המטריצה A תהיה אי-שלילית אם לכל ווקטור $\forall \overline{r} \in \mathbb{R}^n$ מתקיים $\overline{r}^T \cdot A \cdot \overline{r} \geq 0$) קיבלנו כי מתקיים: $H\left(\overline{x}\right) \succeq 0$. מש"ל.

<u>הערה</u>: אם $f\left(\overline{x}\right)$ פונקציה קמורה ממש אז מתקיים $H\left(\overline{x}\right) \succ 0$.

הרצאה Local and Global Minimum – 06 , מינימום מקומי וגלובאלי

<u>הגדרה:</u> (מינימום גלובלי)

נניח כי $C\in\mathbb{R}^n$ תחום (קבוצה) קמור וכן פונקציה $f(\bar{x}):C\to\mathbb{R}$. נאמר כי $\bar{x}\in C$ הוא מינימום גלובלי של הפונקציה f מעל הקבוצה C אם לכל $\forall\bar{y}\in C$ מתקיים:

$$f(\bar{x})\le f(\bar{y})$$

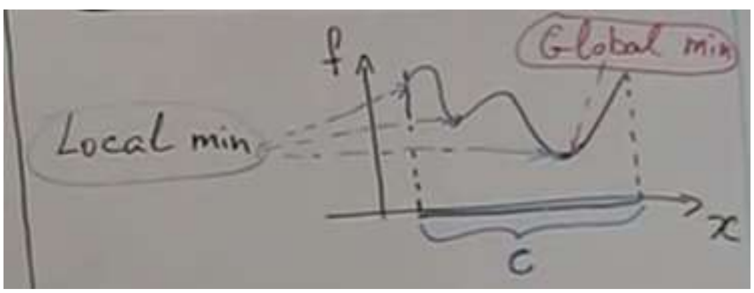
<u>הגדרה:</u> (מינימום מקומי, מינימום לוקאלי)

נניח כי $C\in\mathbb{R}^n$ תחום (קבוצה) קמור וכן פונקציה $f(\bar{x}):C\to\mathbb{R}$. נאמר כי $\bar{x}$ הוא מינימום מקומי (לוקאלי) אם <u>קיים</u> $\varepsilon>0$ כך שמתקיים לכל $\bar{y}\in B_\varepsilon(\bar{x})\bigcap C$:

$$f(\bar{x})\le f(\bar{y})$$

(כאשר $B_\varepsilon(\bar{x})$ הוא סביבה ברדיוס ε של הנקודה $\bar{x}$, כלומר $B_\varepsilon(\bar{x})=\{\bar{z}\mid\|\bar{z}-\bar{x}\|\le\varepsilon\}$)

<u>דוגמה:</u>



נבחין כי גם הנקודה השמאלית ביותר של תחום C היא מינימום מקומי.

<u>משפט:</u>

נניח $C\in\mathbb{R}^n$ תחום קמור. לכל פונקציה קמורה $f(\bar{x}):C\to\mathbb{R}$ מתקיים כי כל מינימום מקומי הוא מינימום גלובלי.

<u>הוכחה:</u> (על דרך השלילה)

נתון כי $C\in\mathbb{R}^n$ תחום קמור וכן נתונה פונקציה קמורה $f(\bar{x}):C\to\mathbb{R}$. נניח בשלילה כי $\bar{x}_0$ הוא מינימום מקומי אך לא מינימום גלובלי, אם כך, אז קיים מינימום גלובלי אחר, כלומר קיים $\bar{y}_0\in C$ כך שמתקיים: $f(\bar{y}_0)<f(\bar{x}_0)$. לפי ההגדרה של פונקציה קמורה, מתקיים (ע"י פיתוח פשוט של ההגדרה של פונקציה קמורה שראינו קודם – למעשה מחליפים תפקידים בין שתי הנקודות $\bar{x}_0$ ו- $\bar{y}_0$):

$$\forall\alpha\in[0,1]:\quad f\left(\bar{x}_0+\alpha(\bar{y}_0-\bar{x}_0)\right)\le f(\bar{x}_0)+\alpha\left(f(\bar{y}_0)-f(\bar{x}_0)\right)$$

כך מצאנו כי $f(\bar{y}_0)<f(\bar{x}_0)$ ולכן מתקיים $f(\bar{y}_0)-f(\bar{x}_0)<0$ ולכן:

$$f\left(\bar{x}_0+\alpha\left(\bar{y}_0-\bar{x}_0\right)\right)\le f\left(\bar{x}_0\right)+\alpha\left(f\left(\bar{y}_0\right)-f\left(\bar{x}_0\right)\right)<f\left(\bar{x}_0\right)\quad\Rightarrow\qquad f\left(\bar{x}_0+\alpha\left(\bar{y}_0-\bar{x}_0\right)\right)<f\left(\bar{x}_0\right)$$

כלומר מצאנו סביבה של נקודות מסביב לנקודה $\bar{x}_0$ עבורן $f\left(\bar{x}_0+\alpha\left(\bar{y}_0-\bar{x}_0\right)\right)<f\left(\bar{x}_0\right)$ ולכן $\bar{x}_0$ היא כלל לא נקודת מינימום מקומי, בסתירה לנתון. לכן ההנחה כי $\bar{x}_0$ היא לא מינימום גלובלי שגויה ולכן הוכחנו את הנדרש. מש"ל.

<u>משפט:</u>

נניח $C\in\mathbb{R}^n$ תחום קמור. לכל פונקציה קמורה <u>ממש</u> $f\left(\bar{x}\right):C\to\mathbb{R}$ יש <u>לכל היותר</u> מינימום גלובלי אחד בלבד.

<u>הוכחה:</u> (על דרך השלילה)

נניח $C\in\mathbb{R}^n$ תחום קמור. נתונה פונקציה קמורה <u>ממש</u> $f\left(\bar{x}\right):C\to\mathbb{R}$ ונניח כי לפונקציה יש לפחות שתי נקודות שנסמנן להיות $\bar{x}_0$ ו- $\bar{y}_0$ מינימום גלובליות. אם שתי הנקודות הן מינימום גלובלי אז ערך הפונקציה בנקודה בהכרח זהה ונסמנו להיות: $f\left(\bar{x}_0\right)=f\left(\bar{y}_0\right)=\mu$. נתון כי הפונקציה קמורה ממש ולכן לפי ההגדרה של פונקציה קמורה ממש ועבור $\alpha=\frac{1}{2}$ נקבל:

$$f\left(\frac{\bar{x}_0+\bar{y}_0}{2}\right)<\frac{f\left(\bar{x}_0\right)+f\left(\bar{y}_0\right)}{2}=\frac{\mu+\mu}{2}=\mu$$

ולכן מצאנו נקודה $\frac{\bar{x}_0+\bar{y}_0}{2}\in C$ אשר עבורה מקבלים ערך נמוך יותר מהערך של המינימום הגלובלי בסתירה להגדרת מנימום גלובלי ולכן לכל פונקציה קמורה ממש קיים לכל היותר מינימום גלובלי אחד.

<u>דוגמה:</u> (פונקציה קמורה ממש ללא מינימום גלובלי)

בתחום $C=\mathbb{R}^+\setminus\{0\}$ לפונקציה $f\left(x\right)=\frac{1}{x}$ אין מינימום גלובלי.

<u>תנאים לאופטימליות (של פתרון, כלומר מציאת נקודה בתחום של פונקציה עבורה נקבל את ערכה המינימלי של הפונקציה):</u>

<u>משפט:</u>

נניח כי $f\left(\bar{x}\right):\mathbb{R}^n\to\mathbb{R}$ פונקציה קמורה. מתקיים:

הנקודה $\bar{x}^*\in\mathbb{R}^n$ היא מינימום גלובלי ב- $\mathbb{R}^n$ $\Leftrightarrow$ $\nabla f\left(\bar{x}^*\right)=0$

<u>הוכחה:</u>

כיוון $\Rightarrow$:

לפי אי-שיוויון הגרדיאנט, מתקיים:

$$f\left(\bar{x}^*+\bar{r}\right)\ge f\left(\bar{x}^*\right)+r^T\nabla f\left(\bar{x}^*\right)$$

אבל נתון לנו כי $\nabla f\left(\bar{x}^*\right)=0$ ולכן מתקיים:

$$\forall\bar{r}\in\mathbb{R}^n:\quad f\left(\bar{x}^*+\bar{r}\right)\ge f\left(\bar{x}^*\right)+r^T\nabla f\left(\bar{x}^*\right)=f\left(\bar{x}^*\right)$$

ולכן $\bar{x}^*$ היא מינימום גלובלי לפי ההגדרה של מינימום גלובלי.

כיוון $\Leftarrow$:

נתון לנו כי $\bar{x}^*$ הוא מינימום גלובלי ומחדו"א אנחנו יודעים כי הגרדיאנט של כל פונקציה בנקודת מינימום (גם גלובלי וגם מקומי) מתאפס, כלומר $\nabla f\left(\bar{x}^*\right)=0$.

מש"ל.

<u>הערה</u>: <u>תנאים הכרחיים עבור פונקציות שאינן קמורות</u>.

1) עבור פונקציות שאינן קמורות, התנאי $\nabla f\left(\bar{x}^*\right)=0$ <u>הוא הכרחי, אך לא מספיק</u>! כלומר:

$$\text{הנקודה } \bar{x}^* \in \mathbb{R}^n \text{ היא מינימום גלובלי ב- } \mathbb{R}^n \;\cancel{\Leftrightarrow}\; \nabla f\left(\bar{x}^*\right)=0$$

למשל עבור הפונקציה $f\left(x\right)=x^3$, בנקודה אפס הגרדיאנט מתאפס אבל ברור כי הנקודה היא לא מינימום גלובלי.

2) עבור פונקציות שאינן קמורות, תנאי הכרחי נוסף שחייב להתקיים בכדי שנוכל לומר שנקודה מסויימת היא מינימום מקומי, חייבים לדרוש שעבור כל $\bar{r} \in \mathbb{R}^n$ נקבל כי הנגזרת הכיוונית מסדר שני בנקודה ה"חשודה כמינימום", תהיה גדולה או שווה לאפס, כלומר $f''_{\bar{r}\bar{r}}\left(\bar{x}^*\right)=\bar{r}^T H\left(\bar{x}\right)\bar{r}\geq 0$ וזו דרישה שקולה למעשה לדרישה על מטריצת ההסיאן להיות אי-שלילית: $H\left(\bar{x}\right)\succeq 0$ (נבחין כי היותה של מטריצת ההסיאן אי-שלילית זה לא תנאי מספיק להיותה של הפונקציה $f\left(\bar{x}\right)$ להיות קמורה).

<u>משפט</u>:

נניח כי $f\left(\bar{x}\right):\mathbb{R}^n \to \mathbb{R}$ פונקציה (לא בהכרח קמורה). מתקיים:

$$\begin{cases}\nabla f\left(\bar{x}^*\right)=0\\ H\left(\bar{x}^*\right)=\nabla^2 f\left(\bar{x}^*\right)\succ 0\end{cases} \Rightarrow \text{הנקודה } \bar{x}^* \in \mathbb{R}^n \text{ היא מינימום } \underline{\text{מקומי}} \text{ ב- } \mathbb{R}^n \text{ של הפונקציה } f$$

(הערה: במקום התנאי $H\left(\bar{x}^*\right)=\nabla^2 f\left(\bar{x}^*\right)\succ 0$ אפשר לדרוש במקום זו שהפונקציה f תהיה קמורה בסביבה $B_\varepsilon\left(\bar{x}^*\right)$)

<u>רעיון ההוכחה</u>: (שיעורי בית)

אם נתון לנו כי מטריצת ההסיאן חיובית בנקודה $\bar{x}^*$ אז גם קיימת סביבה של הנקודה $\bar{x}^*$ בה מטריצת ההסיאן היא חיובית, ולכן בסביבה זו אפשר להגדיר פונקציה קמורה ועבור פונקציה קמורה התנאי $\nabla f\left(\bar{x}^*\right)=0$ הוא תנאי מספיק להיותה של הנקודה $\bar{x}^*$ מינימום <u>גלובלי</u> (בסביבה של הנקודה $\bar{x}^*$) שלה, אבל הסביבה כולה של הפונקציה הקמורה היא סביבה מקומית של הפונקציה f ולכן הוכחנו כי $\bar{x}^*$ הוא מינימום מקומי של הפונקציה f .

בכך סיימנו את ההכנות המתמטיות הנדרשות לקורס ואפשר להתחיל ללמוד אלגוריתמים של אופטימיזציה.

הרצאה Iterative Methods of One Dimensional Optimization – 07, אלגוריתמים איטרטיביים (נומריים) של אופטימיזציה במשתנה יחיד

(החל מדקה 13:15)

לעיתים, ניתן לפתור בעיות אופטימיזציה באופן אנליטי ולהגיע לפתרון אנליטי. פתרון אנליטי טוב כאשר יש לנו מודל מדוייק שמתאר את הבעיה, או משוואות מדוייקות שאפשר לנתח. ברוב המקרים, המצב הוא לא כזה, ולכן אנחנו נאלצים לבצע חישובים נומריים שלעיתים מקרבים אותנו לפתרון באופן מיטבי ולפעמים פחות. אנחנו נתחיל בבעיה אופטימזציה של מציאת מינימום של פונקציה של משתנה יחיד.

<u>אלגוריתם Bisection</u>:

נניח כי יש לנו פונקציה $f(x)$ (משתנה יחיד) אשר יש לה מינימום מקומי (או גלובלי) בנקודה x^* (ונניח כי הפונקציה קמורה בסביבת הנקודה x^*). האלגוריתם מתבסס על חיפוש נקודה בה הנגזרת של הפונקציה שאנחנו מחפשים את המינימום שלה, מתאפסת. אם אנחנו מניחים כי x^* יש מינימום אז משמאל, הנגזרת $f'(x)$, שלילית ממש ומימין לנקודה x^* הנגזרת $f'(x)$ חיובית ממש ואילו בנקודה x^*, הנגזרת מתאפסת ממש. נסתכל על הקטע $[a,b]$ אשר נקודה x^* נמצאת בתוכו, כלומר $a<x^*<b$ וכעת אנחנו מחפשים את השורש של פונקצית הנגזרת, כלומר המרנו את בעית האופטימיזציה שלנו מבעית חיפוש מינימום של פונקציה לחיפוש שורש של פונקציה (אחרת מהפונקציה המקורית).

האלגוריתם למציאת השורש, הוא פשוט:

1) בחר מספר איטרציות התחלתי N (או בחר אורך קטע $[a,b]$ שבו תעצור) ובכל איטרציה בצע:
   a. אם $a=b$ החזר את הנקודה a וסיים את האלגוריתם (או אם אורך הקטע הוא קטן או שווה למה שבחרת).
   b. בחר את הנקודה האמצעית מבין נקודות הקצה $[a,b]$, כלומר בחר את נקודה $\tau=\dfrac{a+b}{2}$.
   c. אם ערך הנגזרת בנקודה τ היא חיובית, החלף את הקטע $[a,b]$ בקטע $[a,\tau]$ וחזור לשלב (a).
   d. אחרת, אם הנגזרת בנקודה τ היא שלילית, החלף את הקטע $[a,b]$ בקטע $[\tau,b]$ וחזור לשלב (a).

<u>סיבוכיות אלגוריתם Bisection</u>:

אם נגדיר את אורך הקטע ההתחלתי שבו אנחנו מחפשים את המינימום המקומי להיות $|b-a|=\Delta$ אז גודל הקטע Δ קטן בצורה אקספוננציאלית ככל שנבצע יותר איטרציות כיוון שבכל איטרציה אנחנו מקטינים את הקטע פי 2, כלומר גודל הקטע לאחר i איטרציות יהיה: $\dfrac{\Delta}{2^i}$.

<u>קשיים במימוש האלגוריתם Bisection</u>:

בכדי להשתמש באלגוריתם Bisection אנחנו צריכים לדעת את הקטע ההתחלתי וכן את הנגזרת של הפונקציה לה אנחנו מחפשים פתרון אופטימלי, אך לא תמיד יש בידינו את הנגזרת או את היכולת לחשב אותה ולכן אלגוריתם זה מוגבל יחסית.

<u>אלגוריתם חיתוך הזהב, Golden section</u>:

נניח כי יש בידינו קטע $[a,b]$ אשר אנחנו יודעים כי נקודת המינימום מוכלת בקטע הזה וכן אנחנו <u>יודעים</u> כי הפונקציה שלה אנחנו מחפשים את המינימום היא <u>קמורה</u> בקטע הזה. בנוסף, נניח כי יש לנו נקודה $a<c<b$ עבורה גילינו כי

$f(c)<f(a)$ וגם $f(c)<f(b)$ האם יש בידינו מספיק מידע על מנת לקבוע שנקודת המינימום נמצאת בקטע $[a,c]$ או בקטע $[c,a]$ בוודאות? לא! למשל:

כלומר, בחירת נקודה אחת בתוך הקטע $[a,b]$ לא מספיקה, אך מתברר כי בחירת שתי נקודות באופן מסויים, מאפשרת לנו כן לדעת באיזה קטע נמצאת נקודת המינימום, הבחירה נעשית למשל באמצעוצ אלגוריתם Golden section .

אופן פעולת האלגוריתם: (נניח כי יש לנו קטע התחלתי $[a,d]$ והמינימום מוכל בקטע זה)

1) אם $a=d$ עצור והחזר את נקודה a .
2) אחרת, מצא בקטע $[a,d]$ עוד שתי נקודות $a<b<c<d$ וחשב את ערך הפונקציה בנקודות הנוספות.
3) אם מתקיים $f(b)<f(c)$ החלף את הקטע $[a,d]$ בקטע $[a,c]$ וחזור לשלב (1).
4) אחרת, כלומר אם $f(b)\geq f(c)$, החלף את הקטע $[a,d]$ בקטע $[b,d]$ וחזור לשלב (1).

נבחין כי בכל פעם שאנחנו מבצעים את שלב (2) אנחנו צריכים לחשב את ערך הפונקציה בעוד שתי נקודות. מתברר שאין צורך בכך, כלומר באיטרציה הראשונה אכן נחשב את הערך של שתי הנקודות $f(b),f(c)$ אבל בכל האיטרציות הבאות מספיק לחשב את ערך הפונקציה בנקודה נוספת אחת (במקום בשתיים) כיוון שבשלב הקודם של האלגוריתם כבר חישבנו את אחת הנקודות $f(b),f(c)$ ואנחנו יכולים להשתמש בערכים אלו.

בנוסף, נחליט שאנחנו לא בוחרים באופן רנדומלי שתי נקודות b,c בכל קטע בשלב (2) אלא נבחר את הנקודות באופן הבא:

(מופיע שרטוט המחשה בסוף ההסבר)

נחליט על פרמטר חלוקה $\tau\in\mathbb{R}$ קבוע אשר נמצא עליו תנאים נוספים בהמשך.

נסמן את אורך הקטע $[a,d]$ באיטרציה האפס להיות Δ_0. נחלק את הקטע $[a,d]$ באופן כזה כך שנקבל:

$$[a,d]=[a,b]\cup[b,c]\cup[c,d]$$

וכן אורך כל אחד מהקטעים $[a,b],[c,d]$ הוא בדיוק $\tau\cdot\Delta_0$.

לאחר מכן, באיטרציה הבאה, אם למשל בחרנו להמשיך בקטע $[a,c]$ נחלק את הקטע $[a,c]$ באופן כזה, כך שאורך הקטע $[b,c]$ (כאשר b היא הנקודה שבחרנו באיטרציה הקודמת) הוא $\tau\cdot\Delta_1$ כאשר $\Delta_1=|c-a|$, כלומר Δ_1 הוא אורך הקטע של האיטרציה הראשונה (לאחר האיטרציה האפס).

מכאן, אנחנו מסיקים כי התנאי לבחירת $\tau\in\mathbb{R}$ מתאים הוא:

$$\tau\cdot\Delta_0=(1-\tau)\Delta_1$$

אך אפשר לבטא את Δ_1 באמצעות Δ_0 באופן הבא: $\Delta_1=(1-\tau)\Delta_0$

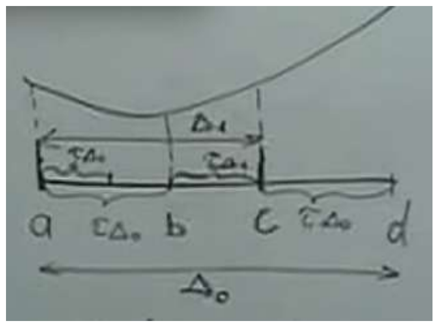
ולכן קיבלנו:

$$\tau\cdot\Delta_0=(1-\tau)\Delta_1=(1-\tau)^2\Delta_0 \;\Rightarrow\; \qquad \tau=(1-\tau)^2$$

וע"י פתירת משוואה ריבועית קלה אנחנו מקבלים:

$$\tau=\frac{3-\sqrt{5}}{2}$$

וזה למעשה פרמטר החלוקה של האלגוריתם. שרטוט להמחשה:



<u>אלגוריתם אינטרפולציה ריבועית, Quadratic Interpolation</u>:

כאשר השתמשנו באלגוריתם Golden section לא ייחסנו חשיבות לסוג הפונקציה שאנחנו מנתחים, לא ייחסנו חשיבות לעד כמה טוב היא מקורבת לפונקציה האמיתית (במידה וזה אכן קירוב) וכן לא התייחסנו עד כמה היא "חלקה"/גזירה. אם הפונקציה היא כן "חלקה" ו"טובה" אז אפשר להשתמש בידע שצברנו על נקודות מסויימות שכבר חישבנו עבור חישוב נקודות שעוד לא חישבנו, כלומר לבצע סוג של אינטרפולציה, כלומר אנחנו מקרבים את הפונקציה שיש ברשותנו לפונקציה אחרת אך שניתנת לביטוי אנליטי ואז אנחנו מחפשים את המינימום בצורה אנליטית – כלומר גוזרים ומשווים לאפס וכו'.

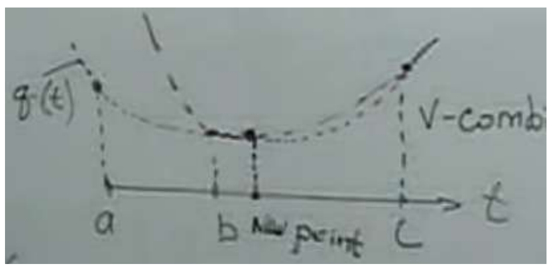
נניח ויש לנו פונקציה של משתנה יחיד $f(x)$ והיא "חלקה" ו"טובה" ונניח שאנחנו יודעים שיש לה מינימום בקטע $[a,c]$ ואנחנו מחפשים את נקודת המינימום (וכמובן גם את ערך הפונקציה בנקודה זו). נניח שאנחנו יודעים את ערכה של הפונקציה בשלוש נקודות, $a<b<c$ ומתקיים: $f(b)<f(c)$ וגם $f(b)<f(a)$ - כלומר יש לנו שלושה ערכים של הפונקציה. נוכל לקרב את הפונקציה $f(x)$ בקטע $[a,c]$ ע"י פונקציה ריבועית כללית: $q(x)=\alpha x^2+\beta x+\gamma$ ע"י פתירת שלוש משוואות בשלושה נעלמים (הנעלמים הם מקדמי המשוואה הריבועית: α,β,γ):

$$\begin{cases} q(a)=f(a)=\alpha a^2+\beta a+\gamma \\ q(b)=f(b)=\alpha b^2+\beta b+\gamma \\ q(c)=f(c)=\alpha c^2+\beta c+\gamma \end{cases}$$

כעת לאחר שיש לנו משוואה ריבועית, $q(x)=\alpha x^2+\beta x+\gamma$, עם מקדמים ידועים, אפשר למצוא את המינימום של משוואה זו ע"י גזירה והשוואת הנגזרת לאפס וכו' וכך למצוא (בצורה אנליטית) את הנקודה שבה מתקבל המינימום של המשוואה הריבועית. האם זו הנקודה שבה מתקבל המינימום של משוואה $f(x)$? לא בהכרח! (מדוע? כיוון שאף אחד לא אמר שמתקיים $f(x)\equiv q(x)$ לכל נקודה בקטע $[a,c]$, אלא אם הפונקציה $f(x)$ היא אכן ריבועית)

נרצה להמשיך למצוא את הנקודה בה מתקבל המינימום של פונקציה $f\left(x\right)$. נבחין כי הנחנו כי מתקיים $f\left(b\right)<f\left(c\right)$ וכן $f\left(b\right)<f\left(a\right)$ ולתבנית זו אנחנו קוראים v-combination כי הצורה ששלושת הנקודות יוצרות על הגרף דומה לצורה "v":

נניח כי נקודת המינימום של המשוואה הריבועית היא $b<m<c$ (כלומר מוכלת בקטע $\left(b,c\right)$) , אזי נוכל להמשיך בצורה איטרטיבית את האלגוריתם על הקטע $\left[b,c\right]$ עם הנקודות b,m,c וזה עדיף כי אנחנו רוצים לשמר את הצורה v-combination כאשר הנקודה החדשה שמצאנו היא הנקודה הנמוכה ביותר ב-"v" כיוון ששימור זה מבטיח את יציבות והתקדמות האלגוריתם.



נבחין שבאלגוריתם זה, גודל הקטע אינו קטן באופן זהה מאיטרציה לאיטרציה, כלומר ייתכן ואנחנו נבצע הרבה איטרציות באלגוריתם אבל האלגוריתם לא יתקרב לפתרון ולכן ניתן באמצע הפעלת האלגוריתם (כאשר מגלים חוסר התקדמות), לעבור למשל לאלגוריתם אחר כדוגמת Golden section.

האלגוריתם Quadratic Interpolation:

1) מצא בקטע $\left[a,c\right]$ נקודה שיוצרת $v-combination$, נסמנה להיות b .
2) מצא באופן אנליטי את המינימום של הפונקציה הריבועית שנוצרה ע"י $v-combination$, נסמנה להיות m .
3) בחר מתוך 4 הנקודות שנוצרו עד כה עוד $v-combination$ וחזור לשלב (2). (אם ההתקדמות באלגוריתם לא מספקת ניתן לעבור לאלגוריתם Golden section).

<u>מתי האלגוריתם Quadratic Interpolation יעיל</u>?

אלגוריתם זה שמיש ויעיל במיוחד כאשר הפונקציה שעבורה אנחנו מחפשים את המינימום היא אכן בעלת קירוב ריבועי טוב בסביבת נקודת המינימום שאנחנו מחפשים.

<u>קצב התכנסות אלגוריתם Quadratic Interpolation:</u>

אם נסמן את ערך הפונקציה בנקודת מרכז ה-"v" באיטרציה ה- k של האלגוריתם להיות f_k וכן נסמן את ערך הפונקציה בנקודת המינימום שאנחנו מחפשים להיות f^* אז נוכל לומר כי האלגוריתם ייתכנס אם יתקיים אי-השיוויון הבא:

$$f_k-f^*\le c_k\left(f_{k-1}-f^*\right)$$

(כאשר ברור כי אם האלגוריתם מתכנס בהכרח מתקיים $c_k < 1$)

אם בשלב מסויים, הערך $c_k = c_{k-1}$, כלומר אנחנו מקבלים c_k קבוע (מאיטרציה לאיטרציה) אז אנחנו אומרים שקצב התכנסות האלגוריתם הוא לינארי. באלגוריתם Quadratic Interpolation מתקיים כי $\lim_{k\to\infty} c_k = 0$ ולהתכנסות זו קוראים סופר-לינארית.

אלגוריתם Cubic Interpolation:

אם נניח כי עבור הפונקציה $f(x)$, אנחנו יודעים לחשב את הנגזרת של הפונקציה בכל נקודה שאנחנו נרצה (בין אם מישהו ייתן לנו את ערך הנגזרת שאנחנו יכולים למדוד זו בצורה כלשהי) אז אפשר להשתמש באינטרפולציה חכמה אף יותר מהאינטרפולציה הריבועית. נניח ואנחנו יודעים כי $f(x)$ מקבלת את המינימום שלה בקטע $[a,c]$ אז במקום למצוא נקודה נוספת וליצור $v-combination$ ולקרב אותה באמצעות משוואה ריבועית, אנחנו יכולים למצוא $v-combination$ בצורה אחרת. אם למשל אנחנו יודעים כי $f'(a) < 0$ וגם $f'(b) > 0$ אז למעשה אנחנו מקבלים גם $v-combination$ וכעת נוכל לקרב את הפונקציה $f(x)$ באמצעות משוואה מסדר 3 עם מקדמים $\alpha,\beta,\gamma,\delta$ שאנחנו מחפשים:

$$q(x) = \alpha x^3 + \beta x^2 + \gamma x + \delta$$

כיוון שיש לנו את 4 המשוואות הבאות (עם 4 נעלמים):

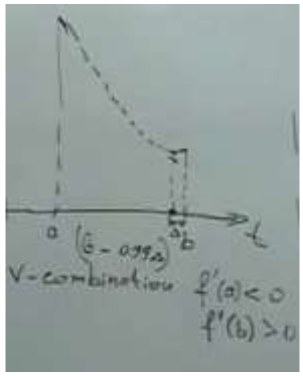
$$\begin{aligned}
q(a) &= f(a) = \alpha a^3 + \beta a^2 + \gamma a + \delta \\
q'(a) &= f'(a) = 3\alpha a^2 + 2\beta a + \gamma \\
q(b) &= f(b) = \alpha b^3 + \beta b^2 + \gamma b + \delta \\
q'(b) &= f'(b) = 3\alpha b^2 + 2\beta b + \gamma
\end{aligned}$$

לאחר מציאת 4 המקדמים $\alpha,\beta,\gamma,\delta$ אפשר למצוא בצורה אנליטית די פשוטה את הנקודה עבורה מתקבל המינימום של הפונקציה $q(x)$ שנסמנה להיות b אז לחפש מבין 3 הנקודות a,b,c פעם נוספת אילו שתי נקודות תיתנה לנו $v-combination$ ולהמשיך משם (זה אפשרי כי הנחנו כי גם עבור נקודה b אנחנו יכולים למצוא את ערך הנגזרת של פונקציה $f(x)$).

קצב התכנסות אלגוריתם Cubic Interpolation:

אם ההתקדמות באלגוריתם לא מספיקה לנו אנחנו כמובן יכולים לבצע אלגוריתמים אחרים כמו Bisection.

ייתכן גם מצב בו האלגוריתם, כאשר הוא מנסה להקטין את הקטע עליו הוא מחפש מינימום, מוצא נקודה חדשה b שמאד קרובה לקצה, אז ייתכן שהנקודה החדשה תהיה קרובה מידי לקצה וזה יכול לגרום לנו לעשות צעד הקטנה קטן, למשל:



נניח כי נקודת המינימום היא מאד מאד קרובה לקצה הימיני של הקטע. אם הנקודה החדשה שמצאנו בצעד כלשהו של האלגוריתם, קרובה עד כדי Δ לקצה הימיני של הקטע אבל עדיין מצד שמאל לנקודת המינימום שלה אז השיפוע שלה הוא קטן מאפס, וזה טוב לנו, כי בצעד הבא אנחנו ניקח את הקטע $\left[b-\Delta, b\right]$ שזה קטע קטן מאד והאלגוריתם מתקדם בצורה טובה. אך אם במקרה קרה שהנקודה החדשה נמצאת במצד ימין של נקודת המינימום, ולכן שיפוע הפונקציה בנקודה זו הוא חיובי, אז בצעד הבא של האלגוריתם, אשר בו אנחנו נדרשים למצוא $v-combination$ אנחנו נחפש בקטע $\left[a, b-\Delta\right]$ ואז למעשה אנחנו מתקדמים מאד לאט באלגוריתם, ולכן נגדיר כי אם הנקודה החדשה היא קרובה מידי לקצה, למשל קרובה יותר מערך $\Delta_{\max}$ שנגדיר מראש, אז אנחנו ניקח נקודה $b-\Delta_{\max}$ (או $a+\Delta_{\max}$ אם הנקודה החדשה קרובה לקצה השמאלי של הקטע) ובכך נוכל "לשמור" האלגוריתם בכך שהוא יבצע התקדמות טובה בכל צעד.

הרצאה 08 – Multidimensional, Unconstrained Optimization Methods, אלגוריתמים לאופטימיזציה ללא אילוצים של פונקציות בעלות מספר משתנים

בעית אופטימיזציה כללית:

למצוא את הנקודה $\bar{x} \in \mathbb{R}^n$ עבורה נקבל את הערך המינימלי של הפונקציה $f(\bar{x}): \mathbb{R}^n \to \mathbb{R}$ שהיא פונקציה של מספר משתנים.

ישנן שיטות רבות לפתרון בעיות אופטימיזציה ושיטות אלו נהוג לחלק למחלקות.

שיטות מסוג Line Search:

שיטה זו היא שיטה איטרטיבית. באיטרציה ה- k האלגוריתם נמצא בנקודה $\bar{x}_k$ (והיא מועמדת להיות נקודת המינימום שאותה אנחנו מחפשים) והיא מחושבת ע"י הנוסחה הבאה:

$$\bar{x}_k = \bar{x}_{k-1} + \alpha_k \cdot \bar{d}_k \qquad , \qquad \alpha_k \in \mathbb{R},\ \bar{d}_k \in \mathbb{R}^n$$

כלומר הנקודה $\bar{x}_k$ היא הנקודה מהאיטרציה הקודמת ($k-1$) עם "הליכה" בכיוון $\bar{d}_k$ ועם גודל צעד (step size) של α_k, כלומר האלגוריתם מחפש את נקודת המינימום באמצעות הליכה על קו כלשהו (ייתכן שכיוון הקו משתנה), של ערכים בפונקציה ובכדי שזה יהיה אלגוריתם טוב נראה שלאורך קו זה ערכי הפונקציה בנקודות לאורך הקו ילכו וירדו (כי אנחנו מחפשים את מינימום הפונקציה). בהמשך נראה כי נוח מאד למשל להגדיר את כיוון ההליכה ככיוון המנוגד לכיוון הגרדיאנט, שכן הגרדיאנט מצביע בכיוון העליה החדה ביותר של ערכי הפונקציה.

בכדי שכיוון ה"הליכה" יהיה בכיוון שבו ערכי הפונקציה יורדים, בהכרח חייב להתקיים התנאי על הנגזרת הראשונה של הפונקציה, כלומר הנגזרת הראשונה חייבת להיות שלילית, ובפונקציה של מספר משתנים אנחנו אומרים כי זו הנגזרת הכיוונית, כלומר נדרוש שהנגזרת הכיוונית בכיוון ההליכה תהיה תמיד שלילית:

$$f'_{\bar{d}_k}(\bar{x}_k) = \left(\nabla f(\bar{x}_k)\right)^T \cdot \bar{d}_k < 0$$

בחירת גודל הצעד α_k:

בכל איטרציה אנחנו נדרשים לספק לאלגוריתם את גודל הצעד שאנחנו מעוניינים לבצע. לבחירת גודל הצעד יש כמה אפשרויות:

1) שיטה שנקראת Exact Line Search:

בשיטה זו אנחנו בוחרים את α_k להיות הערך שמביא את הביטוי $f(\bar{x}_{k+1}) = f(\bar{x}_k + \alpha_k \cdot \bar{d}_k)$ למינימום, כלומר:

$$\alpha_k = \arg\min_{\alpha} f(\bar{x}_k + \alpha \bar{d}_k)$$

למעשה זו בעית אופטימיזציה במשתנה יחיד בפני עצמה וראינו דרכים לפתור בעיה כזו – למשל ע"י אלגוריתם Bisection, או אינטרפולציה וכו'. נשים לב כי פתרון בעית האופטימיזציה הזו היא רק חלק מבעית האופטימיזציה הכללית שאנחנו מנסים לפתור ולכן רצוי לא לבזבז יותר מידי זמן בחיפוש α_k אופטימלי – צריך למצוא איזון בין חיפוש α_k בכל איטרציה לבין הזמן שלוקח לנו למצוא פתרון אופטימלי לבעיה הכללית שלנו.

2) שיטת Inexact Line Search:

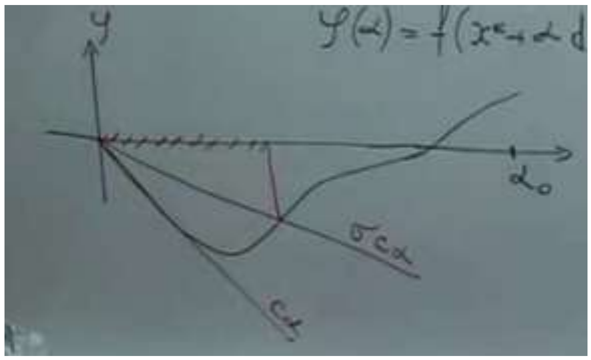
דוגמה לשיטה זו היא עבודה לפי חוק Armijo. בשיטה זו ננסה למצוא את α_k באופן הבא:

$$\alpha_k = \arg\min_{\alpha} \varphi(\alpha) = \arg\min_{\alpha} \left\{ f\left(\bar{x}_k + \alpha \bar{d}_k\right) - f\left(\bar{x}_k\right) \right\}$$

חוק Armijo מציע להסתכל על הבעיה באופן הבא: נבחין כי מתקיים לפי מה שהגדרנו זה עתה $\varphi(0) = 0$ וכן מתקיים (לפי הדרישה על הנגזרת הכיוונית):

$$\varphi'(\alpha)\Big|_{\alpha=0} = \left(\nabla f\left(\bar{x}_k\right)\right)^T \bar{d}_k < 0$$

נסמן את השיפוע של $\varphi(\alpha)$ בנקודה $\alpha = 0$ להיות $c < 0$. כמו כן, החוק אומר שצריך להגדיר קו לינארי נוסף בעל שיפוע $c\sigma < 0$ ואז נוכל לשרטט את הפונקציה $\varphi(\alpha)$ באופן הבא (זו הקו העקום, זו פונקציה עם משתנה ממימד אחד):



(אנחנו מנסים למצוא את הנקודה α עבורה נקבל את הערך המינימלי של הפונקציה $\varphi(\alpha)$)

כעת החוק אומר שאם נתחיל לפתור את בעית האופטימיזציה $\alpha_k = \arg\min_{\alpha} \varphi(\alpha)$ החל מ- α_0 כלשהו, אז ברגע שנגיע לנקודה α_k שבו אנחנו מקבלים כי:

$$\varphi(\alpha_k) < \sigma c \alpha_k$$

במצב זה אנחנו יכולים לעצור ולהחזיר את הערך α_k שמצאנו. ההתקדמות בציר α נעשית ע"י הכפלה בפרמטר $0 < \beta < 1$ (למשל $\beta = 0.2$) ובכך אנחנו מבטיחים להתקדם לכיוון ראשית הצירים. בשיטה זו יש כמה שיקולים שצריך לקחת בחשבון, אם נקבל גודל צעד מאד קטן ($\alpha_k \cong 0$) אז אנחנו נקבל את ההתקדמות בכיוון הנכון ביותר (מבחינת הנגזרת הכיוונית) אך ההתקדמות הכוללת באלגוריתם תהיה מאד קטנה, מצד שני, אם נבחר צעדים גדולים, יש סיכוי גבוה יותר שאנחנו הולכים בכיוון שהוא לא מדוייק מספיק על מנת להגיע לפתרון האופטימלי בצורה המהירה ביותר ולכן הפרמטר σ מאפשר לנו סוג של פשרה. באופן כללי נהוג לבחור $0 < \sigma < 1$. לעיתים נבחר אפילו $\sigma = 10^{-4}$.

3) <u>שיטת Constant step size</u>:
בשיטה זו אנחנו קובעים מראש כי גודל הצעד הוא גודל (קטן יחסית) קבוע $\alpha = Const$. הסיכון בשיטה זו הוא שאם בוחרים בטעות קבוע גדול מידי, ייתכן והאלגוריתם יתבדר(!) וכלל לא יתכנס לפתרון האופטימלי שאנחנו מחפשים. לעומת זאת, אם נבחר גודל צעד קבוע קטן מידי, האלגוריתם יתכנס אך באופן איטי למדי.

4) <u>שיטת Diminishing step size</u>:
בשיטה זו אנחנו מקטינים את גודל הצעד מאיטרציה לאיטרציה, אך צריך לדאוג כי השאיפה של גודל הצעד לאפס <u>לא מתרחשת מהר מידי</u> כי אז האלגוריתם יתכנס באופן איטי מאד לפתרון. נהוג לבחור קצב "הקטנה" באופן הבא:

$$\alpha_k \to 0 \quad , \quad \lim_{l\to\infty} \sum_{i=1}^{l} \alpha_k \to \infty$$

למשל ניתן לבחור $\alpha_k = \frac{1}{k}$ שזה הטור ההרמוני שידוע מחדו"א 1 כי הוא לא מתכנס.

דוגמה: (לשימוש באלגוריתם Line Search לפי Steepest Descent, כלומר הולכים בכיוון הירידה החדה ביותר, או בשם אחד זה נקרא Gradient Descent)

נשתמש באלגוריתם:

$$\bar{x}_k = \bar{x}_{k-1} + \alpha_k \cdot \bar{d}_k \qquad , \qquad \alpha_k \in \mathbb{R},\ \bar{d}_k \in \mathbb{R}^n$$

כאשר נגדיר כי כיוון הצעד, הוא הכיוון המנוגד לכיוון הגרדיאנט (כי הגרדיאנט מצביע לכיוון העליה החדה ביותר בעוד אנחנו מחפשים את הירידה החדה ביותר), כלומר:

$$\bar{d}_k = -\nabla f\left(\bar{x}_k\right)$$

ולכן הצעד שלנו הוא:

$$\bar{x}_k = \bar{x}_{k-1} - \alpha_k \cdot \nabla f\left(\bar{x}_k\right)$$

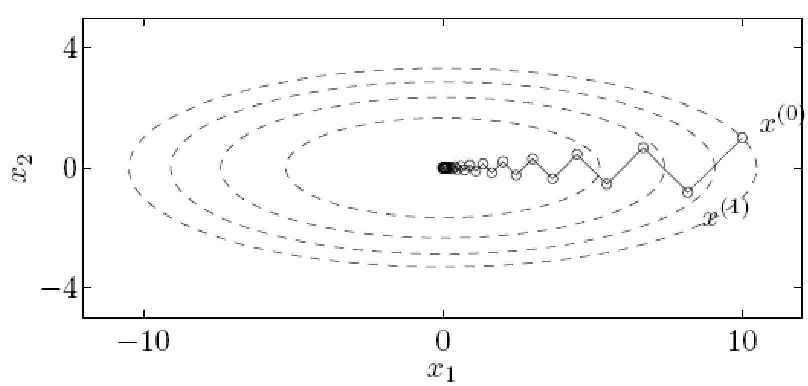
נניח כי $f\left(\bar{x}\right) : \mathbb{R}^2 \to \mathbb{R}$ ואנחנו נשרטט את קווי הגובה של הפונקציה במערכת הצירים $\left(x_1, x_2\right)$.

בשביל לחשב את גודל הצעד בכל איטרציה נשתמש בשיטת Exact Line Search ולכן אם בכל שלב נתקדם על הקו שמנוגד לכיוון הגרדיאנט, אנחנו נבחר את גודל הצעד שייתן לנו את הנקודה לאורך קו (הגרדיאנט) שהפונקציה מקבלת בו את ערכה המינימלי. באיטרציה הבאה, אנחנו שוב נבדוק את כיוון הגרדיאנט בנקודה שבה אנחנו עומדים ונבצע את אותו התהליך של מציאת גודל הצעד, שביצענו, פעם נוספת.

אם למשל קווי הגובה הן אליפסות שמאונכות לאחד הצירים, אנחנו יכולים אולי להניח כי אם הפונקציה היא מהצורה $f\left(\bar{x}\right) = \bar{x}^T A \bar{x}$ אז יש למטריצה A ערכים עצמיים, אחד גדול ואחד קטן, כי אם למשל המטריצה היא אלכסונית (ואז ערכי האלכסון אלו הם הערכים העצמיים) אז נקבל:

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \ \Rightarrow \qquad f\left(\bar{x}\right) = \bar{x}^T A \bar{x} = x_1\lambda_1 + x_2\lambda_2$$

ולכן אם $\lambda_1 \ll \lambda_2$ אז אנחנו רואים כי עבור שינויים קטנים ב- x_2 נקבל שינויים גדולים בערכי הפונקציה ועבור שינויים קטנים ב- x_1 נקבל שינויים קטנים. ואם המטריצה A היא לא אלכסונית אז כיוון האליפסות יהיה במאונך לאחד מהווקטורים העצמים ומקביל לשני (כי ווקטורים עצמיים הם אורותוגנוליים).



(האליפסות מאונכות לציר האנכי ומקבילות לציר האופקי)

אם נמשיך באלגוריתם שלנו, אנחנו נקבל דרך שהיא סוג של "זיג-זג" וגודל הצעדים ילכו ויקטנו מאיטרציה לאיטרציה.

קצב התכנסות האלגוריתם:

לא נוכיח בדוגמה זו את קצב ההתכנסות, אך קצב ההתכנסות של האלגוריתם הוא <u>קצב התכנסות לינארי</u> של הירידה החדה ביותר.

<u>טענה</u> (שלא נוכיח): אם נניח כי ההסיאן של הפונקציה $f\left(\bar{x}\right)$ הוא $H\left(\bar{x}\right)$ והערכים העצמיים שלו הם $\lambda_{\min},\lambda_{\max}$ (זו מטריצה מסדר 2 ולכן יש רק שני ערכים עצמיים) אז מתקיים:

$$\boxed{f\left(\bar{x}_{k+1}\right)-f\left(\bar{x}^{*}\right)\le c\left(f\left(\bar{x}_{k}\right)-f\left(\bar{x}^{*}\right)\right)}$$

כאשר $f\left(\bar{x}^{*}\right)$ זה הערך המינימלי של הפונקציה $f\left(\bar{x}\right)$ וכן $c\in\mathbb{R}$ הוא קבוע.

לקשר זה בין שתי איטרציות באלגוריתם קוראים התכנסות לינארית, והקבוע c נקרא קצב ההתכנסות. במקרה זה מתקיים:

$$\boxed{c=1-\frac{\lambda_{\min}}{\lambda_{\max}}}$$

במקרה בו מתקיים $\lambda_{\min}=\lambda_{\max}$ מתקיים $c=0$ וזה אומר:

$$f\left(\bar{x}_{k+1}\right)-f\left(\bar{x}^{*}\right)\le c\left(f\left(\bar{x}_{k}\right)-f\left(\bar{x}^{*}\right)\right)=0$$

וזה אומר כי האלגוריתם יתכנס כבר לאחר צעד אחד. נוכל לומר כי אם $\lambda_{\min}=\lambda_{\max}$ אז קווי הגובה של הפונקציה $f\left(\bar{x}\right)$ הם עיגולים מושלמים, ולכן כבר בצעד הראשון הכיוון המנוגד לגרדיאנט הוא בדיוק לכיוון המינימום שאותו אנחנו מחפשים ולכן ע"י שיטת Exact Line אנחנו <u>כבר אחרי צעד אחד</u> נמצא את המינימום הנדרש.

אם במקרה אחר, נקבל למשל: $\frac{\lambda_{\min}}{\lambda_{\max}}=10^{-3}$ אז נקבל $c=0.999$ וזה אומר כי האלגוריתם יתכנס בצורה לינארית אך בצורה <u>איטית מאד</u>.

<u>הגדרה</u>: Condition Number

נהוג לסמן $\frac{\lambda_{\max}}{\lambda_{\min}}=\theta$ וזה נקרא condition number (נבחין כי זה היחס ההפוך בין הערכים העצמיים). אנחנו נוהגים לומר כי אם θ הוא גדול אז מטריצת ההסיאן היא ill defined , וקצב ההתכנסות של האלגוריתם הוא איטי ולכן במקרים כאלה נרצה ללמוד שיטות נוספות לאופטימיזציה.

<u>שיטת ניוטון: (Newton method)</u>:

זוהי שיטה הרבה יותר "חזקה" משיטת Line search.

נניח תחילה כי נתונה פונקציה $f\left(x\right):\mathbb{R}\to\mathbb{R}$ וערכה המינימלי מתקבל בנקודה x^{*} ואנחנו נמצאים באיטרציה k של האלגוריתם בנקודה x_k . אם אנחנו נניח למשל כי הפונקציה היא דומה לפונקציה ריבועית אז נרצה לקרב אותה באמצעות פונקציה ריבועית (פולינום). לאחר מכן, נמצא את המינימום של הפולינום באופן אנליטי ונפעיל את האלגוריתם על הנקודה שמצאנו שהיא המינימום של הפולינום. למעשה נקבל את השרטוט הבא:

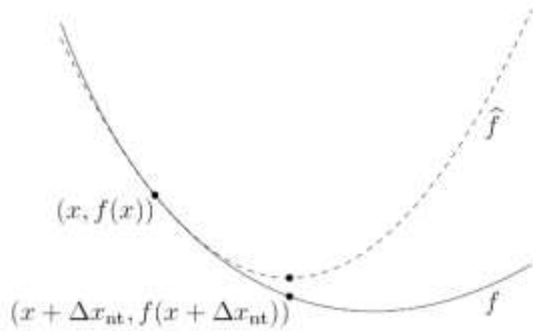


Figure 9.16 The function f (shown solid) and its second-order approximation $\widehat{f}$ at x (dashed). The Newton step $\Delta x_{\rm nt}$ is what must be added to x to give the minimizer of $\widehat{f}$.

<u>השיטה: (שיטת ניוטון)</u>

עבור פונקציה של מספר משתנים $f\left(\bar{x}\right):\mathbb{R}^n \to \mathbb{R}$ ונסמן:

$$\begin{aligned} \bar{g}_k &\equiv \bar{g}\left(\bar{x}_k\right) \equiv \nabla f\left(\bar{x}_k\right) \\ H_k &\equiv H\left(\bar{x}_k\right) \equiv \nabla^2 f\left(\bar{x}_k\right) \end{aligned}$$

וכן פיתח טיילור מסדר שני של הפונקציה $f\left(\bar{x}\right)$ סביב הנקודה $\bar{x}_k$ יסומן להיות:

$$f\left(\bar{x}_k+\bar{d}_k\right)=f\left(\bar{x}_k\right)+\bar{g}_k{}^T\bar{d}_k+\frac{1}{2}\bar{d}^T H\left(\bar{x}_0\right)\bar{d}+\ldots \qquad \Rightarrow q\left(\bar{d}_k\right)\equiv f\left(\bar{x}_k\right)+\bar{g}_k{}^T\bar{d}_k+\frac{1}{2}\bar{d}^T H\left(\bar{x}_0\right)\bar{d}$$

שיטת ניוטון ממזערת את הקירוב הריבועי $q\left(\bar{d}_k\right)$ ובכדי למצוא מינימום של פונקציה ריבועית עם מספר משתנים נדרוש שהגרדיאנט (גוזרים לפי $\bar{d}_k$) יתאפס:

$$Grad\ \left(q\left(\bar{d}_k\right)\right)=\nabla q\left(\bar{d}_k\right)=\bar{g}_k+H_k\bar{d}_k=0$$

משוואה זו נקראת <u>משוואת ניוטון</u>. כמו כן, למדנו כי מתקיים $H_k\bar{d}_k=-\bar{g}_k$, ולכן על מנת לחשב את $\bar{d}_k$ שהוא כיוון ההתקדמות אפשר למעשה לחשב את הביטוי $\bar{d}_k=-\dfrac{\bar{g}_k}{H_k}$ (ע"י מטלב למשל) ואז לבחור את גודל הצעד ע"י אלגוריתם כדוגמת Exact Line search.

נבחין שבשביל למצוא את הנקודה בה הגרדיאנט $\nabla q\left(\bar{d}_k\right)$ מתאפס, אנחנו למעשה פותרים בעית אופטימיזציה במספר משתנים ולכן ניתן לפתור בעיה זו בשיטת Line search. כיוון הצעד הוא:

$$\bar{d}_k=-\frac{\bar{g}_k}{H_k}=-H_k^{-1}\cdot\bar{g}_k$$

כאשר $\bar{d}_k$ נקרא הכיוון של ניוטון (Newton's direction) ואז באיטרציה הבא נקבל:

$$\overline{x}_{k+1} = \overline{x}_k + \alpha_k \overline{d}_k$$

ואת גודל הצעד ניתן לחשב למשל לפי Exact line search או בדרך אחרת.

את שיטת ניוטון ניתן לבצע לא רק באמצעות Line search אלא גם בשיטות האחרות שלא נדון בהן בקורס זה אך אחת מהן היא למשל Trust Region (ניתן למצוא בספר הלימוד).

<u>קצב התכנסות האלגוריתם של שיטת ניטון – התכנסות ריבועית אסימפוטוטית</u>:

ברור כי במידה והפונקציה $f\left(\overline{x}\right)$ היא אכן ריבועית, אז תספיק לנו איטרציה אחת בלבד של האלגוריתם על מנת להגיע לפתרון האופטימלי כיוון שאם היא ריבועית אז הקירוב שלנו (שהוא פולינום ריבועי) הוא בדיוק הפונקציה שאנחנו מחפשים עבורה את המינימום.

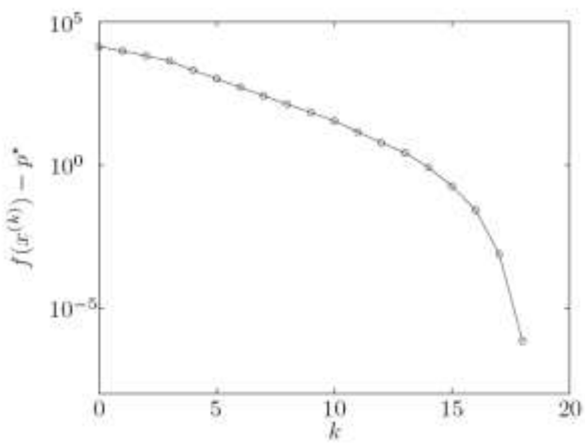
במקרה בו הפונקציה $f\left(\overline{x}\right)$ היא לא ריבועית ובנקודת האופטימום מתקיים כי $H\left(\overline{x}\right) \succ 0$ וגם $H\left(\overline{x}\right)$ רציפה, אז ההתכנסות היא <u>התכנסות ריבועית אסימפוטוטית</u>, כלומר אם באיטרציה מסויימת נגדיר:

$$\delta_k \triangleq \left\| \overline{x}_k - \overline{x}^* \right\| \quad or \quad \delta_k \triangleq \left\| f\left(\overline{x}_k\right) - f\left(\overline{x}^*\right) \right\|$$

אז מתקיים:

$$\delta_{k+1} = c\delta_k^2$$

ולכן אם באיטרציה מסויימת מתקיים $\delta_k = 10^{-2}$ אז עבור $c = 1$ נקבל כבר באיטרציה הבאה $\delta_{k+1} = 10^{-4}$ ולאחר מכן $\delta_{k+2} = 10^{-8}$ וזו התכנסות מאד מהירה (אקספוננציאלית). <u>התכנסות אסימפטוטית מתרחשת רק כאשר האלגוריתם כבר קרוב לפתרון</u> והבעיה היא שאנחנו לא יודעים מתי האלגוריתם מתקרב לפתרון ולכן רק לאחר שאנחנו רואים כי האלגוריתם קרוב לפתרון אנחנו נוכל לדעת כי ההתכנסות החל משלב זה היא מאד מהירה. אפשר לתאר את הגרף של השגיאה בצורה הבאה:



הרצאה 09 – Another view of Newton's Meth. Via solution of system of nonlinear equations , מבט נוסף על שיטת ניוטון באמצעות פתירת מערכת משוואות לא לינאריות

כאשר נתונה לנו בעיית אופטימיזציה, למשל למצוא את $\bar{x}\in\mathbb{R}^n$ אשר יביא את ערך הפונקציה $f(\bar{x}):\mathbb{R}^n\to\mathbb{R}$ למינימום ניתן במקום זאת למצוא את $\bar{x}\in\mathbb{R}^n$ אשר מאפס את הגרדיאנט של הפונקציה, כלומר מקיים $\nabla f(\bar{x})\equiv\bar{g}(\bar{x})=0$ כי למדנו <u>שאלו הם תנאי מספיקים עבור פונקציה שהיא קמורה</u>. כיוון שהגרדיאנט הוא ווקטור, ואנחנו רוצים שהוא יהיה שווה לווקטור האפס, זה אומר שאנחנו למעשה צריכים לדרוש שכל הרכיבים שלו יתאפסו, ולכן למעשה יש לפנינו מערכת משוואות לא לינאריות (כי באופן כללי הגרדיאנט לא חייב להיות לינארי), כלומר אנחנו צריכים לפתור מערכת של n משוואות לא לינאריות:

$$\bar{g}(\bar{x})=\bar{0}\quad\Leftrightarrow\quad g_i(\bar{x})=0\ ,\ \forall i=1,\dots,n$$

נניח ואנחנו עומדים בנקודה מסויימת $\bar{x}_k$, אז אמרנו כי הרכיבים של $\bar{g}(\bar{x}_k)$ הם לא בהכרח לינאריים, אך אם נכתוב את פיתוח טיילור עד <u>סדר ראשון</u> של כל רכיב נקבל n משוואות לינאריות:

$$\bar{g}(\bar{x}_k+\bar{d}_k)\cong g(\bar{x}_k)+H(\bar{x}_k)\bar{d}_k=0$$

ולכן הפתרון למערכת המשוואות נתון ע"י:

$$\bar{d}_k=-H^{-1}(\bar{x}_k)g(\bar{x}_k)$$

וזה בדיוק צעד ניוטון שלמדנו בהקשר של שיטת ניוטון. נבחין כי אנחנו רוצים לדרוש כי כיוון הווקטור $\bar{d}_k$ יהיה בכיוון בו ערכי הפונקציה $f(\bar{x})$ יורדים, כלומר נדרוש כי הנגזרת הכיוונית $f'_{\bar{d}_k}(\bar{x})<0$. כמו כן:

$$f'_{\bar{d}_k}(\bar{x})=\bar{g}_k^T\bar{d}_k\underset{\bar{d}_k=-H^{-1}(\bar{x}_k)g(\bar{x}_k)}{=}-\bar{g}_k^TH_k^{-1}\bar{g}_k<0\quad\Leftrightarrow\quad\bar{g}_k^TH_k^{-1}\bar{g}_k>0$$

וזו דרישה שקולה לדרישה: $H(\bar{x}_k)\succ 0$ (זה דרישה שקולה כי אם $H^{-1}(\bar{x}_k)\succ 0$ אז זה שקול לזה שהערכים העצמים של המטריצה $H^{-1}(\bar{x}_k)$ הם חיובים, והערכים העצמיים של המטריצה $H(\bar{x}_k)$ הם "1 חלקי" הערכים העצמיים של $H^{-1}(\bar{x}_k)$ ולכן זה אכן דרישה שקולה). נבחין כי במידה והפונקציה $f(\bar{x})$ לא קמורה ממש (וייתכן אפילו לא קמורה) אז אין הכרח שהתנאי יתקיים $H(\bar{x}_k)\succ 0$ (לפי משפט שראינו: אם $f(\bar{x})$ קמורה (ממש) אז מתקיים $H(\bar{x}_k)\succeq 0$ ($H(\bar{x}_k)\succ 0$)) אבל אז זה אומר כי יש ערכים עצמיים שהם שליליים ולכן נוכל להשתמש ב"טריק" ולמעשה להוסיף להיסאן מטריצה אלכסונית Δ_k כך שיתקיים:

$$H(\bar{x}_k)+\Delta_k\succ\varepsilon I$$

ומשמעות הביטוי האחרון זה אומר כי <u>כל הע"ע</u> של המטריצה באגף השמאלי, <u>גדולים</u> מהע"ע של המטריצה באגף הימיני. אפשר באותה מידה לכתוב זאת כך (נשאיר את ההוכחה לשיעורי הבית):

$$H(\bar{x}_k)+\Delta_k-\varepsilon I\succ 0$$

וכעת נמצא את כיוון הצעד של שיטת ניוטון באמצעות:

$$\bar{d}_k = -\left(H\left(\bar{x}_k\right) + \Delta_k\right)^{-1} g\left(\bar{x}_k\right)$$

<u>פתרון של מערכת משוואות $A\bar{x} = \bar{b}$ כאשר המטריצה A סימטרית חיובית:</u>

ישנן שיטות רבות לפתירת מערכת משוואות לינאריות ואנחנו צריכים לבחור את השיטה המתאימה ביותר.

<u>שיטת Cholesky factorization (או Cholesky decomposition):</u>

בשיטה זו, אנחנו רוצים לייצג את המטריצה הסימטרית חיובית A בתור מכפלה של מטריצה L במטריצה L^T, כאשר L היא מטריצה משולשית תחתונה, כלומר:

$$A = L \cdot L^T \qquad , \qquad L = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n1} & \cdots & \cdots & l_{nn} \end{pmatrix}$$

לא נסביר את איך השיטה בדיוק עובדת (ניתן למצוא את ההסבר בספר הלימוד) אך נזכיר כי מספר הפעולות של שיטה זו עד שמוצאים מטריצה L בגודל $n \times n$ הוא במימד הבעיה:

$$\mathrm{O}\left(\frac{n^3}{6}\right)$$

נבחין כי מספר פעולות זה הוא מאד נמוך, שכן רק להכפיל שתי מטריצות ריבועיות מסדר $n \times n$ לוקח n^3 פעולות.

<u>דוגמה:</u>

נניח כי לפנינו מערכת המשוואות הבאה: $L\bar{y} = \bar{b}$ (כאשר L היא מטריצה משולשית תחתונה מסדר 3×3 ווקטור הנעלמים הוא $\bar{y}$) אז באופן מפורש, מערכת המשוואות היא:

$$\begin{aligned} &l_{11}y_1 = b_1 \\ &l_{21}y_1 + l_{22}y_2 = b_2 \\ &l_{31}y_1 + l_{32}y_2 + l_{33}y_3 = b_3 \end{aligned}$$

קל לראות שנוכל לחלץ בקלות את הנעלם הראשון y_1 מהמשוואה הראשונה ונציב זאת במשוואה השניה. לאחר כן, ניתן לראות כי לאחר ההצבה, ניתן בקלות לחלץ את הנעלם השני y_2 ולהמשיך באופן דומה עד אשר חילצנו את כל הנעלמים. מספר הפעולות הדרושות לביצוע פתרון זה הוא $\mathrm{O}\left(n^2\right)$ ודרך פתרון זו נקראת forward substitution.

<u>דוגמה:</u>

נניח ואנחנו רוצים לפתור את מערכת המשוואות $A\bar{x} = \bar{b}$ ($\bar{x}$ זה ווקטור הנעלמים וכן A היא מטריצה חיובית) אז נניח כי המרנו את המערכת לפי שיטת Cholesky factorization למערכת הבאה:

$$L \cdot L^T \cdot \bar{x} = \bar{b}$$

נסמן $\bar{y} \triangleq L^T \cdot \bar{x}$ ואז נפתור באמצעות forward substitution את המערכת $L \cdot \bar{y} = \bar{b}$. כעת נשאר לנו לפתור את המשוואה $L^T \cdot \bar{x} = \bar{y}$ בכדי למצוא את ווקטור הנעלמים $\bar{x}$ אבל נבחין כי L^T היא מטריצה משולשית עליונה ולכן נפתור את המערכת באמצעות backward substitution (באופן מאד דומה ל-forward substitution, רק שנתחיל מהמשוואה האחרונה).

בשיטה זו, פתרנו את מערכת המשוואות $A\bar{x} = \bar{b}$ עם סדר גודל של $O\left(n^2\right)$ פעולות.

שיטת Modified Cholesky factorization:

גם אם המטריצה A לא חיובית (שזה תנאי הכרחי לביצוע השיטה) ניתן להשתמש בשיטת Cholesky רק עם שינוי, יש תחילה לחשב את המטריצה האלכסונית Δ אשר ניתן להוסיף למטריצה A בכדי שהמטריצה $A+\Delta$ תהיה חיובית. לא נתאר כאן את אופן החישוב אבל יש פונקציה במטלב שיודעת למצוא את המטריצה Δ המתאימה.

מקרה פרטי של שיטת ניוטון: Least Squares Problem

נתאר מציאת פתרון אופטימלי למערכת משוואות עם יותר משוואות מנעלמים לפי סכום ריבועים מינימלי (Least Squares Problem) לפי השיטה של ניוטון וגאוס (שיטת ניוטון-גאוס). תחילה נתאר את בעית ה-Least Squares Problem ולאחר מכן נציג את הפתרון לפי שיטת ניוטון-גאוס.

נניח כי מערכת המשוואות, מתוארת ע"י פונקציה ווקטורית (ניתן לחשוב על מטריצה כעל פונקציה ווקטורית, כלומר העתקה מרחב ווקטורי אחד למרחב ווקטורי אחר) $\bar{g}\left(\bar{x}\right): \mathbb{R}^n \to \mathbb{R}^m$ ונניח כי אנחנו יודעים כי הפונקציה $\bar{g}\left(\bar{x}\right)$ כמעט מתאפסת בנקודה מסויימת ולכן המערכת שאנחנו רוצים לפתור היא: $\bar{g}\left(\bar{x}\right) \approx \bar{0}$ (כלומר אנחנו מניחים כי לא ניתן למצוא נקודה (ווקטור) $\bar{x}$ עבורה נקבל ממש אפס אך ניתן להתקרב לזה). באופן פורמלי אנחנו מחפשים את הפתרון $\bar{x}$ אשר יביא למינימום את סכום הריבועים של הרכיבים של $\bar{g}\left(\bar{x}\right)$ כלומר:

$$\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right) \equiv \frac{1}{2}\left\|\bar{g}\left(\bar{x}\right)\right\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} g_i^2\left(\bar{x}\right)$$

באופן כללי, אין צורך לבחור דווקא את סכום הריבועים ואפשר גם לפתור את הבעיה:

$$\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right) = \sum_{i=1}^{m} \varphi\left(g_i\left(\bar{x}\right)\right)$$

כאשר $\varphi: \mathbb{R} \to \mathbb{R}$ היא פונקציה כלשהי לבחירתנו. אנחנו ננתח את המקרה הכללי הנ"ל ונרצה למשל למצוא את הגרדיאנט של הפונקציה φ בנקודה מסויימת וגם את ההסיאן. לפונקציות הללו, אנחנו קוראים Penalty Functions שכן הן קובעות מה השגיאה שנוצרת לנו בעקבות פתרון שמצאנו.

אם נסמן:

$$f\left(\bar{x}\right) = \sum_{i=1}^{m} \varphi\left(g_i\left(\bar{x}\right)\right)$$

אז מתקיים (מומלץ להוכיח כשיעורי בית (הוכחנו בכיתה בהרצאה הראשונה את הבעיה הזו בדיוק עם שמות שונים לפונקציות)):

$$\boxed{\begin{aligned}&\nabla f(\bar{x})=\sum_{i=1}^{m}\left(\varphi'\left(g_i(\bar{x})\right)\nabla g_i(\bar{x})\right)\\&H(\bar{x})=\nabla^2 f(\bar{x})=\sum_{i=1}^{m}\left(\varphi''\left(g_i(\bar{x})\right)\nabla g_i(\bar{x})\nabla g_i^T(\bar{x})+\varphi'\left(g_i(\bar{x})\right)\nabla^2 g_i(\bar{x})\right)\end{aligned}}$$

נבחין כי הביטוי $\nabla g_i(\bar{x})\nabla g_i^T(\bar{x})$ הוא מטריצה שכן $\nabla g_i(\bar{x})$ הוא ווקטור עמודה. מטריצה כזו נקראת Rank One Matrix וזה משום שיש לה רק ע"ע אחד שונה מאפס. נרצה לרשום את שתי הנוסחאות לעיל בכתיב מטריצי וקומפקטי.

נסמן תחילה את $\nabla\bar{g}(\bar{x})$ להיות מטריצה שכל אחת מהעמודות שלה היא $\nabla g_i(\bar{x})$. כלומר:

$$\nabla\bar{g}(\bar{x})\triangleq\begin{pmatrix}\nabla g_1(\bar{x}) & \cdots & \nabla g_m(\bar{x})\end{pmatrix}$$

בנוסף, נסמן את הווקטור:

$$\varphi'\left(\bar{g}(\bar{x})\right)=\begin{pmatrix}\varphi'\left(g_1(\bar{x})\right)\\ \vdots\\ \varphi'\left(g_m(\bar{x})\right)\end{pmatrix}$$

נסמן בנוסף את המטריצה האלכסונית:

$$\vartheta''\left(\bar{g}(\bar{x})\right)=\begin{pmatrix}\varphi''\left(g_1(\bar{x})\right) & & 0\\ & \ddots & \\ 0 & & \varphi''\left(g_m(\bar{x})\right)\end{pmatrix}$$

כעת נוכל לרשום באופן מטריצי וקומפקטי את שתי הנוסחאות שרשמנו קודם באופן הבא:

$$\boxed{\begin{aligned}&\nabla f(\bar{x})=\nabla\bar{g}(\bar{x})\varphi'\left(\bar{g}(\bar{x})\right)\\&H(\bar{x})=\nabla^2 f(\bar{x})=\nabla\bar{g}(\bar{x})\vartheta''\left(\bar{g}(\bar{x})\right)\left(\nabla\bar{g}(\bar{x})\right)^T+\sum_{i=1}^{m}\varphi'\left(g_i(\bar{x})\right)\nabla^2 g_i(\bar{x})\end{aligned}}$$

נחזור למקרה הפרטי של בעית Least Squares ולכן הפונקציה Penalty Function היא: $\varphi(t)=\frac{1}{2}t^2$ ולכן הסימונים שלנו מקבלים את הצורה הבאה:

$$\varphi(\bar{t})=\frac{1}{2}\|\bar{t}\|_2^2\quad,\qquad\varphi'(\bar{t})=\bar{t}\qquad,\qquad\varphi''(\bar{t})=1\qquad\Rightarrow\qquad\vartheta''(\bar{t})\equiv I$$

ולכן נקבל בצורה קומפקטית את הבעיה Least Squares:

$$\boxed{\begin{aligned} &f\left(\bar{x}\right)=\sum_{i=1}^{m}\varphi\left(g_i\left(\bar{x}\right)\right)=\frac{1}{2}\left\|\bar{g}\left(\bar{x}\right)\right\|_2^2=\frac{1}{2}\sum_{i=1}^{m}g_i^2\left(\bar{x}\right) \\ &\nabla f\left(\bar{x}\right)=\overbrace{\nabla\bar{g}\left(\bar{x}\right)}^{Matrix}\cdot\overbrace{\bar{g}\left(\bar{x}\right)}^{Column\ Vector} \\ &H\left(\bar{x}\right)=\nabla^2 f\left(\bar{x}\right)=\nabla\bar{g}\left(\bar{x}\right)\left(\nabla\bar{g}\left(\bar{x}\right)\right)^T+\sum_{i=1}^{m}g_i\left(\bar{x}\right)\nabla^2 g_i\left(\bar{x}\right)\end{aligned}}$$

<u>שיטת ניוטון-גאוס (Newton-Gauss Method)</u>:

נניח כי אנחנו מחפשים את הנקודה $\bar{x}$ אשר מביאה למינימום את $f\left(\bar{x}\right)=\frac{1}{2}\left\|\bar{g}\left(\bar{x}\right)\right\|_2^2=\frac{1}{2}\sum_{i=1}^{m}g_i^2\left(\bar{x}\right)$ כלומר:

$$\bar{x}=\arg\min f\left(\bar{x}\right)=\frac{1}{2}\left\|\bar{g}\left(\bar{x}\right)\right\|_2^2=\frac{1}{2}\sum_{i=1}^{m}g_i^2\left(\bar{x}\right)$$

נניח כי האלגוריתם לפתרון שלנו נמצא כבר מאד קרוב לפתרון האופטימלי. נתבונן בביטוי שמצאנו בכתיב מטריצי עבור ההסיאן. נבחין כי אם אנחנו קרובים מאד לפתרון אז כל אחד מהרכיבים בסכום $\frac{1}{2}\sum_{i=1}^{m}g_i^2\left(\bar{x}\right)$ הוא מאד קטן, כלומר הגרדיאנט $\bar{g}\left(\bar{x}\right)$ מאד קטן (קרוב לווקטור האפס) ולכן הביטוי השני בהסיאן $\sum_{i=1}^{m}g_i\left(\bar{x}\right)\nabla^2 g_i\left(\bar{x}\right)$ הוא מאד קטן ולכן הביטוי הדומיננטי בהסיאן הוא הביטוי: $\nabla\bar{g}\left(\bar{x}\right)\left(\nabla\bar{g}\left(\bar{x}\right)\right)^T$. ניתן לרשום:

$$\begin{aligned} &H\left(\bar{x}\right)=\nabla^2 f\left(\bar{x}\right)=\nabla\bar{g}\left(\bar{x}\right)\left(\nabla\bar{g}\left(\bar{x}\right)\right)^T+\cancel{\sum_{i=1}^{m}g_i\left(\bar{x}\right)\nabla^2 g_i\left(\bar{x}\right)}\cong\nabla\bar{g}\left(\bar{x}\right)\left(\nabla\bar{g}\left(\bar{x}\right)\right)^T \\ &\Rightarrow H\left(\bar{x}\right)\cong\nabla\bar{g}\left(\bar{x}\right)\left(\nabla\bar{g}\left(\bar{x}\right)\right)^T\end{aligned}$$

ושיטת ניוטון-גאוס היא למעשה האלגוריתם האיטרטיבי הבא (זו למעשה אלגוריתם ניוטון עם צעד בכיוון קצת שונה):

$$\bar{x}_{k+1}=\bar{x}_k-\alpha_k H^{-1}\left(\bar{x}_k\right)\overbrace{\left(\nabla\bar{g}\left(\bar{x}_k\right)\right)}^{Colmun\ Vector}\overbrace{\left(\bar{g}\left(\bar{x}_k\right)\right)}^{Column\ Vector}$$

אם מטריצת ההסיאן היא סינגולרית (לא בעלת דרגה מלאה), כלומר אם ישנם ערכים עצמיים שהם אפס במטריצת ההסיאן $H\left(\bar{x}\right)\cong\nabla\bar{g}\left(\bar{x}\right)\left(\nabla\bar{g}\left(\bar{x}\right)\right)^T$ (הם אינם יכולים להיות שליליים כי מצאנו כי המטריצה היא אי-שלילית) אז נוכל להוסיף מטריצה אלכסונית Δ ונקבל $\nabla\bar{g}\left(\bar{x}\right)\left(\nabla\bar{g}\left(\bar{x}\right)\right)^T+\Delta$ (ניתן להשתמש בשיטת Modified Cholesky factorization) ולהציב באלגוריתם האיטרטיבי:

$$\bar{x}_{k+1}=\bar{x}_k-\alpha_k\left(H\left(\bar{x}_k\right)+\Delta\right)^{-1}\overbrace{\left(\nabla\bar{g}\left(\bar{x}_k\right)\right)}^{Colmun\ Vector}\overbrace{\left(\bar{g}\left(\bar{x}_k\right)\right)}^{Column\ Vector}$$

באופן אלטרנטיבי ניתן להציב $\Delta=\varepsilon_k I$ ולקבל: $H\left(\bar{x}_k\right)+\varepsilon_k I$ ושיטה זו היא בעלת שם משל עצמה והיא נקראת Levenberg-Marquart Method.

הרצאה 10 – Conjugate Gradient Method, שיטת הווקטורים האורתוגונליים לגרדיאנט

למדנו שאלגוריתם ש"הולך" תמיד נגד כיוון הגרדיאנט, קצב התכנסות שלו (של האלגוריתם) תלוי בערך של condition number וכאשר הוא גדול אז קצב ההתכנסות הוא איטי ויקח הרבה זמן עד שהאלגוריתם יגיע לפתרון. לאחר מכן, מצאנו את שיטת ניוטון ששיפרה את האלגוריתם הקודם אך שיטת ניוטון צריכה לחשב את מטריצת ההסיאן בכל מהלך של האלגוריתם ואם בעית האופטימיזציה שלנו היא פונקציה של הרבה מאד משתנים אז מטריצת ההסיאן תהיה גדולה מאד ונדרש זיכרון עצום בשביל לשמור את המטריצה. כמו כן, בעיה נוספת בשיטת ניוטון זו סיבוכיות החישוב, אם הבעיה היא ב- n מימדים, אז סיבוכיות החישוב היא $O\left(n^3\right)$ ואם נשתמש בשיטת Cholesky נגיע אמנם לסיבוכיות $O\left(\frac{1}{6}n^3\right)$ אבל זו עדיין סיבוכיות די גדולה ולמעשה פתרון של בעיה מסויימת יכול לקחת שעות של חישוב ע"י מחשב (כאשר $n \approx 10,000$) וגם גודל מטריצת ההסיאן כל כך גדול שלא ניתן באופן מעשי לחשב את המטריצה ההופכית של ההסיאן.

אנחנו צריכים שיטה שתמצא פשרה בין פשטות האלגוריתם לבין סיבוכיות האלגוריתם. השיטה שנלמד כעת מנסה למצוא את הפשרה הנ"ל. תחילה נראה איך השיטה עובדת עבור בעיות אופטימיזציה עם פונקציות ריבועיות ולאחר מכן נעבור למקרה הכללי של פונקציות כלשהן.

השיטה Conjugate directions מתבססת בעיקרה על המכפלה הפנימית ועל התכונות של אורתוגנליות אך באופן יותר כללי מאשר מה שלמדנו במרחבים אוקלידיים. בשיטה זו אנחנו למעשה מנסים לקחת בעית אופטימיזציה ב- n משתנים ולהפוך את הבעיה ל- n בעיות אופטימיזציה של משתנה יחיד. נתחיל בהקדמה מתמטית.

<u>הגדרה</u>: מכפלה פנימית

נניח כי $a,b,c \in \mathbb{R}^n$ ווקטורים ונניח כי $\alpha \in \mathbb{C}$ סקלר. נגדיר את המכפלה הסקלרית בין הווקטורים a,b להיות: $\langle a,b \rangle$.

תכונות של כל מכפלה פנימית:

- $\langle a,b \rangle = \overline{\langle b,a \rangle}$
- $\langle a,b+c \rangle = \langle a,b \rangle + \langle a,c \rangle$
- $\langle \alpha a,b \rangle = \alpha \langle a,b \rangle$
- $\langle a,a \rangle \geq 0$
- $\langle a,a \rangle = 0$ אמ"מ $a = \bar{0}$.

הנורמה מוגדרת להיות:

$$norm\ a \triangleq \|a\| = \sqrt{\langle a,a \rangle}$$

<u>דוגמאות למכפלה פנימית</u>:

- במרחב $\mathbb{R}^n$: (המכפלה הפנימית הסטנדרטית)

$$\langle a,b \rangle \triangleq a^T \cdot b = \sum_{i=1}^{n} a_i b_i$$

- במרחב הפונקציות של משתנה יחיד (או רבים) שמוגדרות בתחום $I = [a,b]$: (המכפלה הפנימית הסטנדרטית)

$$\left\langle f\left(x\right),g\left(x\right)\right\rangle \triangleq \int_a^b f\left(x\right)\cdot\overline{g\left(x\right)}dx$$

- במרחב $\mathbb{R}^n$, נניח כי $\bar{x},\bar{y}\in\mathbb{R}^n$ וכן Q מטריצה <u>חיובית</u> מסדר $n\times n$ אז ניתן להגדיר את המכפלה הפנימית הבאה:

$$\left\langle \bar{x},\bar{y}\right\rangle_Q \triangleq \bar{x}^TQ\bar{y}$$

(יש להוכיח כי המכפלה הפנימית שהגדרנו היא אכן מכפלה פנימית באמצעות זה שנוכיח כי היא מקיימת את כל התכונות שראינו קודם שכל מכפלה פנימית חייבת לקיים – מומלץ לבדוק כשיעורי בית)

<u>תהליך גרהם-שמידט:</u>

תהליך גרהם-שמידט הוא תהליך שמקבל קבוצה של וקטורים בלתי תלויים של מרחב בעל מכפלה פנימית ומחזיר קבוצה של וקטורים אורתונורמליים של מרחב זה.

נניח כי קבוצת הווקטורים שהתהליך מקבל היא: $\left\{\bar{x}_1,\bar{x}_2,...,\bar{x}_n\right\}$ אז הבסיס האורתונורמלי שנקבל ע"י התהליך הוא:

$$\begin{aligned}
&\bar{y}_1=\bar{x}_1\\
&\bar{y}_k=\bar{x}_k-\sum_{i=1}^{k-1}\left\langle \bar{x}_i,\frac{\bar{y}_i}{\left\|\bar{y}_i\right\|}\right\rangle\cdot\frac{\bar{y}_i}{\left\|\bar{y}_i\right\|}=\bar{x}_k-\sum_{i=1}^{k-1}\frac{\left\langle \bar{x}_i,\bar{y}_i\right\rangle}{\left\|\bar{y}_i\right\|^2}\cdot\bar{y}_i
\end{aligned}$$

<u>וקטורים אורתוגונלים ביחס למכפלה הפנימית עם המטריצה Q (Q-conjugate (Q-orthogonal) directions):</u>

נניח כי $\bar{d}_i,\bar{d}_j\in\mathbb{R}^n$ וקטורים, אז נאמר כי הם אורתוגונליים לפי המכפלה הפנימית לפי המטריצה Q אם מתקיים:

$$\left\langle \bar{d}_i,\bar{d}_j\right\rangle_Q=\bar{d}_i^TQ\bar{d}_j=0$$

אם נניח כי נתונה לנו קבוצה של וקטורים בלתי תלויים לינארית $\left\{\bar{\xi}_1,\bar{\xi}_2,...,\bar{\xi}_n\right\}$ אז ע"י תהליך גרהם-שמידט אפשר למצוא קבוצת וקטורים אורתוגונליים $\left\{\bar{d}_1,\bar{d}_2,...,\bar{d}_n\right\}$ באופן הבא:

$$\begin{aligned}
&\bar{d}_1=\bar{\xi}_1\\
&\bar{d}_k=\bar{\xi}_k-\sum_{i=1}^{k-1}\left\langle \bar{\xi}_i,\frac{\bar{d}_i}{\left\|\bar{d}_i\right\|}\right\rangle\cdot\frac{\bar{d}_i}{\left\|\bar{d}_i\right\|}=\bar{\xi}_k-\sum_{i=1}^{k-1}\frac{\left\langle \bar{\xi}_i,\bar{d}_i\right\rangle}{\left\|\bar{d}_i\right\|^2}\cdot\bar{d}_i=\bar{\xi}_k-\sum_{i=1}^{k-1}\frac{\bar{\xi}_i^TQ\bar{d}_i}{\bar{d}_i^TQ\bar{d}_i}\cdot\bar{d}_i
\end{aligned}$$

נראה בקרוב כי הווקטורים האורתוגונליים הללו מאפשרים לנו באמצעות שיטה פשוטה למדי למצוא את המינימום של פונקציה ריבועית.

<u>מציאת מינימום של פונקציה ריבועית:</u>

נניח כי Q מטריצה <u>חיובית</u> מסדר $n\times n$ וכן וקטור $\bar{b}\in\mathbb{R}^n$ אז נוכל להגדיר פונקציה ריבועית $f:\mathbb{R}^n\to\mathbb{R}$ (התנאי שהמטריצה תהיה חיובית אומר שבהכרח לפונקציה הריבועית יש מינימום ולא מקסימום, בדיוק כמו פונקציה ריבועית ממימד אחד אשר יהיה לה מקסימום אם הנגזרת השניה תהיה חיובית) באופן הבא:

$$f\left(\bar{x}\right)=\bar{x}^TQ\bar{x}+\bar{b}^T\cdot\bar{x}=\frac{1}{2}\left\|\bar{x}\right\|_Q^2+\bar{b}^T\cdot\bar{x}$$

בנוסף, נניח ובידינו יש קבוצה של n וקטורים אורתוגונליים (בנינו ע"י תהליך גרהם-שמידט) לפי המטריצה Q:

$$\left\{\bar{d}_i\right\}_{i=1}^{n}=\left\{\bar{d}_1,...,\bar{d}_n\right\}$$

נוכל למצוא צירוף לינארי של $\bar{x}$ באמצעות הבסיס שבנינו $\left\{\bar{d}_i\right\}_{i=1}^{n}$ ונקבל:

$$\bar{x}=\sum_{i=1}^{n}\alpha_i\bar{d}_i \qquad , \qquad \forall 1\le i\le n:\ \alpha_i\in\mathbb{R}$$

(כלומר $\bar{\alpha}\in\mathbb{R}^n$ הוא וקטור ואז נוכל לומר כי $\bar{x}$ הוא פונקציה של הווקטור $\bar{\alpha}$)

וכעת נוכל לכתוב את הפונקציה הריבועית באופן הבא:

$$f\left(\bar{x}\left(\bar{\alpha}\right)\right)=\frac{1}{2}\left\|\bar{x}\left(\bar{\alpha}\right)\right\|_Q^2+\bar{b}^T\cdot\bar{x}=\sum_{i=1}^{n}\left(\frac{1}{2}\alpha_i^2\left\|\bar{d}_i\right\|+\alpha_i\cdot\bar{b}^T\bar{d}_i\right)$$

ואם נסמן $\varphi_i\left(\alpha_i\right)\triangleq\frac{1}{2}\alpha_i^2\left\|\bar{d}_i\right\|+\alpha_i\cdot\bar{b}^T\bar{d}_i$ אז נקבל את בעית האופטימיזציה הבאה:

$$\min_{\bar{\alpha}\in\mathbb{R}^n} f\left(\bar{x}\left(\bar{\alpha}\right)\right)=\sum_{i=1}^{n}\min_{\alpha_i\in\mathbb{R}}\varphi_i\left(\alpha_i\right)=\min_{\alpha_1\in\mathbb{R}}\varphi_1\left(\alpha_1\right)+...+\min_{\alpha_n\in\mathbb{R}}\varphi_n\left(\alpha_n\right)$$

כלומר הצלחנו לבטא את הפונקציה הריבועית שלנו בתור פונקציה פריקה (separable) לפי α_i וזה אומר שהפכנו בעית אופטימיזציה אחת ב- n משתנים ל- n בעיות אופטימיזציה של משתנה אחד וזה טוב לנו, כי אנחנו רואים כי מציאת המינימום של בעיה אחת, בהכרח לא משפיע על בעיה אחרת (כי המטרה היא למזער את הסכום כולו בסופו של דבר). בשביל לפתור כל אחת מ- n בעיות האופטימיזציה נוכל לבצע אלגוריתם איטרטיבי של אופטימיזציה במשתנה יחיד (למשל Line search).

תכונת היריעה של מרחבים אפיניים (Expanding Manifold Property):

באמצעות השיטה Conjugate directions method שראינו לעיל, ניתן לפתור בעיות גם במרחבים אפיניים, כלומר הנקודה $\bar{x}$ יכולה להיות מתוארת גם באופן הבא:

$$\bar{x}=\bar{x}_0+\sum_{i=1}^{n}\alpha_i\bar{d}_i \qquad , \qquad \forall 1\le i\le n:\ \alpha_i\in\mathbb{R}$$

ותהליך האופטימיזציה הוא זהה למה שראינו לעיל.

לאחר שבידינו n בעיות האופטימיזציה, אם למשל נסתכל על מצב שבו פתרנו רק $k<n$ בעיות אז אנחנו מקבלים את הפתרון החלקי הבא:

$$\bar{x}=\bar{x}_0+\sum_{i=1}^{k}\alpha_i\bar{d}_i$$

הגדרה:

תת-מרחב אפיני G_k הוא המרחב:

$$G_k = \left\{ \bar{x} \mid \bar{x} = \bar{x}_0 + \sum_{i=1}^{k} \alpha_i \bar{d}_i \right\} \quad , \quad \forall 1 \le i \le k : \ \alpha_i \in \mathbb{R}$$

כלומר G הוא מרחב לינארי מוזז ע"י הווקטור $\sum_{i=1}^{k} \alpha_i \bar{d}_i$.

תכונת היריעה של מרחבים אפיניים אומרת כי לאחר הצעד ה- k בשיטה Conjugate directions method (כלומר פתרנו k בעיות אופטימיזציה מתוך ה- n) אנחנו נקבל את הפתרון החלקי <u>האופטימלי</u> הבא:

$$\bar{x}_k = \arg\min_{\bar{x} \in G_k} f\left(\bar{x}\right)$$

כלומר בנקודה $\bar{x}_k$ נקבל את הערך האופטימלי האפשרי לאחר k איטרציות באלגוריתם.

הרצאה 11 – Conjugate Gradient Method 2, שיטת הווקטורים האורתוגונליים לגרדיאנט - המשך

ראינו בהרצאה הקודמת כי ניתן בצורה פשוטה למצוא את המינימום של הפונקציה הריבועית $f\left(\bar{x}\right)=\bar{x}^{T}Q\bar{x}+\bar{b}^{T}\bar{x}$ באמצעות זה שנמצא את הווקטורים האורתוגונליים לפי המטריצה החיובית Q .

נסמן את הגרדיאנט של הפונקציה הריבועית בנקודה $\bar{x}_k$ להיות:

$$\bar{g}_k \triangleq \bar{g}\left(\bar{x}_k\right)=\nabla f\left(\bar{x}_k\right)=Q\bar{x}_k+\bar{b}$$

אם בהרצאה הקודמת הנחנו שבידינו קבוצת ווקטורים אורתוגונליים $\left\{\bar{d}_i\right\}_{i=0}^{n}$ שיצרנו באמצעות תהליך גרהם-שמידט, כעת נרצה לבטא את הווקטורים האורתוגונליים $\left\{\bar{d}_i\right\}_{i=0}^{n}$ באמצעות הגרדיאנטים $\bar{g}_k$ ולראות איך השיטה Conjugate Gradient נראית באופן הזה. נבחר את הווקטור הראשון:

$$\bar{d}_0=-\bar{g}_0=-\left(Q\bar{x}_0+\bar{b}\right)$$

הצעד ה-k-י באלגוריתם הוא:

$$\bar{x}_{k+1}=\bar{x}_k+\gamma_k\bar{d}_k \qquad , \qquad \gamma_k\in\mathbb{R}$$

ואת הערך γ_k נמצא בשיטת Exact line Search (בשיטת Conjugate direction חייבים למצוא את γ_k בשיטה זו בלבד כיוון שזו השיטה הכי מדוייקת ובמקרה זה השיטה Exact Line Search הכרחית). ואז נמצא את הווקטור האורתוגונלי הבא $\bar{d}_{k+1}$ באמצעות תהליך גרהם-שמידט:

$$\bar{d}_{k+1}=-\bar{g}_{k+1}+\sum_{j=0}^{k}\frac{\bar{g}_{k+1}^{T}Q\bar{d}_j}{\bar{d}_j^{T}Q\bar{d}_j}\bar{d}_j$$

אחת היתרונות העיקריים של השיטה Conjugate Gradient על פני השיטה הכללית יותר שהיא Conjugate Direction נעוץ ברישום של הווקטורים האורתוגונליים $\bar{d}_k$, כאשר רושמים את הווקטורים באמצעות הגרדיאנט זה נהיה הרבה יותר פשוט וקומפקטי.

נבחין כי הצעד באלגוריתם הוא $\bar{x}_{k+1}=\bar{x}_k+\gamma_k\bar{d}_k$ ולכן מתקיים:

$$\bar{d}_j=\frac{1}{\gamma_j}\left(\bar{x}_{j+1}-\bar{x}_j\right)$$

ומכאן נקבל:

$$Q\bar{d}_j=\frac{1}{\gamma_j}Q\left(\bar{x}_{j+1}-\bar{x}_j\right)\underset{\bar{g}_k=\left(Q\bar{x}_k+\bar{b}\right)}{=}\frac{1}{\gamma_j}\left(\bar{g}_{j+1}-\bar{b}-\left(\bar{g}_j-\bar{b}\right)\right)=\frac{1}{\gamma_j}\left(\bar{g}_{j+1}-\bar{g}_j\right)$$

בנוסף ניזכר כי לפי תכונת היריעה של מרחבים אפיניים אנחנו יודעים כי בצעד ה-k-י באלגוריתם, אנחנו מגיעים לפתרון האופטימלי האפשרי $\bar{x}_k$ בתת המרחב האפיני G_k (המרחב שנפרש ע"י כל הווקטורים שמצאנו עד לאותו שלב באלגוריתם) וכיוון שבשיטה Conjugate Gradient אנחנו גורמים לזה שכל וקטור שאנחנו יוצרים הוא צירוף של גרדיאנטים קודמים אז זה אומר שבכל שלב באלגוריתם אנחנו מוציאים את הפתרון האופטימלי במרחב נפרש ע"י הגרדיאנטים הקודמים שמצאנו, וכמו שראינו בהרצאה הקודמת, כל ווקטור חדש שאנחנו מוצאים הוא אורתוגונלי לכל הקודמים ולכן גם הגרדיאנט החדש שנמצא בצעד הבא הוא אורתוגונלי לכל הגרדיאנטים האחרים ולמרחב כולו שנפרש על ידם, כלומר באופן פורמלי נקבל:

$$\bar{g}_{k+1} \perp \{\bar{g}_k, \bar{g}_{k-1}, ..., \bar{g}_0\}$$

כעת מהחישוב שעשינו עבור $Q\bar{d}_j$ ומהעובדה שמצאנו כי מתקיים $\bar{g}_{k+1} \perp \{\bar{g}_k, \bar{g}_{k-1}, ..., \bar{g}_0\}$ אנחנו מסיקים כי מתוך כל הסכום בביטוי:

$$\bar{d}_{k+1} = -\bar{g}_{k+1} + \sum_{j=0}^{k} \frac{\bar{g}_{k+1}^T Q \bar{d}_j}{\bar{d}_j^T Q \bar{d}_j} \bar{d}_j$$

נשאר רק האיבר האחרון (של הסכום) כי הגרדיאנט $\bar{g}_{k+1}^T$ אורתוגונלי לכל הגרדיאנטים הקודמים (לאחר הצבת החישוב שמצאנו עבור $Q\bar{d}_j$) ולכן המונה מתאפס כי הוא מכפלה סקלרית בין ווקטורים אורתוגנליים ולכן נקבל:

$$\bar{d}_{k+1} = -\bar{g}_{k+1} + \frac{\bar{g}_{k+1}^T Q \bar{d}_k}{\bar{d}_k^T Q \bar{d}_k} \bar{d}_k \underset{Q\bar{d}_k = \frac{1}{\gamma_k}(\bar{g}_{k+1} - \bar{g}_k)}{=} -\bar{g}_{k+1} + \frac{\bar{g}_{k+1}^T (\bar{g}_{k+1} - \bar{g}_k)}{\bar{d}_k^T (\bar{g}_{k+1} - \bar{g}_k)} \bar{d}_k$$

ואם נסמן את הסקלר שמוכפל בווקטור $\bar{d}_k$ להיות β_k נקבל את הביטוי הפשוט הבא:

$$\bar{d}_{k+1} = -\bar{g}_{k+1} + \beta_k \bar{d}_k$$

ננסה לפשט את הביטוי עבור β_k . במכנה נבחין כי $\bar{g}_{k+1} \perp \bar{d}_k^T$ ולכן המכפלה הסקלרית ביניהם מתאפסת ולכן המכנה נותר רק $\bar{d}_k^T \cdot (-\bar{g}_k)$ ואם נציב במפורש את $\bar{d}_k$ מתוך האיטרציה הקודמת באלגוריתם נקבל:

$$\bar{d}_k = -\bar{g}_k + \underbrace{\frac{\bar{g}_k^T (\bar{g}_j - \bar{g}_{j-1})}{\bar{d}_{k-1}^T (\bar{g}_j - \bar{g}_{j-1})}}_{\beta_{k-1}} \bar{d}_{k-1} = -\bar{g}_k + \beta_{k-1} \bar{d}_{k-1}$$

(נציב את הביטוי שקיבלנו זה עתה עבור $\bar{d}_k$ במכנה) ולכן המכנה הוא:

$$\left(-\bar{g}_k + \beta_{k-1} \bar{d}_{k-1}\right)^T (-\bar{g}_k)$$

וכיוון שמתקיים $\bar{d}_{k-1} \perp \bar{g}_k$ (כי מתקיים $\bar{d}_{k-1}^T \cdot \bar{g}_k = 0$) נשארנו במכנה עם הסקלר:

$$\left(-\bar{g}_k{}^T\right)(-\bar{g}_k) = \|\bar{g}_k\|_2^2$$

כלומר הצלחנו לפשט את המכנה של β_k וקיבלנו:

$$\beta_k = \frac{\bar{g}_{k+1}^T\left(\bar{g}_{k+1} - \bar{g}_k\right)}{\left\|\bar{g}_k\right\|_2^2}$$

הנוסחה האחרונה פותחה ע"י Polak-Ribiere ונקראת על שמם, וכן אם נבחין כי מתקיים גם $\bar{g}_k \perp \bar{g}_{k+1}$ אז נוכל לרשום את הנוסחה האחרונה גם באופן הבא:

$$\beta_k = \frac{\left\|\bar{g}_{k+1}\right\|_2^2}{\left\|\bar{g}_k\right\|_2^2}$$

ונוסחה זו קרויה על שם Fletcher-Reevs.

שתי הנוסחאות האלה במקרה שלנו מתלכדות אך כאשר נשתמש בשיטה Conjugate Gradient עבור פונקציה כללית (ולאו דווקא ריבועית) אנחנו נראה כי העובדה $\bar{g}_k \perp \bar{g}_{k+1}$ לא מתקיימת ולכן שתי הנוסחאות האחרונות שונות.

כמו כן, עבור חישובים נומריים, דווקא הנוסחה הראשונה מתנהגת טוב יותר ולכן עבור חישובים נומריים לרוב משתמשים בנוסחה של Polak-Ribiere.

<u>סיכום השיטה Conjugate Gradient</u>:

עבור הפונקציה הריבועית $f\left(\bar{x}\right) = \frac{1}{2}\bar{x}^T Q\bar{x} + \bar{b}^T\bar{x}$ השיטה היא כלהלן:

1) מתחילים לבנות את קבוצת הווקטורים האורתוגונליים לפי הווקטור הראשון:
$\bar{d}_0 = -\bar{g}_0 = \nabla f\left(\bar{x}_0\right) = Q\bar{x}_0 + \bar{b}$

2) בכל שלב באלגוריתם מחשבים את הנקודה הבאה לפי הנוסחה: $\bar{x}_{k+1} = \bar{x}_k + \gamma_k \bar{d}_k$ ואת הסקלר γ_k מוצאים בצורה אנליטית (כיוון שזו פונקציה ריבועית אז ניתן למצוא באופן אנליטי) לפי הנוסחה:

$$\gamma_k = -\frac{\bar{g}_k^T\bar{d}_k}{\bar{d}_k^T Q\bar{d}_k} = -\frac{f'_{\bar{d}_k}\left(\bar{x}_k\right)}{f''_{\bar{d}_k\bar{d}_k}\left(\bar{x}_k\right)}$$

(זו נוסחה דומה לנוסחה של מציאת מינימום של פונקציה ריבועית במשתנה יחיד – הוכיחו זאת לעצמכם) ניתן גם להגיע לפתרון זה באמצעות איטרציה אחת (כיוון שמדובר בפונקציה ריבועית) בשיטת ניוטון.

3) את הגרדיאנט הבא מוצאים לפי $\bar{g}_{k+1} = Q\bar{x}_{k+1} + \bar{b}$ ונבחין כי ניתן לחשב $\bar{g}_{k+1}$ ללא חישוב נוסף כיוון שמתקיים $Q\bar{x}_{k+1} = Q\bar{x}_k + \gamma_k Q\bar{d}_k$ ונבחין כי את הביטוי $Q\bar{d}_k$ כבר חישבנו כאשר חישבנו את γ_k.

4) מוצאים את וקטור הכיוון הבא, הווקטור האורתוגנולי הבא לפי הנוסחה $\bar{d}_{k+1} = -\bar{g}_{k+1} + \beta_k\bar{d}_k$ כאשר מתקיים (לא רק עבור פונקציה ריבועית):

$$\beta_k = \frac{\bar{g}_{k+1}^T\left(\bar{g}_{k+1} - \bar{g}_k\right)}{\left\|\bar{g}_k\right\|_2^2} \quad , \qquad Polak - Ribiere$$

$$\beta_k = \frac{\left\|\bar{g}_{k+1}\right\|_2^2}{\left\|\bar{g}_k\right\|_2^2} \quad , \qquad Fletcher - Reevs$$

עבור הפונקציה הכללית $f\left(\bar{x}\right)$ השיטה היא כלהלן:

1) מתחילים לבנות את קבוצת הווקטורים האורתוגונליים לפי הווקטור הראשון:
$$\bar{d}_0 = -\bar{g}_0 = \nabla f\left(\bar{x}_0\right)$$
2) בכל שלב באלגוריתם מחשבים את הנקודה הבאה לפי הנוסחה: $\bar{x}_{k+1} = \bar{x}_k + \gamma_k \bar{d}_k$ ואת הסקלר γ_k מוצאים ע"י שיטת Line Search Exact (הסברנו כי בשיטת Conjugate Gradient חייבים להשתמש בשיטה זו).

3) מוצאים את הגרדיאנט הבא: $\bar{g}_{k+1} = \nabla f\left(\bar{x}_{x+1}\right)$.

4) מוצאים את וקטור הכיוון הבא, הווקטור האורתוגנולי הבא לפי הנוסחה $\bar{d}_{k+1} = -\bar{g}_{k+1} + \beta_k \bar{d}_k$ כאשר מתקיים (לא רק עבור פונקציה ריבועית):
$$\beta_k = \frac{\bar{g}_{k+1}^T\left(\bar{g}_{k+1} - \bar{g}_k\right)}{\left\|\bar{g}_k\right\|_2^2} \quad , \qquad Polak - Ribiere$$
$$\beta_k = \frac{\left\|\bar{g}_{k+1}\right\|_2^2}{\left\|\bar{g}_k\right\|_2^2} \quad , \qquad Fletcher - Reevs$$

<u>קצב ההתכנסות של השיטה Conjugate Gradient</u>:

כמו שהראנו כי בשיטת Conjugate Direction, לאחר בדיוק n איטרציות של האלגוריתם, כאשר n זה מספר המשתנים של בעית האופטימיזציה, נגיע לפתרון האופטימלי אז גם בשיטת Conjugate Gradient עבור פונקציה ריבועית האלגוריתם יתכנס לפתרון האופטימלי לאחר n איטרציות כאשר n זה מספר המשתנים של בעית האופטימיזציה. בבעיות מסויימות, אשר בהן יש מאות אלפי משתנים, לפעמים עוצרים את האלגוריתם עוד לפני שהוא ביצע את כל האיטרציות הדרושות.

נסמן את הערך העצמי המקסימלי של מטריצת ההסיאן להיות $M = \lambda_{\max}$ ואת הערך העצמי המינימלי להיות $m = \lambda_{\min}$.

נסמן את $\bar{x}^*$ להיות הערך עבורו הפונקציה שעבורה אנחנו מחפשים את המינימום מקבלת אכן את המינימום אזי ניתן לתאר את קצב ההתכנסות של האלגוריתם Conjugate Gradient ע"י קצב ההתקדמות הבא:
$$\left\|\bar{x}_{k+1} - \bar{x}^*\right\| = c\left\|\bar{x}_k - \bar{x}^*\right\| = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}\left\|\bar{x}_k - \bar{x}^*\right\|$$

בנוסף, כאשר מתקיים $M \gg m$ אפשר להניח כי מתקיים: $c \underset{M \gg m}{=} 1 - 2\sqrt{\frac{m}{M}}$ (נוכל גם להיזכר כי היחס $\theta = \frac{\lambda_{\max}}{\lambda_{\min}}$

נקרא גם Condition Number ובסימונים כעת מתקיים: $c \underset{\lambda_{\max} \gg \lambda_{\min}}{=} 1 - 2\sqrt{\frac{1}{\theta}}$) ואם נשווה את קצב ההתכנסות של האלגוריתם Conjugate Gradient לקצב ההתכנסות של האלגוריתם Steepest Descent (או בשמו האחר Gradient Descent) אז מתקיים קצב ההתכנסות של Steepest Descent הוא:
$$f\left(\bar{x}_{k+1}\right) - f\left(\bar{x}^*\right) \le c\left(f\left(\bar{x}_k\right) - f\left(\bar{x}^*\right)\right)$$

ומתקיים:
$$c = 1 - 2\frac{m}{M} \quad , \qquad M = \lambda_{\max}, \qquad m = \lambda_{\min}$$

וכמובן נזכור כי ככל שהקצב c קרוב יותר לאפס אז קצב ההתכנסות מהיר יותר.

דוגמה:

אם נניח כי $\theta = \frac{M}{m} = 10^6$ (כאשר θ זה ה-condition number) אז נקבל:

$$c_{Conjugate\ Gradient} = 1 - 2\sqrt{\frac{m}{M}} = 0.998$$

$$c_{Gradient\ Descent} = 1 - 2\frac{m}{M} = 0.999998$$

כלומר צריך לבצע בערך 1000 צעדים של שיטת Gradient Descent על מנת לקבל את התוצאות של שיטת Conjugate Gradient – כלומר שיטת Conjugate Gradient בדוגמה זו יעילה פי 1000 מהשיטה Gradient Descent!

<u>שיטה להקטנת הפרמטר Condition number (Preconditioning):</u>

ראינו כי גם בשיטת Conjugate Gradient וגם בשיטת Gradient Descent (Steepest Descent) קצב ההתכנסות מושפע מהפרמטר $\theta = \frac{\lambda_{\max}}{\lambda_{\min}}$ (condition number) וראינו כי כאשר הפרמטר θ הוא קטן, אז מקבלים קצב התכנסות מהיר ולכן נרצה למצוא דרכים כיצד להקטין את הפרמטר θ.

<u>מקרה פרטי של שיטת Preconditioning:</u>

נניח ואנחנו מחפשים את המינימום של פונקציה ריבועית:

$$f(\bar{x}) = \frac{1}{2}\bar{x}^T Q \bar{x} + \bar{b}^T x$$

כאשר מניחים כי מטריצה Q היא חיובית $Q \succ 0$ ולכן גם סימטרית.

נחליף משתנים בפונקציה באופן הבא: נניח כי ברשותנו מטריצה הפיכה S אז נסמן:

$$\bar{x} = S\bar{y}$$

ולכן קיבלנו את החלפת המשתנים:

$$\varphi(\bar{y}) = f(S\bar{y}) = \frac{1}{2}\bar{y}^T S^T Q S\bar{y} + b^T S\bar{y}$$

כלומר במקום המטריצה Q יש לנו את המטריצה S^TQS. החלפת משתנים זו יעילה לנו כאשר אנחנו יודעים כי ה-condition number של המטריצה S^TQS הוא קטן מה-condition number של המטריצה Q ועכשיו אפשר להשתמש בשיטה Conjugate Gradient בצורה יעילה ומהירה יותר כי מובטח לנו קצב התכנסות מהיר יותר מאשר אם היינו מפעילים את השיטה על הפונקציה המקורית עם המטריצה Q.

באופן מעשי, פעמים רבות אנחנו נתקלים במטריצה Q ישנם ערכים גדולים מאד וקטנים מאד על האלכסון הראשי (מה שמצביע גם על ערכים עצמיים גדולים מאד וקטנים מאד) ולכן לאחר שנכפול במטריצה S מתאימה נוכל להקטין את ההפרש בין הערכים על האלכסון ואז הערכים העצמיים יהיו קרובים יותר אחד לשני וכתוצאה מזה גם ה-condition number יקטן ואז קצב ההתכנסות יהיה מהיר יותר.

<u>דוגמה:</u>

נניח כי המטריצה Q היא כמעט מטריצה אלכסונית (כלומר רוב הערכים נמצאים על האלכסון וערכים מעטים נמצאים במקומות אחרים) ונסמן כי המטריצה $Q' = diagonal\ Q$ היא המטריצה Q כאשר מאפסים את כל הערכים שהם לא על האלכסון הראשי אז המטריצה S הדרושה היא:

$$S = \left(\sqrt{Q'}\right)^{-1} = \begin{pmatrix} \frac{1}{\sqrt{q_{11}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{q_{nn}}} \end{pmatrix} \quad \Rightarrow \quad S = S^T$$

ולכן נקבל:

$$S^T Q' S = I \quad \Rightarrow \quad \lambda_{\max} = \lambda_{\min} = 1 \Rightarrow \quad \theta = 1$$

<u>הערה</u>: במקרה הכללי ביותר, אם מניחים כי Q היא סימטרית אז אם נוכל לחשב את $S = \left(\sqrt{Q}\right)^{-1}$ ואז שוב נקבל $\theta = 1$.
שיטה זו, שיטה של Preconditioning יעילה מאד אך ניתן להשתמש בה רק במקרים יחסית ספציפיים.

<u>הכללה של שיטת Preconditioning:</u>

ביטאנו את השיטה (Precondioning) באמצעות החלפת משתנים, וכעת נבטא את השיטה בצורתה המקורית: נסמן $W = S^T S$.
בשיטת Conjugate Gradient הכיוון הראשון אליו מתקדמים באלגוריתם עבור פונקציה ריבועית נקבע ע"י:

$$\bar{d}_0 = W \cdot \bar{g}_0$$

כאשר הגרדיאנט של פונקציה ריבועית הוא כפי שראינו כבר בעבר:

$$f\left(\bar{x}\right) = \frac{1}{2}\bar{x}^T Q \bar{x} + \bar{b}^T x \Rightarrow \quad \bar{g}\left(\bar{x}\right) = Q\bar{x} + \bar{b}$$

הצעד הבא הוא:

$$\bar{x}_{k+1} = \bar{x}_k + \gamma_k \bar{d}_k$$

הכיוון הבא הוא:

$$\bar{d}_{k+1} = W \cdot \bar{g}_{k+1} + \beta_k \bar{d}_k$$

כאשר:

$$\beta_k = \frac{\bar{g}_{k+1}^T \cdot W \cdot \bar{g}_{k+1}}{\bar{g}_k^T \cdot W \cdot \bar{g}_k}$$

בשיטה זו אנחנו לא משתמשים בצורה מפורשת במטריצה S אלא במטריצה $W = S^T S$ שלעיתים חוסך חישובים מיותרים.

<u>שיטת הקטימה של ניוטון (Truncated Newton's Method):</u>

בשיטת ניוטון למדנו כיצד למצוא את המינימום של פונקציות ריבועיות (או פונקציות שאנחנו רוצים לקרב באמצעות פונקציה ריבועית) במספר רב של משתנים בצורה יעילה. ניזכר כי את כיוון הצעד הבא מצאנו לפי מערכת ניוטון (משוואת ניוטון):

$$\begin{cases} H\left(\bar{x}_k\right)\bar{d}_k = -\bar{g}\left(\bar{x}_k\right) \\ \min\limits_{\bar{d}_k} \nabla q\left(\bar{d}_k\right) = \min\limits_{\bar{d}_k} \bar{g}^T\left(\bar{x}_k\right)\bar{d}_k + \frac{1}{2}\bar{d}_k^{\,T} H\left(\bar{x}_k\right)\bar{d}_k \end{cases}$$

(ניזכר כי $q\left(\bar{d}_k\right)$ הוא פיתוח טיילור מסדר שני (ולכן היא פונקציה ריבועית) של הפונקציה המקורית $f\left(\bar{x}\right)$ שאנחנו מנסים למזער)

וראינו כי בעיית מציאת המינימום של פונקציה ריבועית שקולה לפתירת מערכת משוואות כפי שנראה שוב מיד.

<u>טענה</u>:

המשוואה $H\left(\bar{x}_k\right)\bar{d}_k = -\bar{g}\left(\bar{x}_k\right)$ (שנותנת לנו פתרון עבור $\bar{d}_k$, כי $\bar{d}_k = -\frac{\bar{g}\left(\bar{x}_k\right)}{H\left(\bar{x}_k\right)}$) היא תנאי <u>מספיק</u> למינימום של הביטוי $\nabla q\left(\bar{d}_k\right)$. כלומר אם נציב $\bar{d}_k = -\frac{\bar{g}\left(\bar{x}_k\right)}{H\left(\bar{x}_k\right)}$ נקבל את המינימום של הביטוי $\nabla q\left(\bar{d}_k\right)$.

<u>הוכחה</u>:

אם נחשב את הגרדיאנט של הביטוי $\nabla q\left(\bar{d}_k\right) = \frac{1}{2}\bar{d}_k^{\,T} H\left(\bar{x}_k\right)\bar{d}_k + \bar{g}^T\left(\bar{x}_k\right)\bar{d}_k$ ביחס לווקטור $\bar{d}_k$ נקבל כי הגרידאנט מתאפס עבור $\bar{d}_k = -\frac{\bar{g}\left(\bar{x}_k\right)}{H\left(\bar{x}_k\right)}$ ולכן הוכחנו את הנדרש. מש"ל.

נסביר את הצורך בשיטת הקטימה. לעיתים אנחנו נדרשים לפתור בעית אופטימיזציה ובמהלכה אנחנו נדרשים למצוא את כיוון ההתקדמות, כלומר לפתור את משוואת ניוטון: $H\left(\bar{x}_k\right)\bar{d}_k = -\bar{g}\left(\bar{x}_k\right)$, אך פתירת משוואה זו היא יקרה (מבחינת זמן) ולכן לעיתים נעדיף לפתור את הבעיה השקולה:

$$\min_{\bar{d}_k} \frac{1}{2}\bar{d}_k^{\,T} H\left(\bar{x}_k\right)\bar{d}_k + \bar{g}^T\left(\bar{x}_k\right)\bar{d}_k$$

ואת הבעיה הזו נפתור באמצעות שיטת Conjugate Gradient עד השלב שבו נקבל <u>קירוב מספיק טוב</u> ולא באופן מלא כדי לחסוך זמן בחישובים ולכן שיטה זו נקראת Truncated Newton’s Method.

אנחנו למעשה רוצים למצוא $\bar{d}_k$ אשר יקיים: $H\left(\bar{x}_k\right)\bar{d}_k = -\bar{g}\left(\bar{x}_k\right)$ או במילים אחרות יקיים $H\left(\bar{x}_k\right)\bar{d}_k + \bar{g}\left(\bar{x}_k\right) = 0$.

כפי שאמרנו, אנחנו לא נמצא את $\bar{d}_k$ שאכן מביא למינימום את הביטוי $\frac{1}{2}\bar{d}_k^{\,T} H\left(\bar{x}_k\right)\bar{d}_k + \bar{g}^T\left(\bar{x}_k\right)\bar{d}_k$ (ולכן גם פותר את משוואת ניוטון) אלא נמצא בקירוב, אך מה זה קירוב טוב? אנחנו נעצור כאשר יתקיים:

$$\left\| H\left(\bar{x}_k\right)\bar{d}_k + \bar{g}\left(\bar{x}_k\right) \right\| \le \sigma \left\| H\left(\bar{x}_0\right)\bar{d}_0 + \bar{g}\left(\bar{x}_0\right) \right\|$$

כאשר σ הוא <u>קבוע שנקבע לעצמנו</u>. באופן, הזה, אנחנו מצאנו את כיוון הצעד המקורב שיש לעשות באלגוריתם על מנת לפתור את בעית האופטימיזציה הכללית יותר שאנחנו מנסים לפתור.

<u>הערה (זמן החישוב של אלגוריתמי אופטימיזציה)</u>:

בזמן פתירת בעיית אופטימיזציה למעשה אנחנו נדרשים לבצע הרבה פעולות חישוב. למשל, בשיטת ניטון אנחנו נדרשים הרבה פעמים לכפול את מטריצת ההסיאן בווקטור הכיוון ומתברר כי אם אנחנו נמצאים בשלב כלשהו באלגוריתם שפותר בעית אופטימיזציה עם סדר גודל גדול מאד של משתנים (למשל מאות אלפי משתנים), אז כל החישובים הבאים לוקחים אותו סדר גודל של מאמץ חישובי:

- חישוב של מטריצת ההסיאן כפול ווקטור הכיוון
- חישוב ערך הגרדיאנט של הפונקציה המקורית
- חישוב ערך הפונקציה עצמה בנקודה שבה אנחנו עומדים

כל החישובים לעיל לוקחים את אותה כמות הזמן לערך (עד כדי קבוע כפלי קטן יחסית).

בנוסף, בשביל ליעל את החישוב (למשל במטלב) אם למשל $H = A^T BA$ (ההסיאן הוא תוצאה של מכפלה של מטריצות) ואנחנו רוצים לחשב $H\overline{z}$ רצוי לרשום במטלב את השורה הבאה:

$$Answer = A^T\left(B\left(A\overline{z}\right)\right)$$

במקום:

$$Answer = A^T BA\overline{z}$$

כי באפשרות השניה מטלב יחשב תחילה כפל מטריצות ולא יזהה שהוא יכול לבצע רק כפל ווקטור במטריצה במקום.

בנוסף, נבחין כי למדנו כי מתקיים בצורה הדיפרנציאלית $H\left(\overline{x}\right)d\overline{x} = d\overline{g}\left(\overline{x}\right)$, אז אם נסמן $\varepsilon\overline{z}$ אז נוכל לומר כי מתקיים:

$$H\left(\overline{x}\right)\overline{z} \approx \frac{\overline{g}\left(\overline{x}+\varepsilon\overline{z}\right)-\overline{g}\left(\overline{x}\right)}{\varepsilon}$$

כלומר מצאנו שיטה לחשב באופן מקורב את התוצאה של מכפלת הסיאן בווקטור כלשהו ע"י הגרידאנט ולכן זה יקח פחות כוח חישוב.

הרצאה 12 – Sequential Subspace Optimization (SESOP) and Quasi-Newton's Method

גם שיטת SESOP (שהיא היא הרחבה של שיטת Conjugate Gradient) וגם השיטה Quasi-Newton's Method שתיהן שיטות שמיועדות לבעיות אופטימיזציה שהפונקציה בהן היא פונקציה של מספר רב של משתנים.

<u>שיטת Sequential Subspace Optimization (SESOP)</u>:

שיטה זו, היא כאמור הרחבה של שיטת Conjugate Gradient שהיא שיטה איטרטיבית למציאת הנקודה $\bar{x}$ עבור הפונקציה שאנחנו מנסים למצוא את המינימום שלה, $f\left(\bar{x}\right)$, מקבלת את המינימום הזה. כעת, נניח כי $f\left(\bar{x}\right)$ היא חלקה (לפחות גזירה ברציפות) ובעית האופטימיזציה היא כרגיל:

$$\bar{x} = \underset{x\in\mathbb{R}^n}{\arg\min}\ f\left(\bar{x}\right) \qquad , \qquad \bar{x}\in\mathbb{R}^n$$

כאשר הגרדיאנט של הפונקציה $f\left(\bar{x}\right)$ בזמן האיטרציה ה- k הוא $\bar{g}_k\left(\bar{x}_k\right) = \nabla f\left(\bar{x}_k\right)$.

בנוסף נזכיר כי ה**<u>כיוון</u>** (שהלכנו בו על מנת להגיע לנקודה $\bar{x}_k$) (ללא התחשבות בגודל הצעד) הוא:

$$\bar{d}_{k-1} = \bar{x}_k - \bar{x}_{k-1}$$

(למה זה הביטוי עבור $\bar{d}_{k-1}$? זה הביטוי לכיוון שהכיוון הקודם ניתן לביטוי גם כחיבור, או במקרה שלנו כחיסור של שני הווקטורים של הצעדים שביצענו באלגוריתם)

בשיטה SESOP אנחנו מצמצמים את מרחב החיפוש של הנקודה הבאה $\bar{x}_{k+1}$ באמצעות שני הווקטורים: $\bar{d}_{k-1}$ ו- $\bar{g}_k\left(\bar{x}_k\right)$, כלומר אנחנו מחפשים את הנקודה $\bar{x}_{k+1}$ בתת המרחב האפיני שנפרש ע"י הגרדיאנט בנקודה הנוכחית שבה אנחנו נמצאים וע"י ווקטור הכיוון שעל ידו הגענו לנקודה הנוכחית.

תחילה נבנה את המטריצה (באיטרציה ה- k של האלגוריתם Conjugate Gradient) שעמודותיה הן:

$$P_k = \begin{pmatrix} \bar{g}_k\left(\bar{x}_k\right) & \bar{d}_{k-1} & \bar{d}_{k-2} & \cdots \end{pmatrix}$$

(כלומר עמודה ראשונה זה ווקטור הגרדיאנט $\bar{g}_k\left(\bar{x}_k\right)$, עמודה שניה זה ווקטור הכיוון $\bar{d}_{k-1}$ וניתן (אך לא חובה) להשתמש גם בווקטורי הכיוון או בגרדיאנטים שמצאנו עד השלב ה- k שאנחנו נמצאים בו עכשיו, המטריצה P_k חייבת להכיל לכל הפחות שתי עמודות).

<u>האלגוריתם SESOP באיטרציה ה- k</u>:

1) נמצא את כיוון וגודל ההתקדמות $\bar{\alpha}_k$ מהנקודה $\bar{x}_k$, ע"י לפתור את תת-בעית האופטימיזציה:

$$\alpha_k = \underset{\bar{\alpha}}{\arg\min}\ f\left(\bar{x}_k + P\cdot\bar{\alpha}\right)$$

2) והצעד הבא יהיה:

$$\bar{x}_{k+1} = \bar{x}_k + P\cdot\alpha_k$$

הערה:

ניתן לחשוב כי כמות הפעולות שמתבצעת באיטרציה יחידה של אלגוריתם SESOP היא משמעותית גדולה מזו של Conjugate Gradient אך לרוב מבצעים את האלגוריתם הזה על פונקציות מסויימות (שנראה מיד) ובפונקציות אלו (שהן די נפוצות) אלגוריתם זה (שמשתמש בעיקרון ה-Subspace) בכל זאת מאד יעיל.

שימוש בשיטת Fast Subspace Optimization:

ביישומים רבים הפונקציה שעבורה אנחנו מחפשים את המינימום, $f\left(\bar{x}\right)$, היא מהצורה:

$$f\left(\bar{x}\right)=\varphi\left(A\bar{x}\right)$$

כאשר ניתן לחשב את הפונקציה φ יחסית מהר ("זולה לחישוב") והמטריצה A זו מטריצה שמתפקדת למעשה כהעתקה לינארית והכפל שלה בווקטור זו הפעולה האיטית ("יקרה לחישוב"). נראה כי בשיטת Subspace Optimization אין צורך בחישובים מיותרים הכוללים את המטריצה A ולכן זו שיטה מהירה במקרה זה.

בשיטת Subsapce אנחנו למעשה מחפשים את גדול וכיוון הצעד הבא , α_k, באופן הבא:

$$\alpha_k=\arg\min_{\bar{\alpha}} f\left(\bar{x}_k+P\cdot\bar{\alpha}\right)=\arg\min_{\bar{\alpha}}\varphi\left(A\left(\bar{x}_k+P\cdot\bar{\alpha}\right)\right)=\arg\min_{\bar{\alpha}}\varphi\left(A\bar{x}_k+AP\cdot\bar{\alpha}\right)$$

ובשביל לפתור בעיה זו אנחנו צריכים לחשב את הפונקציה φ בהרבה נקודות שונות כתלות ב- α. נסמן $R=AP$ ונבחין שמטריצה R היא בעלת מימדים יחסית קטנים (ביחס ל- A) (כי המימדים תלויים במטריצה P וראינו כי זו יכולה להיות גם מטריצה עם שתי עמודות בלבד) ולכן ניתן לחשב אותה במהירות, כלומר בזמן החישוב אנחנו נחשב פעם אחת את $A\bar{x}_k$ ו<u>פעם אחת</u> את $R=AP$ ולאחר מכן נבצע רק חישוב מסוג $R\cdot\bar{\alpha}$. בנוסף המטריצה P משנה לכל היותר עמודה אחת מאיטרציה לאיטרציה ואין צורך לחשב את כולה כל פעם מחדש ולכן בכל איטרציה אנחנו מחשבים עמודה אחת במטריצה R . בנוסף נבחין כי באיטרציה הבאה אנחנו צריכים לדעת את ערכו של הביטוי $A\cdot\bar{x}_{k+1}$ ולכאורה זה יגזול מאיתנו זמן חישוב גדול אך נבחין כי מתקיים:

$$A\cdot\bar{x}_{k+1}=A\bar{x}_k+R\cdot\bar{\alpha}_k$$

(ע"י כפל של מטריצה A משמאל במשוואה שראינו קודם: $\bar{x}_{k+1}=\bar{x}_k+P\cdot\alpha_k$)

ולסיכום, בכל איטרציה של האלגוריתם לפי Subspace Optimization אנחנו צריכים לחשב את הפונקציה φ מספר פעמים אך בכל פעם אנחנו צריכים לחשב בסך הכול עמודה אחת במטריצה AP .

נתבונן בנוסף בגרדיאנט ובהסיאן של הפונקציה φ ביחס לווקטור $\bar{\alpha}$: (נסמן $\bar{u}\triangleq A\bar{x}_k+R\bar{\alpha}$)

$$\nabla_{\alpha}\varphi\left(\bar{u}\right)=R^T\nabla_{\bar{u}}\varphi\left(\bar{u}\right)$$
$$\nabla^2_{\alpha\alpha}\varphi\left(\bar{u}\right)=R^T\nabla^2_{\bar{u}\bar{u}}\varphi\left(\bar{u}\right)R$$

שיטות מסוג Quasi-Newton's Method:

שיטות מסוג זה שימושיות כאשר בעיית האופטימיזציה שלנו היא ממימד לא גבוה יותר מידי, למשל עד כמה אלפי משתנים. בשיטת ניוטון המקורית, ראינו כי מספר הפעולות הנדרשות בכפל מטריצות (או פתירת מערכת משוואות) הוא סדר גודל של $\mathrm{O}\left(n^3\right)$ ובשיטות מסוג זה אנחנו נראה כי מספר הפעולות הוא בסדר גודל $\mathrm{O}\left(n^2\right)$ כאשר עיקר החיסכון בחישובים נעוץ

בעובדה שאנחנו לא נחשב את מטריצת ההסיאן בצורה מדוייקת אלא רק בצורה מקורבת ולכן שיטה זו שימושית מאד כאשר החישוב של מטריצת ההסיאן הוא קשה עד בלתי אפשרי מבחינה מעשית.

נניח כי $f\left(\bar{x}\right)$ היא "חלקה" (לפחות גזירה ברציפות) ובעית האופטימיזציה היא כרגיל:

$$\bar{x} = \arg\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right) \qquad , \qquad \bar{x}\in\mathbb{R}^n$$

כאשר הגרדיאנט של הפונקציה $f\left(\bar{x}\right)$ בזמן האיטרציה ה- k הוא $\bar{g}_k\left(\bar{x}_k\right) = \nabla f\left(\bar{x}_k\right)$.

נסמן את המטריצה ההופכית של מטריצת ההסיאן בצעד ה- k של האלגוריתם להיות:

$$B_k = \left(H\left(\bar{x}_k\right)\right)^{-1} = \left(\nabla^2 f\left(\bar{x}_k\right)\right)^{-1}$$

ונדאג לכך שהמטריצה B_k תהיה מטריצה סימטרית וכן חיובית ($B_k \succ 0$) כך שכיוון ההתקדמות יהיה תמיד בכיוון הירידה של הפונקציה $f\left(\bar{x}\right)$ (כיוון שאם $B_k \succ 0$ אז באזור הנקודה שאנחנו נמצאים הפונקציה $f\left(\bar{x}\right)$ תהיה קמורה כי אם $B_k \succ 0$ אז גם $H\left(\bar{x}\right)\succ 0$).

<u>האלגוריתם של שיטות מסוג Quasi-Newton’s Method בכל איטרציה k</u>:

1) אם תנאי העצירה מתקיימים, עצור. אחרת, חשב את הכיוון הניוטוני המקורב:
$$\bar{d}_k = -B_k \cdot \bar{g}\left(\bar{x}_k\right)$$
2) מצא את גודל הצעד בכיוון ההתקדמות, α_k, ע"י אלגוריתם Line Search (ולאו דווקא Exact Line Search אלא אפשר למשל להשתמש בחוק Armijo וזה יתרון על פני שימוש בשיטת Conjugate Gradient ששם אנחנו מחוייבים להשתמש ב-Exact Line Search אשר מצריך יותר חישובים).
3) בצע את הצעד הבא:
$$\bar{x}_{k+1} = \bar{x}_k + \alpha_k \bar{d}_k$$
4) חשב את הגרדיאנט בנקודה החדשה $\bar{x}_{k+1}$, כלומר את $\bar{g}\left(\bar{x}_{k+1}\right)$.
5) חשב את המטריצה B_{k+1} באמצעות הפרמטרים הבאים: $\left\{B_k, \bar{x}_{k+1}-\bar{x}_k, \bar{g}\left(\bar{x}_{k+1}\right)-\bar{g}\left(\bar{x}_k\right)\right\}$. (מיד נראה כיצד לחשב זאת בצורה יעילה) וחזור לשלב 1 באלגוריתם.

<u>כיצד לחשב את המטריצה ההופכית של ההסיאן המקורב בכל איטרציה k באלגוריתם – המטריצה B_k</u>:

נתחיל דווקא בהסבר כיצד לחשב את ההסיאן המקורב (ולא את המטריצה ההופכית של ההסיאן) בכל איטרציה.

נסמן תחילה:

$$\bar{p}_k \triangleq \bar{x}_{k+1} - \bar{x}_k$$
$$\bar{q}_k \triangleq \bar{g}\left(\bar{x}_{k+1}\right) - \bar{g}\left(\bar{x}_k\right)$$

ניזכר כי לפי ההגדרה הדיפרנציאלית של מטריצת ההסיאן מתקיים:

$$H\left(\bar{x}\right)d\bar{x} = d\bar{g}\left(\bar{x}\right)$$

והביטוי הזה מתאר את השינוי בגרדיאנט אם נזוז בכיוון אינפיניטיסימלי $d\bar{x}$. בפונקציה $f\left(\bar{x}\right)$ כללית כל שינוי קטן ב- $d\bar{x}$ ייתן לנו (לרוב) שינוי קטן ב- $d\bar{g}\left(\bar{x}\right)$ ושינוי גדול ב- $d\bar{x}$ יכול לתת לנו שינוי גדול ב- $d\bar{g}\left(\bar{x}\right)$. נבחין כי בפונקציה $f\left(\bar{x}\right)$ ריבועית, המצב אחר, גם עבור שינוי קטן וגם עבור שינוי גדול (לא יותר מידי) ב- $d\bar{x}$ נקבל שינוי יחסית קטן ב- $d\bar{g}\left(\bar{x}\right)$ (כי הגרדיאנט בפונקציה ריבועית הוא לינארי). תחילה נתאר את החישוב עבור פונקציה $f\left(\bar{x}\right)$ ריבועית.

<u>חישוב מטריצת ההסיאן המקורבת לפונקציה ריבועית</u>:

מתקיים (לפי ההגדרות עבור $\bar{p}_k, \bar{q}_k$ שראינו זה עתה):

$$H\left(\bar{x}_{k+1}\right)\bar{p}_k = \bar{q}_k$$

(המשוואה האחרונה נקראת משוואת המיתר (secant equation))

נניח כי בידינו החישוב המקורב הקודם של מטריצת ההסיאן, $H\left(\bar{x}_k\right)$, ואנחנו רוצים להשתמש במידע זה על מנת לחשב את המטריצה המקורבת בצעד הבא. חישוב/עדכון המטריצה בצעד הבא יכול להיות מחושב בשיטת Rank one update, אשר שיטה זו אומרת כי אנחנו מעדכנים את המטריצה ע"י הוספת התוצאה של מכפלה פנימית בין וקטור עמודה <u>כלשהו לבין וקטור שורה כלשהו</u> (שתוצאה זו היא מטריצה) למטריצה בצעד הקודם. כלומר מתקיים:

$$Rank\ One\ update:\quad H\left(\bar{x}_{k+1}\right) = H\left(\bar{x}_k\right) + \bar{u}\cdot\bar{v}^T \qquad \left(\bar{u} - Column\ Vector\right)$$

וממשוואת המיתר נקבל:

$$H\left(\bar{x}_{k+1}\right)\bar{p}_k = \left(H\left(\bar{x}_k\right) + \bar{u}\cdot\bar{v}^T\right)\bar{p}_k = \bar{q}_k$$

נבחין כי מטריצת ההסיאן היא מטריצה סימטרית ($f\left(\bar{x}\right)$ גזירה ברציפות ומשפט שוורץ מתקיים ואפשר להחליף סדר גזירה) אז אנחנו רוצים שגם המטריצה המקורבת תהיה סימטרית ולכן אנחנו צריכים לדאוג שמטריצה $\bar{u}\cdot\bar{v}^T$ תהיה סימטרית (כי אנחנו מוסיפים אותה למטריצה סימטרית ולכן בשביל להישאר עם מטריצה סימטרית אנחנו חייבים שהמטריצה $\bar{u}\cdot\bar{v}^T$ תהיה גם כן סימטרית) , אך זה אפשרי אם ורק אם $\bar{u} = \alpha\cdot\bar{v}$ כאשר $\alpha\in\mathbb{R}$, כלומר $\bar{u}\propto\bar{v}$.

וע"י פישוט קל של המשוואה האחרונה נקבל:

$$\bar{u}\left(\bar{v}^T\bar{p}_k\right) = \bar{q}_k - H\left(\bar{x}_k\right)\bar{p}_k$$

ולכן עבור כל וקטור $\bar{v}^T$ שאינו אורתוגנלי לווקטור $\bar{p}_k$, כלומר $\bar{v}^T \not\perp \bar{p}_k$ (ולכן $\bar{v}^T\bar{p}_k = \langle\bar{v}, \bar{p}_k\rangle \neq 0$) , נקבל:

$$\bar{u} = \frac{1}{\bar{v}^T\bar{p}_k}\left(\bar{q}_k - H\left(\bar{x}_k\right)\bar{p}_k\right)$$

נבחין כי $\frac{1}{\bar{v}^T\bar{p}_k}$ זה סקלר ולכן נוכל לבחור למשל:

$$\bar{v} = \bar{q}_k - H\left(\bar{x}_k\right)\bar{p}_k$$

כלומר קיבלנו:

$$\bar{u}=\frac{1}{\bar{v}^T\bar{p}_k}\bar{v}$$

ובסך הכול נקבל כי מטריצת ההסיאן בצעד הבא של האלגוריתם:

$$H(\bar{x}_{k+1})=H(\bar{x}_k)+\frac{1}{\bar{v}^T\bar{p}_k}\bar{v}\cdot\bar{v}^T$$

אך למעשה אנחנו מעוניינים במטריצה ההופכית של מטריצת ההסיאן. ישנן נוסחאות שמוצאות את המטריצה ההופכית של המטריצה החדשה בהינתן זה שהמטריצה החדשה חושבה ע"י שיטת Rank one update, למשל נוסחת Sherman-Morrison אשר יודעת לחשב את מטריצת ההסיאן ההופכית, $B_{k+1}=\left(H(\bar{x}_{k+1})\right)^{-1}$ בהינתן שהמטריצה $H(\bar{x}_{k+1})$ חושבה בשיטת Rank one update.

כעת, בהינתן זה שאנחנו יודעים את המטריצה ההופכית, ממשוואת המיתר שראינו: $H(\bar{x}_{k+1})\bar{p}_k=\bar{q}_k$, נכפול מצד שמאל במטריצה ההופכית, $B_{k+1}=\left(H(\bar{x}_{k+1})\right)^{-1}$, ונקבל:

$$B_{k+1}\bar{q}_k=\bar{p}_k$$

ונבחין כי המשוואה האחרונה דומה מאד למשוואת המיתר: $H(\bar{x}_{k+1})\bar{p}_k=\bar{q}_k$, ולכן נוכל לרשום ע"י החלפת המקומות של הווקטורים $\bar{p}_k,\bar{q}_k$ ונקבל בדומה לביטוי שקיבלנו עבור $H(\bar{x}_{k+1})$ את הביטוי עבור $B_{k+1}=\left(H(\bar{x}_{k+1})\right)^{-1}$:

$$\bar{v}=\bar{p}_k-B_k\bar{q}_k$$
$$B_{k+1}=B_k+\frac{1}{\bar{v}^T\bar{q}_k}\bar{v}\cdot\bar{v}^T$$

(לא הגדרנו מהו הווקטור $\bar{v}$ ורק אמרנו שהוא ווקטור כלשהו!)

ובזאת מצאנו שיטה לחשב את המטריצה $B_{k+1}=\left(H(\bar{x}_{k+1})\right)^{-1}$ אך זו לא השיטה היחידה שקיימת ויש שיטות אחרות ויעילות יותר ויש עוד פרטים שיש לדאוג להם על מנת שהמטריצה B_{k+1} תהיה חיובית והאלגוריתם הבא שנראה למציאת B_{k+1} נחשב לאלגוריתם הטוב ביותר שהומצא עד היום:

<u>שיטות מסוג Broyden family of Quasi-Newton's Method</u>:

נזכיר תחילה את הסימון:

$$\bar{p}_k\triangleq\bar{x}_{k+1}-\bar{x}_k$$
$$\bar{q}_k\triangleq\bar{g}(\bar{x}_{k+1})-\bar{g}(\bar{x}_k)$$

כעת נסמן בנוסף:

$$\begin{aligned}
&\bar{s}_k \triangleq B_k \bar{q}_k \\
&\tau_k \triangleq \bar{s}_k^T \bar{q}_k \qquad , \qquad \tau_k \in \mathbb{R} \\
&\mu_k \triangleq \bar{p}_k^T \bar{q}_k \qquad , \qquad \mu_k \in \mathbb{R} \\
&\bar{v}_k = \frac{\bar{p}_k}{\mu_k} - \frac{\bar{s}_k}{\tau_k}
\end{aligned}$$

כאשר נבחין כי τ_k, μ_k הם סקלרים. לכן בשיטות מסוג Broyden family of Quasi-Newton's Method מתקיים:

$$B_{k+1} = B_k + \frac{\bar{p}_k \cdot \bar{p}_k^T}{\mu_k} - \frac{\bar{s}_k \cdot \bar{s}_k^T}{\tau_k} + \xi_k \cdot \tau_k \cdot \bar{v}_k \cdot \bar{v}_k^T$$

כאשר החישוב האחרון הוא עדכון שמתבסס על Rank two update. כמו כן, $\xi_k \in \mathbb{R}$ הוא פרמטר.

<u>שיטת BFGS ושיטת DFP</u>:

אחת השיטות החשובות מסוג Broyden family of Quasi-Newton's Method נקראת BFGS על שם: Broyden, Fletcher, Goldfrab ו-Shanno ושיטה מוכרת נוספת מסוג זה נקראת DFP על שם Davidon, Fletcher ו-Powell. כמו כן, בכל שיטה הפרמטר $\xi_k \in \mathbb{R}$ הוא שונה למשל:

$$\begin{aligned}
&BGFS \;\Rightarrow\; \xi_k = 1 \\
&DFP \;\Rightarrow\; \xi_k = 0
\end{aligned}$$

נבחין כי בנוסחה האחרונה עבור B_{k+1} מספר פעולות החישוב הנדרש הוא $\mathrm{O}\left(n^2\right)$ כאשר $n \times n$ הוא המימד של מטריצה B_{k+1}.

<u>הערות</u>:

- השיטה BFGS היא מאד שימושית ואפשר למצוא אותה גם במטלב (Matlab) ב-Matlab Optimization Toolbox וזה נתון לבחירתנו באיזו מן השיטות לעיל להשתמש לחישובים. נזכיר כי שיטות אלו שימושיות כאשר חישוב ערך הפונקציה והגרדיאנט בנקודה הן יקרות יחסית, כי אם הפעולות האלה זולות אולי כדאי דווקא לחזור לשיטות שלמדנו לפני שהן Conjugate Gradient או Subspace Optimization.
- הנוסחה $B_{k+1} = B_k + \frac{\bar{p}_k \cdot \bar{p}_k^T}{\mu_k} - \frac{\bar{s}_k \cdot \bar{s}_k^T}{\tau_k} + \xi_k \cdot \tau_k \cdot \bar{v}_k \cdot \bar{v}_k^T$ מקיימת את משוואת המיתר $H\left(\bar{x}_{k+1}\right)\bar{p}_k = \bar{q}_k$ והמטריצה B_{k+1} היא סימטרית ומקיימת $B_{k+1} \succ 0$ וכמו כן מתקיים כי המטריצה B_{k+1} היא בעלת הנורמה המינימלית במובן של: $\min\left\|B_{k+1} - B_k\right\|$.
- במידה והפונקציה $f\left(\bar{x}\right)$ היא לא קמורה, אז בשביל שנקבל $B_{k+1} \succ 0$ ע"י הנוסחה $B_{k+1} = B_k + \frac{\bar{p}_k \cdot \bar{p}_k^T}{\mu_k} - \frac{\bar{s}_k \cdot \bar{s}_k^T}{\tau_k} + \xi_k \cdot \tau_k \cdot \bar{v}_k \cdot \bar{v}_k^T$ יש לדרוש כי ה-Line Search יקיים:

$$f'_{\bar{d}_k}\left(\bar{x}_k\right) < f'_{\bar{d}_k}\left(\bar{x}_{k+1}\right)$$

<u>שלב האתחול בשיטות מסוג Broyden family of Quasi-Newton's Method:</u>

אם איננו יודעים את מטריצת ההסיאן בשלב ההתחלה אז נוכל להתחיל את השיטה ע"י: $B_0 = I_{[n\times n]}$

ואם אנחנו יודעים את מטריצת ההסיאן באופן כזה או אחר אז נקרב את המטריצה B_0 למטריצה ההופכית של הסיאן כמה שנוכל:
$B_0 \approx H\left(\overline{x}_0\right)$

<u>הערה</u>:

- אם נשתמש בשיטת BFGS Quasi-Newton's Method על פונקציה ריבועית, ונשתמש ב-Exact Line Search (שעבור פונקציה ריבועית הוא למעשה חישוב אנליטי) אז באופן מפתיע אנחנו נקבל את אות המסלול באופטימיזציה שהיינו מקבלים לו היינו משתמשים בשיטת Conjugate Gradient ולכן נפתור את בעיית האופטימיזציה (שהיא במימד n) בדיוק ב- n צעדים ולאחר n הצעדים הללו תהיה בידינו מטריצת ההסיאן המדוייקת של הפונקציה המקורית (אפילו אם בשלב האתחול של השיטה נתחיל עם מטריצה מקורבת להסיאן $B_0 = I_{[n\times n]}$).
- אם נשתמש בשיטה BFGS Quasi-Newton's Method על פונקציה שאינה ריבועית, בדר"כ נקבל פתרון לאחר פחות איטרציות מאשר אם היינו משתמשים בשיטת Conjugate Gradient וכן שיטת BFGS אינה דורשת את שיטת Exact Line Search ולכן היא חסכונית במידה משמעותית משיטת Conjugate Gradient מבחינת כמות החישובים.
- קצב ההתכנסות של השיטה: ראינו שבשיטת Gradient Descent עבור פונקציה שאינה לינארית, קצב ההתכנסות הוא לינארי ובשיטה ניוטון על פונקציה לא לינארית קצב ההתכנסות הוא ריבועי. בשיטת Quasi-Newton קצב ההתכנסות הוא סופר-לינארי. נזכיר כי אם נסמן את נקודת המינימום להיות $\overline{x}^*$ אז קצב התכנסות לינארי אומר כי מתקיים:

$$Linear\,Convergence\ \ Rate:\ \forall k:\frac{\left\|\overline{x}_{k+1}-\overline{x}^*\right\|}{\left\|\overline{x}_k-\overline{x}^*\right\|}=Const\in\mathbb{R}$$

וקצב התכנסות סופר לינארי מקיים:

$$Super-Linear\,Convergence\ \ Rate:\lim_{k\to\infty}\frac{\left\|\overline{x}_{k+1}-\overline{x}^*\right\|}{\left\|\overline{x}_k-\overline{x}^*\right\|}\to 0$$

הרצאה 13 – Summary Of Unconstrained Optimization And Intro To Optimization With Constraints

סיכום השיטות לאופטימיזציה ללא אילוצים:

1. שיטות אופטימיזציה של בעיות במימד אחד:
   a. שיטת Golden Section – שיטה זו שימושית בבעיות אופטימיזציה ממימד אחד שבהן אנחנו יכולים רק לחשב את ערכי הפונקציה אך לא יכולים לחשב את ערכי הנגזרת (מסיבות כאלה ואחרות).
   b. שיטת Bisection – שיטה זו הייתה שימושית כאשר היינו יכולים לחשב את הנגזרת של הפונקציה בכל נקודה, ותנאי הכרחי שנקודה תהווה מינימום הוא התאפסות הנגזרת.
   c. שיטת Quadratic/Cubic Interpolation – בשיטה זו אנחנו מניחים כי בסביבת נקודת המינימום של הפונקציה, הפונקציה דומה לפולינום מדרגה שנייה או שלישית (וזה אכן נכון לרוב הפונקציות החלקות בסביבת נקודת המינימום). שימושים לשיטה זו:
      i. ניתן גם להשתמש בשיטה זו יחד עם שיטת Golden Section ובכל פעם שיש לנו 3 נקודות (בשיטת Golden Section) לחשב את המינימום באמצעים אנליטיים כי באמצעות 3 נקודות אפשר לקרב פונקציה ריבועית.
      ii. ניתן גם להשתמש בשיטת זו יחד עם שיטת Bisection ולקרב את הפונקציה שלנו בעזרת ערכי הנגזרת ולא רק ע"י ערכי הפונקציה.
   d. שיטת Inexact Line Search – למדנו ליישם שיטה זו באמצעות Backtracking Method יחד עם חוק Armijo .
2. שיטות אופטימיזציה של בעיות במספר מימדים:
   a. שיטת Steepest Descent / Gradient Descent – בכל צעד באלגוריתם הולכים בכיוון שמנוגד לכיוון הגרדיאנט (כי הגרדיאנט מצביע על כיוון העליה החדה ביותר) ולאחר כל בחירת כיוון מחפשים את גודל הצעד המתאים, כאשר בעיה זו היא למעשה בעיה במימד אחד ולכן משתמשים באחת משיטות האופטימיזציה במימד אחד. אחד החסרונות הבולטים של שיטה זו, זה כאשר נקודת המינימום נמצא במעיין "ואדי צר" (ill condition function) ואז כיווני הגרדיאנט מתנהגים כמו "זיג-זג" והתכנסות האלגוריתם היא מאד איטית.
   b. שיטת Newton – שיטה טובה עבור ill condition function. עבור בעיות מסוג Least Square Problems השתמשנו ב"תת-שיטה" של שיטת Newton וקראנו לשיטה זו Newton-Gauss Method. שיטת Newton דורשת בכל צעד של האלגוריתם לפתור מערכת גדולה של משוואות לינאריות (בגודל $n\times n$ עבור בעיות אופטימיזציה ממימד n). אם n גדול מידי (סדר גודל של אלפים/עשרות אלפים) שיטה זו לא ישימה במחשבים של ימינו.
   c. שיטת Conjugate Gradient – זו שיטה שראינו למזעור של פונקציות ריבועיות וכן פונקציות לינאריות אחרות (לא ריבועיות אך "חלקות").
      i. שיטת Truncated Newton – בשיטה זו ראינו שאפשר לפתור בעית אופטימיזציה בשיטת Newton אך את מערכת המשוואות בנוצרת בשיטת Newton נפתור לא באופן מלא אלא באופן מקורב ע"י שיטת Conjugate Gradient.

   גם בשיטת Conjugate Gradient וגם בשיטת Truncated Newton ניתן להשתמש לעיתים בפעולה שלמדנו בשם Preconditioning על מנת להקטין את ה-condition number של הבעיה ובכך לגרום לקצב התכנסות האלגוריתם לפתרון להיות מהיר יותר ובכך לפתור מהר יותר את בעית האופטימיזציה.

   d. שיטות Quasi-Newton – בכל איטרציה בשיטה זו, אנחנו בונים מטריצה אשר מקרבת את מטריצת ההסיאן האמיתית או למעשה מקרבת את הפונקציה ההופכית של ההסיאן. שיטה מסוג Quasi-Newton שלמדנו היא שיטת BFGS. שיטה מסוג Quasi-Newton היא ככל הנראה השיטה היעילה ביותר לכל בעית אופטימיזציה עד מימד של כמה אלפים בודדים של משתנים.
   e. שיטת Sequential Subspace Optimization (SESOP) – שיטה זו היא למעשה עוד הכללה של שיטת Conjugate Gradient למזעור של פונקציות לא לינאריות ולא ריבועיות וכל איטרציה של שיטה זו, אנחנו

למעשה משתמשים בגרדיאנט של הנקודה בה אנחנו עומדים ובכל הכיוונים שהשתמנו בהם עד לאיטרציה הנוכחית על מנת לחשב את הכיוון וגודל הצעד באיטרציה הנוכחית. נבחין כי בכדי לחשב כל צעד אנחנו פותרים בעית אופטימיזציה אבל במספר יחסית קטן של משתנים, כי המשתנים הם כל האיטרציות הקודמות ולא כמספר המשתנים המקורי של הבעיה. במקרה של פונקציות ריבועיות השיטה הזו תתהנג באופן זהה כמו שיטת Conjugate Gradient. שיטה זו יעילה מאד כאשר הפונקציה שאותה אנחנו ממזערים היא מהצורה: $f\left(\overline{x}\right)=\varphi\left(A\overline{x}\right)$, כלומר הפונקציה היא פונקציה לא לינארית (φ) של אופרטור לינארי כלשהו (A) וכן הקושי החישוב נעוץ דווקא בחישוב של האופרטור הלינארי ($A\overline{x}$) ולא בחישוב הפונקציה φ ואז כאשר אנחנו מחפשים את גודל וכיוון הצעד הבא, אנחנו מחפשים אותו ב-Subspace ולא צריכים לבצע את החישוב $A\overline{x}$ כל פעם מחדש (שהוא, כאמור, החישוב היקר מבחינת זמן). שיטה זו יכולה להיות יעילה מאד בבעיות אופטימיזציה אפילו עם מאות אלפי משתנים.

f. שיטת Nelder-Mead Simplex Method – שיטה זו <u>לא נלמדה בקורס</u>. שיטה זו טובה לבעיות במספר משתנים שלא עולה על עשרות בודדות. שיטה זו לא דורשת לדעת את הנגזרות של הפונקציה אותה ממזערים אלא רק את ערכיה בנקודות שנרצה.

<u>אופטימיזציה עם אילוצים</u>:

אופטימיזציה עם אילוצים זה החלק המרכזי והעיקרי של החלק השני של הקורס. תחילה נלמד מה הם התנאים ההכרחיים עבור נקודה להיות הנקודה בה מתקבל המינימום של הפונקציה אותה אנחנו ממזערים ולאחר מכן נלמד שיטות ואלגוריתמים יעילים למציאת הנקודות הללו.

בעיות אופטימיזציה עם אילוצים הן מהצורה:

$$\begin{aligned}&\min_{\overline{x}\in\mathbb{R}^n} f\left(\overline{x}\right)\\ &Subject\,To\ (S.T): \qquad g_i\left(\overline{x}\right)\le 0 \qquad , \qquad i=1,....,m\\ &\qquad\qquad\qquad\qquad\quad h_j\left(\overline{x}\right)=0 \qquad , \qquad j=1,....,k\end{aligned}$$

כאשר אנחנו נוהגים לקרוא לפונקציה שאותה אנחנו ממזערים f להיות פונקציה המטרה (Objective function או Cost function). הפונקציות $g_i:\mathbb{R}^n\to\mathbb{R}$ נקראות Inequality Constraints והפונקציות $h_j:\mathbb{R}^n\to\mathbb{R}$ נקראות Equality Constraints. בבעיות מסוג זה אנחנו מחפשים את הנקודה $\overline{x}\in\mathbb{R}^n$ אשר מביאה למינימום את הפונקציה f אך גם מקיימת את כל אי-השיוויונות בהצבה בפונקציות g_i ומקיימת את כל השיוויונות בהצבה בפונקציות h_j .

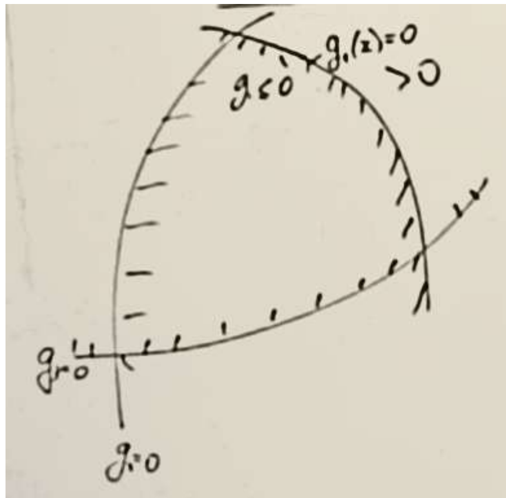
אנחנו נתחיל בשיטות לפתירת בעיות אופטימיזציה עם פונקציות מסוג Inequality Constraints בלבד ולאחר מכן נכליל לבעיות אופטימיזציה גם עם פונקציות Equality Constraints .

<u>בעיות אופטימיזציה עם פונקציות מסוג Inequality Constraints (בלבד)</u>:

בעיות אלו באופן כללי מיוצגות באופן הבא:

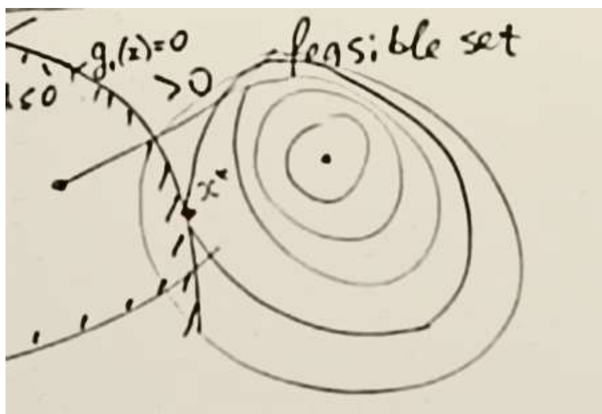
$$\begin{aligned}&\min_{\overline{x}\in\mathbb{R}^n} f\left(\overline{x}\right)\\ &Subject\,To\ (S.T): \qquad g_i\left(\overline{x}\right)\le 0 \qquad , \qquad i=1,....,m\end{aligned}$$

ננסה תחילה להבין את הגיאומטריה של הבעיה. אנחנו נצייר את קווי הגובה של הפונקציה $f\left(\overline{x}\right)$ ואת קווי הגובה של הפונקציות $g_i\left(\overline{x}\right)$. נניח ומתקיים $m=3$ ומתקיים "קווי הגובה" של הפונקציות $g_i\left(\overline{x}\right)$ הן:



כלומר הפונקציות $g_i\left(\overline{x}\right)$ בעלות תחום משותף שבו שלושתן מקבלות ערכים קטנים מאפס בכל נקודה בתחום זה. לתחום זה קוראים Feasible Set. כל נקודה $\overline{x}\in Feasible\ Set$ מקיימת את כל אילוצי הבעיה.

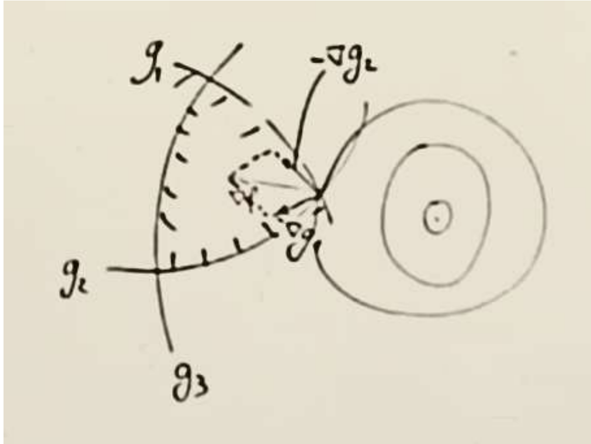
כעת נצייר את קווי הגובה של הפונקציה, $f\left(\overline{x}\right)$, אותה אנחנו רוצים למזער:



במרכז יש את נקודת המינימום (מודגשת), אך מהציור ברור כי רק הנקודה המודגשת משמאל ($\overline{x}^*$) מקיימת את האילוצים של הפונקציות $g_i\left(\overline{x}\right)$ והיא הנקודה בה מתקבל המינימום החוקי של הפונקציה $f\left(\overline{x}\right)$ בבעיה הנתונה. אנחנו יודעים כי כיוון הגרדיאנט (למשל של $f\left(\overline{x}\right)$) הוא תמיד מאונך לקווי הגובה של הפונקציה ולכן גם בנקודה $\overline{x}^*$ יש גרדיאנט שמאונך לקו הגובה המתאים. בנוסף, נביט על הגרדיאנט של הפונקציה $g_1\left(\overline{x}\right)$ בנקודה $\overline{x}^*$, ברור כי מתקיים $g_1\left(\overline{x}^*\right)=0$ (ניתן לראות בציור כי מצד אחד של הנקודה הפונקציה היא חיובית ומצד שני היא שלילית, וכיוון שהיא רציפה אז לפני משפט ערך הביניים על הקו הנ"ל הפונקציה מתאפסת) אך כיוון הגרדיאנט הוא בדיוק <u>בכיוון ההפוך</u> של כיוון הגרדיאנט של הפונקציה $f\left(\overline{x}\right)$, ובאופן פורמלי:

$$\nabla f\left(\overline{x}^*\right)=-\lambda\nabla g_1\left(\overline{x}^*\right)\qquad\Rightarrow\qquad\nabla f\left(\overline{x}^*\right)+\lambda\nabla g_1\left(\overline{x}^*\right)=0\qquad,\qquad\lambda\in\mathbb{R}$$

במקרה הנ"ל ראינו כי רק אילוץ אחד "נוגע" בנקודה $\overline{x}^*$, אך תיתכנה בעיות בהן כמה אילוצים יגעו בנקודה האופטימילית שאנחנו מחפשים, למשל שני אילוצים:



כלומר במקרה זה מתקיים (ובאופן כללי):

$$\nabla f\left(\bar{x}^*\right)+\sum_{i\in I_a}\lambda_i\nabla g_i\left(\bar{x}^*\right)=0 \quad , \quad I_a\equiv\{\text{Active Constraints}\},\ \lambda_i\in\mathbb{R}$$

כאשר I_a מייצגת את קבוצת האילוצים הפעילים (Active Constraints) אשר באמת נוגעים בנקודה (בניגוד לפונקצית האילוץ $g_3\left(\bar{x}\right)$ בדוגמה לעיל).

נגדיר את פונקצית הלגרנז'יאן:

$$L\left(\bar{x},\bar{\lambda}\right)\triangleq f\left(\bar{x}\right)+\sum_{i\in I}\lambda_i g_i\left(\bar{x}^*\right)$$

ולכן נוכל למעשה לבטא את המשוואה $\nabla f\left(\bar{x}^*\right)+\sum_{i\in I_a}\lambda_i\nabla g_i\left(\bar{x}^*\right)=0$ (שראינו זה עתה) באמצעות הלגרנז'יאן:

$$\nabla_{\bar{x}}L\left(\bar{x}^*,\bar{\lambda}^*\right)=0 \quad \Leftrightarrow \quad \nabla f\left(\bar{x}^*\right)+\sum_{i\in I_a}\lambda_i\nabla g_i\left(\bar{x}^*\right)=0$$

<u>התנאים ההכרחיים מסדר ראשון לאופטימליות של פתרון של Karush-Kuhn-Tucker (תנאיי KKT):</u>

נניח כי הנקודה $\bar{x}^*$ היא הנקודה הפותרת את בעיית האופטימיזציה הבאה:

$$\begin{aligned}&\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right)\\ &Subject\,To\ (S.T): \quad g_i\left(\bar{x}\right)\le 0 \quad , \quad i=1,\ldots,m\end{aligned}$$

ונסמן את הגרדיאנטים $\nabla g_i\left(\bar{x}^*\right)$, כאשר $i\in I_a$ (כאשר $I_a\equiv\{\text{Active Constraints}\}$), וכן נניח כי הם בלתי תלויים לינארית אחד בשני. במצב זה, קיים $\bar{\lambda}^*$ (ווקטור כופלי לגרנז') אשר עבורו נקבל:

$$\nabla_{\bar{x}}L\left(\bar{x}^*,\bar{\lambda}^*\right)=0 \quad , \quad \bar{g}\left(\bar{x}^*\right)\le 0$$

(כאשר $\bar{g}\left(\bar{x}\right)$ זה <u>סימון</u> ווקטורי לכל פונקציות האילוצים)

ועבור אילוצים מסוג Inequality Constraints מתקיים כי $\lambda_i^* \geq 0$. בנוסף, עבור אילוצים שהם לא בקבוצה I_a (אילוצים לא פעילים) מתקיים $\lambda_i^* = 0$.

אפשר לסכם את כל התנאים הללו:

$$\begin{aligned}
&\nabla_{\bar{x}} L\left(\bar{x}^*, \bar{\lambda}^*\right) = 0 \qquad , \qquad \bar{g}\left(\bar{x}^*\right) \leq 0 \\
&\lambda_i^* \geq 0 \;\; , \qquad i \in I_a \\
&\lambda_i^* = 0 \;\; , \qquad i \notin I_a
\end{aligned}$$

לתנאי אחד בלבד (תנאי Complimentary Slackness):

$$\sum_{i=1}^{m} \lambda_i^* g_i\left(\bar{x}^*\right) = 0$$

(כי עבור כל אילוץ לא פעיל, מתקיים $\lambda_i^* = 0$ ועבור כל אילוץ פעיל מתקיים $g_i\left(\bar{x}^*\right) = 0$)

<u>שיטת פונקציית עונשין (Penalty function method) עבור אילוצים מסוג Inequality Constraints</u>:

את בעיית האופטימיזציה:

$$\begin{aligned}
&\min_{\bar{x} \in \mathbb{R}^n} f\left(\bar{x}\right) \\
&Subject\,To\ (S.T): \qquad g_i\left(\bar{x}\right) \leq 0 \qquad , \qquad i = 1, \ldots, m
\end{aligned}$$

ניתן גם לרשום באופן הבא:

$$\begin{aligned}
&\min_{\bar{x} \in G} f\left(\bar{x}\right) \\
&G = \left\{ \bar{x} \in \mathbb{R}^n \mid g_i\left(\bar{x}\right) \leq 0 \;\; , \; i = 1, \ldots, m \right\}
\end{aligned}$$

אנחנו נגדיר פונקציית עונשין ($\varphi : \mathbb{R} \to \mathbb{R}$) לכל אחת מפונקציות האילוצים $g_i\left(\bar{x}\right)$, כך שעבור כל $g_i\left(\bar{x}\right)$ (אשר לא מקיים את האילוץ) תקבל ערך ("עונשין") שונה, וככל שערך זה גדול יותר כך זה אומר כי הנקודה הנוכחית שאנחנו נמצאים בה פחות קרובה לפתרון האופטימלי. כלומר יהיו אילוצים שנרצה לעמוד בהם באופן נוקשה ויהיו אילוצים שנאפשר הפרה שלהם (אם העונש לא גדול מידי). נגדיר פרמטר $p \in \mathbb{R}$ באופן הבא:

עבור $p \to 0$ נקבל כי פונקציית העונשין φ_p, נותנת עונש "קל" עבור חריגה של $g_i\left(\bar{x}\right) = t \geq 0$ ועבור $p \to \infty$ העונש עבור חריגה הוא אינסופי. למשל אפשר לחשוב על פונקציית העונשין הבאה:

$$\varphi_p\left(t\right) = e^{pt} - 1 \;\; , \qquad t, p \in \mathbb{R}$$

בנוסף, עבור $t \leq 0$ נרצה שהעונש (ערך הפונקציה $\varphi\left(t\right)$) יהיה מינימלי (ולכן הפוקציה שהוגדרה לעיל לא מתאימה לחלוטין לדרישות שלנו). הדרישות שלנו מפונקציית העונשין הן:

1) פונקציה קמורה.
2) פונקציה מונוטונית עולה.
3) פונקציה "חלקה".

4) תקיים: $\lim_{t\to\infty}\varphi'(t)=\infty$ וכן $\lim_{t\to-\infty}\varphi'(t)=0$.

5) תקיים: $\varphi(0)=0$ וגם $\varphi'(0)=1$ (דרישה זו היא לשם נוחות)

נגדיר את פרמטר העונשין (Penalty Parameter): $p\in\mathbb{R}^+\setminus\{0\}$

נגדיר גם את פונקצית העונשין $\varphi_p(t)$ להיות:

$$\varphi_p(t)=\frac{1}{p}\varphi(pt)$$

נראה שעבור $t>0$ ועבור ערך p גדול אנחנו מקבלים עונש "כבד", נבדוק את הנגזרת הראשונה ונקבל:

$$\frac{\partial}{\partial t}\left(\varphi_p(t)\right)=\frac{1}{p}\varphi'(pt)\cdot p=\varphi'(pt)$$

ולכן עבור ערך p גדול, הארגומנט pt הוא מאד גדול ולכן לפי הדרישה שלנו מתקיים: $\lim_{t\to\infty}\varphi'(t)=\infty$, כלומר אפילו חריגה קלה מהתחום המותר עבור t תגרום לעונש "כבד". גם במקרה ההפוך, אם p הוא מאד גדול, וכן $t<0$ אז פונקצית העונשין שואפת לאפס.

<u>דוגמאות לפונקציות עונשין:</u>

1) $\varphi(t)=e^t-1$

2) $\varphi(t)=\begin{cases}\frac{1}{2}t^2+t & ,t\ge-\frac{1}{2}\\ -\frac{1}{4}\log(-2t)-\frac{3}{8} & ,t>-\frac{1}{2}\end{cases}$ (ניתן לבדוק כי הפונקציה הנ"ל מקיימת את כל הדרישות מפונקצית עונשין כפי שהגדרנו אותן)

<u>פונקצית העונשין האופטימלית/אידיאלית (Ideal Penalty Function):</u>

באופן כללי אפשר לומר כי פונקצית העונשין האופטימלית/אידיאלית (Ideal Penalty Function) מקיימת:

$$\lim_{p\to\infty}\varphi_p(t)=\varphi_\infty(t)=\left\{\begin{matrix}0 & ,t\le0\\ \infty & ,t>0\end{matrix}\right\}=Ideal\ \ Penalty\ \ Function$$

<u>פונקצית העונשין המצרפית (Penalty Aggregate):</u>

נוכל להגדיר את הפונקציה שאותה אנחנו ממזערים בבעית האופטימיזציה באופן הבא:

$$F_p(\bar{x})=f(\bar{x})+\sum_{i=1}^{m}\varphi_p\left(g_i(\bar{x})\right)$$

הרעיון הכללי: בעיה זו ניתן לפתור ע"י בחירת ערך התחלתי p שהוא יחסית נמוך ואז בעיית אופטימיזציה זו הופכת לבעית אופטימיזציה ללא אילוצים, ואז מתחילים להעלות את הפרמטר p ומקבלים בעית אופטימיזציה עם יותר ויותר אילוצים. אם נצליח למצוא פתרון עבור $p\to\infty$ אז נוכל לומר שפתרנו את בעית האופטימיזציה באופן מוחלט.

<u>בעית האופטימיזציה המצרפית האידיאלית (Ideal Penalty Aggregate)</u>:

$$F_\infty\left(\bar{x}\right) = f\left(\bar{x}\right) + \sum_{i=1}^{m} \varphi_\infty\left(g_i\left(\bar{x}\right)\right) = \begin{cases} f\left(\bar{x}\right) & , \quad \bar{x} \in G = \{Feasible\ Set\} \\ \infty & , \quad \bar{x} \notin G = \{Feasible\ Set\} \end{cases}$$

כלומר שתי בעיות האופטימיזציה הבאות הן שקולות:

$$\min_{\bar{x}\in\mathbb{R}} F_\infty\left(\bar{x}\right) \quad \Leftrightarrow \quad \min_{\bar{x}\in G} f\left(\bar{x}\right)$$

<u>אלגוריתם שיטת פונקצית העונשין (Penalty Function algorithm)</u>:

1) התחל עם פרמטר עונשין p קטן יחסית בנקודה $\bar{x}_0$.
2) באיטרציה ה- k בצע את השלבים הבאים:
   a. הנקודה הבאה היא $\bar{x}_{k+1} \approx \arg\min F_{p_k}\left(\bar{x}\right)$, כלומר פותרים בעית <u>אופטימיזציה ללא אילוצים</u> עבור פרמטר p_k מסויים.
   b. הגדל את פרמטר העונשין p, באופן הבא: $p_{k+1} = \alpha \cdot p_k$ (כאשר $2 \le \alpha \le 10$) עד אשר $10^5 \le p \le 10^6$ (לא בהכרח המספרים הללו).

<u>הערה</u>:

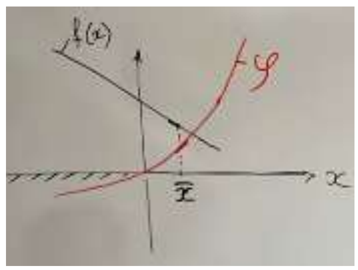
בשיטת פונקצית העונשין אנחנו אכן יכולים לקבל פתרון מדוייק יחסית, אך עבור ערכים גדולים של פרמטר השגיאה, אנחנו נאלצים לפתור בעית אופטימיזציה ללא האילוצים קשה יותר ויותר ולכן אנחנו מאלצים את הדיוק של הפתרון האמיתי של בעית האופטימיזציה המקורית שניסינו לפתור (הבעיה עם האילוצים). בהרצאה הבאה נראה איך להתמודד עם מצב שכזה.

הרצאה – Penalty Function Method and Augmented Lagrangian Method For Constrained Optimization

נתבונן בבעית אופטימיזציה פשוטה (ממימד אחד):

$$\begin{aligned} &\min_{x\in\mathbb{R}} f(x) \\ &Subject\,To\ (S.T): \qquad x \le 0 \\ &Penalty\ Function: \qquad \varphi(x) = e^{x} - 1 \end{aligned}$$

ולשם המחשה:



ונפתור את הבעיה באמצעות פונקצית עונשין ואנחנו רואים כי פונקצית העונשין מקבלת ערכים חיובים עבור $x > 0$.

ובעית האופטמיזציה היא למעשה:

$$\bar{x} \triangleq \arg\min_{x} \{F(x) = f(x) + \varphi(x)\}$$

<u>טענה (ללא הוכחה)</u>:

הנקודה $\bar{x}$ (הנקודה בה היא הנקודה בה הפונקציה $F(x) = f(x) + \varphi(x)$ מקבלת את המינימום שלה) מתקיים $f'(\bar{x}) = -\varphi'(\bar{x})$.

<u>הערה</u>: הטענה האחרונה היא למעשה טענה שקולה שבנקודת המינימום הלגרנז'יאן מתאפס.

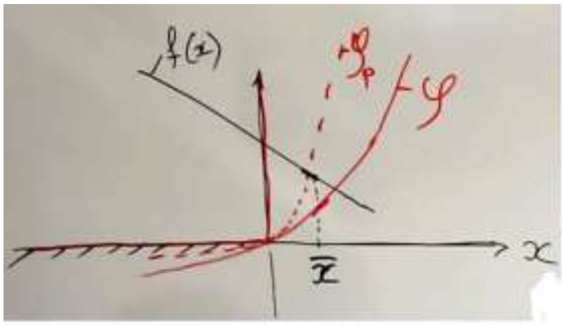
נבחין כי הפתרון במצב זה לא מדוייק, שכן, אנחנו יודעים כי הפתרון לבעית האופטימיזציה כפי שמתוארת בשרטוט:

$$\begin{aligned} &\min_{x\in\mathbb{R}} f(x) \\ &Subject\,To\ (S.T): \qquad x \le 0 \end{aligned}$$

היא למעשה $x = 0$. נגדיר כעת פונקצית העונשין עם פרמטר עונשין:

$$\varphi_p(t) = \frac{1}{p}\varphi(pt)$$

ולכן כאשר $p \to \infty$ אנחנו מקבלים כי $\varphi_p \to \varphi_\infty$ ואז מקבלים פונקצית עונשין אידיאלית:



$$\varphi_\infty = \begin{cases} 0 & , x \in Feasible\ Set \\ \infty & , x \notin Feasible\ Set \end{cases}$$

ולכן אנחנו למעשה מנסים לפתור את בעית האופטימיזציה הבאה:

$$\bar{x}_p \triangleq \arg\min_x \left\{ F_p(x) = f(x) + \varphi_p(x) \right\}$$

כאשר ערך הפרמטר p הולך וגדל.

נבחין כי כאשר p הולך וגדל, פתרון בעית האופטימיזציה ללא האילוצים הולכת ונהיית קשה יותר (מדוע?), כיוון שפונקצית העונשין נהיית יותר ויותר חדה. בכדי להתגבר על הבעיה הזו קיימת השיטה הבאה:

<u>שיטת Augmented Lagrangian Penalty – Multiplier method</u>:

פתרון בעית האופטימיזציה שנוצרת בעקבות שימוש בשיטת פונקצית העונשין נהיה יותר ויותר קשה כאשר שיפוע פונקצית העונשין הולך וגדל (וזה קורה כאשר p הולך וגדל), ולכן אנחנו רוצים להכניס פרמטר נוסף לפונקצית העונשין, פרמטר אשר "ישלוט" על שיפוע פונקצית העונשין. פרמרטר זה נהוג לסמן μ והוא נקרא ה<u>מכפיל</u>. פונקצית העונשין הנשלטת נקראת Penalty-multiplier function ומסומנת להיות $\varphi_{p\mu}(x)$ והיא מקיימת:

$$\varphi'_{p\mu}(0) = 0$$

דוגמה לפונקצית עונשין שנשלטת ע"י מכפיל כזה היא:

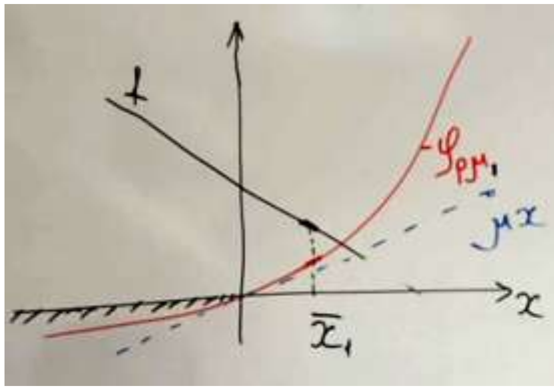
$$\varphi_{p\mu} = \mu \varphi_p(x)$$

בעיית האופטימיזציה שהצגנו קודמת:

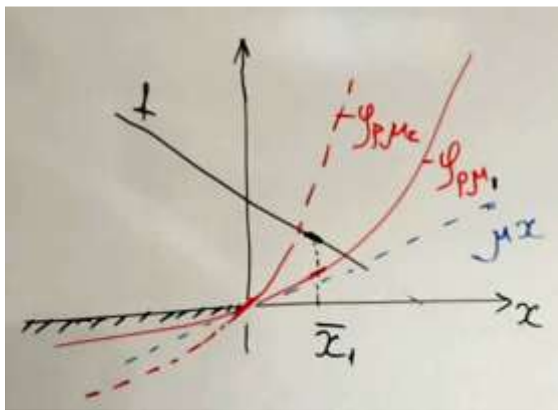
$$\begin{aligned} &\min_{x \in \mathbb{R}} f(x) \\ &Subject\ To\ (S.T): \quad x \le 0 \\ &Penalty\ Function: \quad \varphi(x) = e^x - 1 \end{aligned}$$

כאשר בכל צעד אנחנו בוחרים מכפיל μ חדש (גדול יותר מהקודם), כאשר באיטרציה הראשונה נסמנו להיות μ_1 והבעיה שאנחנו פותרים היא:

$$\bar{x}_1 \triangleq \arg\min_x \left\{ F_{p\mu_1}(x) = f(x) + \varphi_{p\mu_1}(x) \right\}$$



ובאיטרציה הבאה אנחנו בוחרים את המכפיל הבא להיות $\mu_2 = \varphi'_{p\mu_1}\left(\bar{x}_1\right)$, כלומר נקבל באיטרציה הבאה פונקצית עונשין שקרובה יותר לפונקצית עונשין אידיאלית כאשר השיפוע שלה בראשית הוא בדיוק $\mu_2 = \varphi'_{p\mu_1}\left(\bar{x}_1\right)$:



ועכשיו באיטרציה הבאה אנחנו מקבלים את הבעיה:

$$\bar{x}_2 \triangleq \arg\min_{x}\left\{F_{p\mu_2}\left(x\right) = f\left(x\right) + \varphi_{p\mu_2}\left(x\right)\right\}$$

ובדוגמה זו, למעשה באיטרציה השניה כבר נקבל את הפתרון $x_2 = 0$ (שזה הפתרון האופטימילי של הבעיה המקורית שהצגנו) וזה קרה כיוון שלפי הטענה, בנקודה האופטימלית מתקיים: $f'\left(\bar{x}\right) = -\varphi_{p\mu}\left(\bar{x}\right)$, וכיוון שבחרנו את $\mu_2 = \varphi'_{p\mu_1}\left(\bar{x}_1\right)$ אז זה אומר כי פונקצית העונשין תהיה בעלת שיפוע זהה לשיפוע הפונקציה $f\left(x\right)$ בדיוק בראשית.

<u>האלגוריתם לשיטת Augmented Lagrangian Penalty – Multiplier method:</u>

נניח ובעית האופטימיזציה המקורית היא:

$$\begin{aligned}&\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right)\\&Subject\,To\ (S.T):\qquad g_i\left(\bar{x}\right)\le 0\qquad ,\qquad i=1,....,m\end{aligned}$$

נגדיר את פונקצית העונשין המצרפית:

$$F_{p\mu}\left(\bar{x}\right) = f\left(\bar{x}\right) + \sum_{i=1}^{m}\varphi_{p\mu_i}\left(g_i\left(\bar{x}\right)\right)$$

בכל איטרציה של השיטה נמצא את הצעד הבא באמצעות פתירת בעית האופטימיזציה הבאה:

$$\bar{x}_{k+1} = \arg\min_{\bar{x}} F_{p_k \bar{\mu}_k}\left(\bar{x}\right)$$

ונעדכן את המכפיל בכל איטרציה להיות:

$$\mu_{i,k+1} = \varphi'_{p_k \mu_{i,k}}\left(g_i\left(\bar{x}_{k+1}\right)\right)$$

(למעשה יש לנו ווקטור של מכפילים ולעיל אנחנו רואים כיצד מעדכנים כל רכיב בווקטור המכפילים)

ובכל איטרציה אנחנו גם מגדילים את פרמטר העונשין p_k באופן הבא: $p_{k+1} = \alpha \cdot p_k$ כאשר $2 \le \alpha \le 10$.

<u>בעיות אופטימיזציה עם אילוצים מסוג Equality Constraints (בלבד)</u>:

נניח ובעית האופטימיזציה היא (במימד אחד):

$$\begin{aligned} &\min_{x \in \mathbb{R}} f\left(x\right) \\ &Subject\,To\ (S.T): \qquad h_j\left(x\right) = 0 \qquad , \qquad j = 1,....,k \end{aligned}$$

במקרה זה, אנחנו נרצה פונקצית עונשין שתקבל ערכים חיוביים גם עבור ערכי נקודות x שחורגות מהכיוון החיובי וגם מהכיוון השלילי של האילוץ $h_j\left(x\right) = 0$, ולכן הכי מתבקש להגדיר פונקצית עונשין מהצורה:

$$\varphi : \mathbb{R} \to \mathbb{R} \qquad , \qquad \varphi(t) = \frac{1}{2}t^2$$

ואם נשתמש בשיטת Augmented Lagrangian Penalty – Multiplier method אז נקבל את פונקצית העונשין:

$$\varphi_{p\mu}\left(t\right) = \mu t + \frac{p}{2}t^2$$

ופונקצית העונשין המצרפית במקרה זה היא:

$$F_{p\mu}\left(x\right) = f\left(x\right) + \sum_{i=1}^{m} \varphi_{p\mu_i}\left(h_i\left(x\right)\right) = f\left(x\right) + \mu^T \bar{h}\left(x\right) + \frac{p}{2}\left\|\bar{h}\left(x\right)\right\|^2$$

כמו כן, בדוגמה זו, המכפיל (<u>הבעיה היא ממימד אחד בלבד</u>) משתנים מאיטרציה לאיטרציה בצורה הבאה:

$$\varphi'_{p\mu}\left(t\right) = \mu + pt \qquad \Rightarrow \qquad \mu_{k+1} = \varphi'_{p_k \mu_k}\left(h\left(x_{k+1}\right)\right) = \mu_k + p \cdot h\left(x_{k+1}\right)$$

הרצאה 14 – Lagrange Multipliers and Penalty Function Method. Augmented Lagrangian

בהרצאה הקודמת הגדרנו את פונקצית העונשין המצרפית ($F_p\left(\bar{x}\right)$). ניזכר כי בעית האופטימיזציה הכללית (עם אילוצים) היא:

$$\underset{\bar{x}\in\mathbb{R}^n}{\arg\min}\, F_p\left(\bar{x}\right)=\underset{\bar{x}\in\mathbb{R}^n}{\arg\min}\left(f\left(\bar{x}\right)+\sum_{i=1}^{m}\varphi_p\left(g_i\left(\bar{x}\right)\right)\right)$$

כאשר נסמן את הפתרון של בעית אופטימיזציה כללית זו להיות $\bar{x}_p$. נגדיר תנאים הכרחיים לאופטימליות של פתרון שנציע.

<u>תנאי מסדר ראשון (על הנגזרת הראשונה) לפתרון אופטימלי (Karush-Kuhn-Tucker) (תנאי KKT):</u>

כמו שראינו בעבר עבור בעית אופטימיזציה ללא אילוצים, תנאיי KKT הוא התאפסות פגרדיאנט פונקצית המטרה בנקודה האופטימלית. כיוון שהמרנו את בעית האופטימיזציה שהייתה לנו עם אילוצים לבעית אופטימיזציה ללא אילוצים, אז כעת בעיה האופטימיזציה ללא אילוצים שיש לנו היא פונקצית העונשין המצרפית ולכן תנאיי KKT הוא שוב התאפסות הגרדיאנט בנקודת הפתרון:

$$\nabla F_p\left(\bar{x}_p\right)=\nabla f\left(\bar{x}_p\right)+\sum_{i=1}^{m}\varphi'_p\left(g_i\left(\bar{x}_p\right)\right)\cdot\nabla g_i\left(\bar{x}_p\right)=0$$

נסמן את הסקלר: $\lambda_{i,p}\left(\bar{x}_p\right)\triangleq\varphi'_p\left(g_i\left(\bar{x}_p\right)\right)$ ואז נקבל:

$$\nabla F_p\left(\bar{x}_p\right)=\nabla f\left(\bar{x}_p\right)+\sum_{i=1}^{m}\lambda_{i,p}\left(\bar{x}_p\right)\cdot\nabla g_i\left(\bar{x}_p\right)=0$$

נבחין כי הביטוי האחרון הוא בדיוק הגרדיאנט של הלגרנג'יאן שראינו בהרצאה הקודמת. בנוסף, ראינו כי כאשר $p\to\infty$ אז אנחנו מקבלים כי הפתרון שאנחנו מציעים שואף לפתרון האופטימלי, כלומר $\bar{x}_p\to\bar{x}^*$, כלומר הפתרון $\bar{x}_p$ מתכנס לפתרון האמיתי של בעית האופטימיזציה המקורית עם אילוצים.

ננתח מה הם הסקלרים $\lambda_{i,p}\left(\bar{x}_p\right)=\varphi'_p\left(g_i\left(\bar{x}_p\right)\right)$ (לכל $1\le i\le m$) כאשר $p\to\infty$. עבור אילוץ לא פעיל (Non-Active Constrain) מתקיים (לפי הדרישה על פונקצית העונשין):

$$\lim_{p\to\infty}\lambda_{i,p}\left(\bar{x}_p\right)=\varphi'_p\left(g_i\left(\bar{x}_p\right)\right)=0$$

ועבור אילוצים פעילים (Active Constrains) נביט על הגרדיאנט של הלגרנג'יאן בצורה מטריצית. נביט על המטריצה $\nabla G\left(\bar{x}\right)$ שעמודותיה הן הגרדיאנטים של כל אחד מהאילוצים:

$$\nabla G\left(\bar{x}\right)=\left(\underbrace{\nabla g_{i_1}\left(\bar{x}\right)\quad\nabla g_{i_2}\left(\bar{x}\right)\quad\cdots}_{\substack{\nabla G_A\left(\bar{x}\right)\\ A\equiv Active}}\;\underbrace{\cdots\quad\nabla g_{i_1}\left(\bar{x}\right)\quad\nabla g_{i_1}\left(\bar{x}\right)}_{\substack{\nabla G_N\left(\bar{x}\right)\\ N\equiv Non-Active}}\right)$$

ואז מהדרישה של התנאי מסדר ראשון: $\nabla f\left(\bar{x}_p\right)+\sum_{i=1}^{m}\lambda_{i,p}\left(\bar{x}_p\right)\cdot\nabla g_i\left(\bar{x}_p\right)=0$ נקבל בצורה מטריצית:

$$-\nabla f\left(\bar{x}_p\right)=\sum_{i=1}^{m}\left(\nabla G_A\left(\bar{x}_p\right)\lambda_{i,p}\left(\bar{x}_p\right)+\nabla G_N\left(\bar{x}_p\right)\lambda_{i,p}\left(\bar{x}_p\right)\right)$$

עבור אילוצים לא פעילים, אנחנו מקבלים כי נגזרת פונקצית העונשין $\varphi'_p\left(\bar{x}_p\right)$ היא שואפת לאפס כיוון שאין חריגה בתחום זה (אילוץ לא פעיל) אך עבור אילוצים פעילים, אנחנו נמצאים בנקודות $\bar{x}_p$ שיכולות לחרוג (לפחות לפני שאנחנו מגיעים לפתרון האופטימלי) ולכן מעניינת אותנו הנגזרת $\varphi'_p\left(\bar{x}_p\right)$, והיא אף יותר מעניינת כאשר $p\to\infty$ כי אז פונקצית העונשין מתקרבת לפונקצית עונשין אידיאלית והנגזרת שואפת לאינסוף או אפילו נהיית לא מוגדרת. נראה כי הנגזרת תלויה בהאם המטריצה $\nabla G\left(\bar{x}\right)$ היא מטריצה מדרגה מלאה, או שבמילים אחרות אפשר לומר כי המטריצה היא Full Column Rank.

עבור $p\to\infty$ ראינו כי עבור אילוצים לא פעילים מתקיים $\lim_{p\to\infty}\lambda_{i,p}\left(\bar{x}_p\right)=0$ ולכן $\lim_{p\to\infty}\nabla G_N\left(\bar{x}_p\right)\lambda_{i,p}\left(\bar{x}_p\right)=0$ כלומר מהמשוואה האחרונה אנחנו מקבלים כעת:

$$-\nabla f\left(\bar{x}_p\right)=\sum_{i=1}^{m}\nabla G_A\left(\bar{x}_p\right)\lambda_{i,p}\left(\bar{x}_p\right)$$

אם המטריצה $\nabla G\left(\bar{x}\right)$ היא בעלת דרגה מלאה (ואז עמודותיה בלתי תלויות לינארית) חייב להיות פתרון יחיד למשוואה האחרונה וכן עבור שינויים "קטנים" במטריצה $\nabla G\left(\bar{x}\right)$ נקבל כי השינויים בפתרון ($\lambda_{i,p}\left(\bar{x}_p\right)$) יהיו "קטנים" גם כן. כמו כן, כאשר $p\to\infty$ נקבל $\lambda_{i,p}\left(\bar{x}_p\right)\to\lambda^*_{i,p}\left(\bar{x}_p\right)$, כלומר נקבל את הפתרון האופטימלי.

שיטת פונקציות עונשין (Penalty function method) עבור אילוצים מסוג Equality Constraints:

ניזכר כי בעית האופטימיצזיה עם אילוצים מסוג Equality Constraints היא מהצורה:

$$\begin{aligned}&\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right)\\&Subject\,To\ (S.T): \qquad h_j\left(\bar{x}\right)\le 0 \qquad , \qquad j=1,....,l\end{aligned}$$

במקרה זה, אנחנו נרצה פונקצית עונשין שתקבל ערכים חיוביים גם עבור ערכי נקודות x שחורגות מהכיוון החיובי וגם מהכיוון השלילי של האילוץ $h_j\left(x\right)=0$, ולכן הכי מתבקש להגדיר פונקצית עונשין מהצורה:

$$\psi:\mathbb{R}\to\mathbb{R} \qquad , \qquad \psi\left(t\right)=\frac{1}{2}t^2$$

בנוסף, נרצה גם כאן להגדיר פרמטר עונשין, p, אשר יקיים:

$$\lim_{p\to\infty}\psi_p\left(t\right)=\begin{cases}0 & ,t=0\\ \infty & ,t\neq 0\end{cases}$$

למשל ניקח את פונקצית העונשין:

$$\psi(t)=t^2$$

ונוסיף לה פרמטר עונשין בצורה הבאה:

$$\psi_p(t)=p\psi(t)$$

או למשל:

$$\psi_p(t)=\frac{1}{p}\psi(pt)$$

ונבחין כי עבור $\psi(t)=t^2$, שתי הדוגמאות עבור $\psi_p(t)$ תהיינה זהות ונקבל:

$$\psi_p(t)=pt^2$$

ניתן כעת גם להגדיר את פונקצית העונשין המצרפית:

$$F_p(\overline{x})=f(\overline{x})+\sum_{j=1}^{l}\psi_p\left(h_j(\overline{x})\right)$$

בצורה דומה למקודם, נוכל להגדיר תנאי מסדר ראשון הכרחי לאופטימליות של פתרון. אם נניח כי הפתרון האופטימלי הוא $\overline{x}^*$ אז קיים פתרון $\lambda^*\left(\overline{x}^*\right)\in\mathbb{R}^l$ כך שמתקיים:

$$\nabla F_p\left(\overline{x}^*\right)=f\left(\overline{x}^*\right)+\sum_{j=1}^{l}\lambda_j^*\left(\overline{x}^*\right)\cdot h_j\left(\overline{x}^*\right)=0$$

ראינו שני תנאים מסדר ראשון, כביכול שונים עבור מקרה של אילוצים מסוג Inequality ומסוג Equality אך למעשה אפשר להכליל את שני התנאים באמצעות זה שנאמר כי התנאי הוא כי הגרדיאנטים של האילוצים הפעילים הם בלתי תלוים לינארית וכופלי לגרנז' עבור אילוצים מסוג Inequality יהיו אי-שליליים ועבור אילוצים מסוג Equality כופלי לגרנז' יכולים להיות חיובים או שליילים.

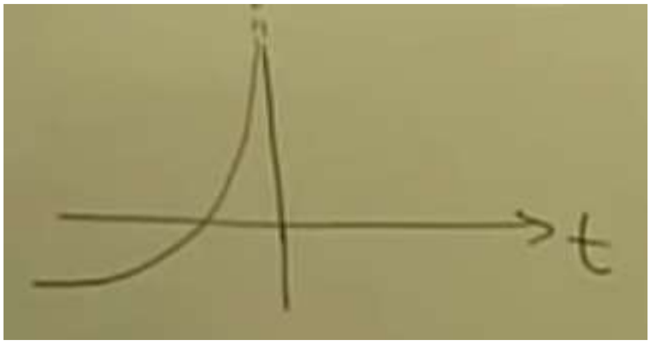
מקרה פרטי של פונקצית עונשין, פונקצית Barrier, (שיטת Barrier, שיטת המחסום):

נניח ובעית האופטימיזציה היא:

$$\begin{aligned}&\min_{x\in\mathbb{R}^n} f(\overline{x})\\&Subject\,To\ (S.T): \qquad g_i(\overline{x})\le 0 \qquad , \qquad i=1,\dots,m\end{aligned}$$

נניח ועבור נקודות $\overline{x}$ שלא עומדות באילוצים $g_i(\overline{x})\le 0$, הפונקציות (האילוצים) $g_i(\overline{x})$ כלל לא ניתנים להגדרה.

נגדיר את פונקציה המחסום (Barrier) במשתנה יחיד:



כלומר באפס יש לה אסימפטוטה ששואפת לאינסוף. ומשפחה פרמטרית של פונקציות מסוג כזה תהיה בעלת פרמטר p . עבור $p \to \infty$ נרצה כי הפונקציה תקיים:

$$\lim_{\substack{p\to\infty \\ t\to 0^-}} \varphi_p(t) = \begin{cases} \infty & ,t=0 \\ 0 & ,t<0 \end{cases}$$

ניתן גם להגדיר באופן הבא:

$$\lim_{t\to 0^-} \varphi(t) = \infty$$

$$\varphi_p(t) = \frac{1}{p}\varphi(t)$$

<u>דוגמה לפונקצית מחסום:</u>

$$\varphi(t) = -\log(-t)$$

פונקצית העונשין המצרפית, כמו קודם, היא:

$$F_p(\bar{x}) = f(\bar{x}) + \sum_{i=1}^{m} \varphi_p(g_i(\bar{x}))$$

היתרון העיקרי בשיטה זו (Barrier Method) הוא שהנקודה $\bar{x}_p$ תמיד נמצאת בתחום, כלומר תמיד עומדת בדרישת האילוץ, כלומר תמיד נמצאת ב-Feasible Set.

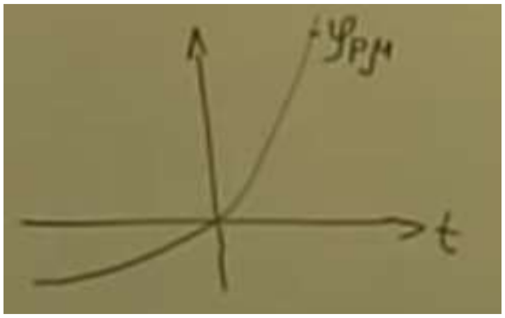
<u>הערה:</u>

כיוון שפונקצית העונשין אינה מוגדרת עבור $t>0$ אסור לחרוג מתחום זה בעת פתרון בעית האופטימיזציה, ולכן למשל כאשר מבצעים Line Search צריך לוודא כי לא חורגים מהתחום המותר.

<u>הרחבה של שיטת פונקצית העונשין, שיטת Augmented Lagrangian method</u>:

המטרה בשיטה זו היא למצוא פתרון מדוייק לבעית האופטימיזציה (עם אילוצים) ללא השאפת פרמטר העונשין p לאינסוף.

נגדיר פונקצית עונשין נשלטת הנקראת Penalty-multiplier function ומסומנת להיות $\varphi_{p\mu}(t)$. פונקציה זו מקיימת את הצורה:



מטרת הפרמטר הנוסף μ הוא לדרוש כי הנגזרת הראשונה של פונקצית העונשין בראשית תהיה בדיוק:

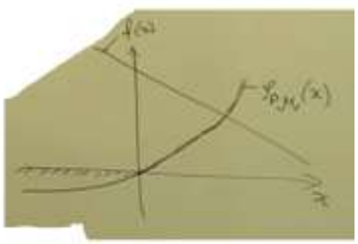
$$\varphi'_{p\mu}(0)=\mu$$

בקרוב נראה איך דרישה זו מתקשרת לשיטת האופטימיזציה באמצעות כופלי לגרנז'.

תחילה נבין את השיטה לבעית אופטימיזציה ממימד אחד. נניח כי בעית האופטימיזציה היא:

$$\begin{aligned}&\arg\min_{x\in\mathbb{R}} f(x) \quad , && x\in\mathbb{R}\\ &Subject\,To\ (S.T): && x\le 0\end{aligned}$$

ובנוסף נניח כי $f(x)$ היא <u>פונקציה לינארית יורדת</u>. ברור כי הפתרון האופטימלי הוא $x^*=0$ אך נרצה להגיע לפתרון זה באמצעות פונקצית העונשין שהגדרנו. בצורה גרפית נקבל: (באיטרציה הראשונה של האלגוריתם פונקצית העונשין היא $\varphi_{p\mu_1}$)



נגדיר את פונקצית העונשין המצרפית:

$$F_p(x,\mu)=f(x)+\varphi_{p\mu}(x)$$

כעת אנחנו מחפשים את הנקודה x_1 אשר בה מתקיים $-f'(x)=\varphi'_{p\mu_1}(x)$ (לפי התנאי מסדר ראשון, כלומר זה התנאי שגרדיאנט הפונקציה המצרפית יתאפס). כעת, מטרנו היא למצוא פונקצית עונשין חדשה, לאיטרציה הבאה באלגוריתם, $\varphi_{p\mu_2}$ אשר תקרב אותנו לפתרון האופטימלי שהוא כאמור $x^*=0$. כיוון שאנחנו יודעים כי הפונקציה אותה אנחנו ממזערים היא לינארית, אז אם נבחר פונקצית עונשין אשר בראשית שלה יש שיפוע זהה והפוך בסימן לנגזרת של הפונקציה $f(x)$ נקבל את הפתרון האופטימלי כבר באיטרציה הבאה. לפי האמור לעיל, נגדיר באיטרציה הבאה כי פונקצית העונשין <u>בראשית</u>, היא בעלת שיפוע שזהה לשיפוע של פונקצית העונשין בנקודה x_1. באופן פורמלי הגדרנו את הפרמטר μ באיטרציה הבאה לפי:

$$\mu_2=\varphi'_{p\mu_2}(0)=\varphi'_{p\mu_1}(x_1)$$

כעת נעבור לבעית אופטימיזציה במספר מימדים. נניח ובעית האופטימיזציה הבא:

$$\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right)$$
$$Subject\,To\ (S.T): \qquad g_i\left(\bar{x}\right)\le 0 \qquad , \qquad i=1,....,m$$

ופונקצית העונשין המצרפית היא:

$$F_p\left(\bar{x},\bar{\lambda}\right)=f\left(\bar{x}\right)+\sum_{i=1}^{m}\varphi_{p\lambda_i}\left(g_i\left(\bar{x}\right)\right)$$

<u>האלגוריתם Augmented Lagrangian</u>:

באיטרציה ה- k אנחנו פותרים את בעית האופטימיזציה:

$$\bar{x}_{k+1}=\arg\min_{\bar{x}\in\mathbb{R}^n} F_p\left(\bar{x},\bar{\lambda}_k\right)$$

מעדכנים את כופלי לגרנז' באופן הבא:

$$\lambda_{i,k+1}=\varphi'_{p\lambda_{k,i}}\left(g_i\left(\bar{x}_k\right)\right)$$

מעדכנים את פרמטר העונשין באופן הבא:

$$p_{k+1}=\min\left\{\alpha p_k,p_{\max}\right\}, \qquad \alpha>1,\quad 2\le\alpha\le 10, \qquad p_{\max}\in\left[100,1000\right]$$

(את פרמטר העונשין אנחנו משנים באופן מתון. אמנם ככל שהפרמטר גדול יותר כך אנחנו מקבלים פתרון מדוייק יותר אך לוקח גם יותר זמן להגיע אליו בכל איטרציה ולכן אנחנו עושים פשרה בין דיוק הפתרון לבין המהירות שמגיעים אליו).

<u>הערה</u>:

על מנת שהאלגוריתם יעבוד בצורה יעילה, רצוי כי הפרמטר $\lambda_{i,k+1}$ לא ישתנה בצורה חדה מידי ולכן כדאי להגביל את השינוי בו בתחום:

$$\frac{1}{3}\le\frac{\lambda_{i,k+1}}{\lambda_{i,k}}\le 3$$

<u>הערה</u>:

השיטה Augmented Lagrangian עובדת עם אילוצים מסוג Inequality, אך מה נעשה כאשר בעית האופטימיזציה היא עם אילוצים מסוג Equality? נוכל לומר כי מתקיים:

$$h\left(x\right)=0 \qquad \Leftrightarrow \qquad h\left(x\right)\le 0\ \ and\ \ h\left(x\right)\ge 0$$

ועכשיו קיבלנו אילוצים מסוג Inequality, כלומר נפתור שתי בעיות אופטימיזיציה עם אילוצים מסוג Inequality.

אפשרות נוספת לפתור בעית אופטימיזציה עם אילוצים מסוג Equality נוכל להגדיר Penalty-multiplier function מהצורה:

$$\varphi_{p\mu}\left(t\right)=pt^2+\mu t$$

וקל לראות כי בראשית נקבל כי השיפוע הוא תמיד μ כפי שנדרש בשיטת Augmented Lagrangian:

$$\varphi'_{p\mu}(0)=\mu$$

<u>סיכום שיטת Augmented Lagrangian</u>:

בשיטת Augmented Lagrangian ראינו ניתן למזער את פונקצית העונשין המצרפית $F_p\left(\bar{x},\bar{\lambda}^*\right)$, ולהגיע לפתרון בעית האופטימיזציה ללא השאפת פרמטר העונשין, p לאינסוף. התנאי מסדר ראשון של פונקצית העונשין המצרפית:

$$\nabla_{\bar{x}}F_p\left(\bar{x},\bar{\lambda}\right)=\nabla_{\bar{x}}f\left(\bar{x}\right)+\sum_{i=1}^{m}\varphi'_{p\lambda_i}\left(g_i\left(\bar{x}\right)\right)\cdot\nabla_{\bar{x}}g_i\left(\bar{x}\right)=0$$

ובפתרון האופטימלי הוא למעשה:

$$\nabla_{\bar{x}}F_p\left(\bar{x}^*,\bar{\lambda}^*\right)=\nabla_{\bar{x}}L\left(\bar{x}^*,\bar{\lambda}^*\right)=0$$

וזה אכן נכון כי עבור אילוצים פעילים נקבל:

$$g_i\left(\bar{x}^*\right)=0 \qquad \Rightarrow \qquad \varphi'_{p\lambda_i^*}\left(g_i\left(\bar{x}^*\right)\right)=\varphi'_{p\lambda_i^*}(0)\triangleq\lambda_i^*$$

ועבור אילוצים לא פעילים, מתקיים $\lambda_i^*=0$ ולכן נקבל:

$$g_i\left(\bar{x}^*\right)<0 \qquad \Rightarrow \qquad \varphi'_{p\lambda_i^*}\left(g_i\left(\bar{x}^*\right)\right)=0\triangleq\lambda_i^*$$

הרצאה 15 – Minmax, Game Theory, Lagrangian Duality

נרצה למצוא ביטוי שיבטא את דואליות הלגרנז'יאן נותנת לנו מבט נוסף על בעיות אופטימיזציה רבות ותעזור לנו לפשט את הבעיות. דואליות הלגרנז'יאן תאפשר לנו לדעת עד כמה אנחנו קרובים לפתרון וכן לתת גבול תחתון לפונקציה שאותה אנחנו מנסים למזער. דואליות הלגרנז'יאן היא הבסיס להרבה שיטות אופטימיזציה יעילות.

משפט Minmax:

נניח כי פונקצית המטרה היא בעלת שני משתנים, כאשר כל משתנה הוא ווקטור, כלומר $f:\left(\mathbb{R}^n,\mathbb{R}^n\right)\to\mathbb{R}$. מתקיים:

$$\min_{\bar z}\left(\max_{\bar\omega} f\left(\bar z,\bar\omega\right)\right)\ge\max_{\bar\omega}\left(\min_{\bar z} f\left(\bar z,\bar\omega\right)\right)$$

פירוש למשפט ה-Minmax בתורת המשחקים:

נתאר מהלך משחק:

נניח כי שחקן z בחר נקודה $\bar z\in Z$ ושחקן ω בוחר נקודה $\bar\omega\in\Omega$. לאחר מכן,המשחק אומר כי שחקן z משלם לשחקן ω סכום של $f\left(\bar z,\bar\omega\right)$ שקלים (ורק שחקן z משלם ל-ω). המטרה של שחקן ω לקבל הכי הרבה כסף שהוא יכול והמטרה של שחקן z לשלם כמה שפחות לשחקן ω .

במשחק יכולות להיות שתי אפשרויות לשחקן שמשחק ראשון, או שחקן ω או שחקן z .

- שחקן ω משחק שני: אם שחקן ω משחק שני (כלומר קודם שחקן z בוחר נקודה $\bar z$ ורק לאחר מכן שחקן ω בוחר נקודה $\bar\omega$) אז שחקן z (שבוחר ראשון) יבחר נקודה $\bar z$ כך יתקיים: $\min_{\bar z} f\left(\bar z,\bar\omega\right)$ (שחקן z למעשה חוזה כי שחקן ω יבחר את נקודה $\bar\omega$ כך שיתקיים $\max_{\bar\omega}\left(\min_{\bar z} f\left(\bar z,\bar\omega\right)\right)$). כעת שחקן ω כבר יודע מה שחקן z בחר, ולכן רוצה למקסם את $f\left(\bar z,\bar\omega\right)$ (כלומר למקסם את הסכום שהוא יקבל) ולכן יבחר $\max_{\bar\omega}\left(\min_{\bar z} f\left(\bar z,\bar\omega\right)\right)$ (בדיוק כפי ששחקן z "חזה").
- שחקן ω משחק ראשון: אם שחקן ω משחק ראשון, אז הוא כמובן יבחר את הנקודה $\max_{\bar\omega} f\left(\bar z,\bar\omega\right)$ (בכדי לקבל את כמות הכסף הגדולה ביותר האפשרית מבחינתו), ואז שחקן z , ירצה לבחור נקודה כך שכמות הכסף שהוא יצטרך לשלם תקטן ולכן יבחר $\min_{\bar z}\left(\max_{\bar\omega} f\left(\bar z,\bar\omega\right)\right)$ ולכן שחקן z ישלם סך הכול $\min_{\bar z}\left(\max_{\bar\omega} f\left(\bar z,\bar\omega\right)\right)$.

לפי משפט Minmax:

$$\min_{\bar z}\left(\max_{\bar\omega} f\left(\bar z,\bar\omega\right)\right)\ge\max_{\bar\omega}\left(\min_{\bar z} f\left(\bar z,\bar\omega\right)\right)$$

ולכן אם שחקן ω משחק ראשון אז הוא יקבל סכום גדול יותר. (צריך לוודא שזה אכן הניתוח הנכון למשחק)

משפט נקודת האוכף (Saddle point):

אם קיימת נקודה $\left(\bar z^*,\bar\omega^*\right)$ כך שמתקיים:

$$\forall\bar z,\bar\omega\in\mathbb{R}^n:\quad f\left(\bar z^*,\bar\omega\right)\le f\left(\bar z^*,\bar\omega^*\right)\le f\left(\bar z,\bar\omega^*\right)$$

אז מתקיים:

$$\min_{\bar{z}}\left(\max_{\bar{\omega}} f\left(\bar{z},\bar{\omega}\right)\right)=\max_{\bar{\omega}}\left(\min_{\bar{z}} f\left(\bar{z},\bar{\omega}\right)\right)$$

(דוגמה: לפונקציה $f\left(z,\omega\right)=z^2-\omega^2$ יש נקודת אוכף)

<u>הוכחה:</u>

נתון לנו כי קיימת נקודה $\left(\bar{z}^*,\bar{\omega}^*\right)$ כך שמתקיים:

$$f\left(\bar{z}^*,\bar{\omega}\right)\underset{(left)}{\le} f\left(\bar{z}^*,\bar{\omega}^*\right)\underset{(right)}{\le} f\left(\bar{z},\bar{\omega}^*\right)$$

מתקיים:

$$\min_{\bar{z}}\left(\max_{\bar{\omega}} f\left(\bar{z},\bar{\omega}\right)\right)\le\max_{\bar{\omega}} f\left(\bar{z}^*,\bar{\omega}\right)\underset{(left)}{=} f\left(\bar{z}^*,\bar{\omega}^*\right)\underset{(right)}{=}\min_{\bar{z}} f\left(\bar{z}^*,\bar{\omega}\right)\le\max_{\bar{\omega}}\left(\min_{\bar{z}} f\left(\bar{z}^*,\bar{\omega}\right)\right)$$

כלומר קיבלנו כי:

$$\min_{\bar{z}}\left(\max_{\bar{\omega}} f\left(\bar{z},\bar{\omega}\right)\right)\le\max_{\bar{\omega}}\left(\min_{\bar{z}} f\left(\bar{z}^*,\bar{\omega}\right)\right)$$

אך המשפט Minmax אומר כי מתקיים:

$$\min_{\bar{z}}\left(\max_{\bar{\omega}} f\left(\bar{z},\bar{\omega}\right)\right)\ge\max_{\bar{\omega}}\left(\min_{\bar{z}} f\left(\bar{z},\bar{\omega}\right)\right)$$

ולכן בהכרח מתקיים:

$$\min_{\bar{z}}\left(\max_{\bar{\omega}} f\left(\bar{z},\bar{\omega}\right)\right)=\max_{\bar{\omega}}\left(\min_{\bar{z}} f\left(\bar{z},\bar{\omega}\right)\right)$$

מש"ל.

<u>אי-שיוויון Minmax של הלגרנז'יאן, Minmax of Lagrangian:</u>

כעת נראה את הקשר בין משפט ה-Minmax לבין פתרון בעיות אופטימיזציה עם אילוצים. נניח ובעית האופטימיזציה היא:

$$\begin{aligned}&\min_{x\in\mathbb{R}^n} f\left(\bar{x}\right)\\&Subject\,To\ (S.T):\qquad g_i\left(\bar{x}\right)\le 0\qquad,\qquad i=1,....,m\end{aligned}$$

נזכיר כי הלגרנז'יאן הוא:

$$L\left(\bar{x},\bar{\lambda}\right)=f\left(\bar{x}\right)+\sum_{i=1}^{m}\lambda_i g_i\left(\bar{x}\right)$$

נביט בביטוי הבא:

$$\max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} L\left(\bar{x}, \bar{\lambda}\right) = \max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} f\left(\bar{x}\right) + \sum_{i=1}^{m} \lambda_i g_i\left(\bar{x}\right)$$

אם $\bar{x} \in Feasible\ Set$ אז כל האילוצים חלים עליו ולכן $\forall 1 \leq i \leq m: \ g_i\left(x\right) \leq 0$ ולכן $\sum_{i=1}^{m} \lambda_i g_i\left(\bar{x}\right) \leq 0$ (כאשר נזכור כי $\lambda_i \geq 0$).

ולכן למעשה קיבלנו כי אם $\sum_{i=1}^{m} \lambda_i g_i\left(\bar{x}\right) \leq 0$ אז $L\left(\bar{x}, \bar{\lambda}\right)$ הוא חסם תחתון של פונקצית המטרה $f\left(\bar{x}\right)$. מכל האמור לעיל נוכל לקבל (ע"י בחירת λ_i מתאימים):

$$\max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} L\left(\bar{x}, \bar{\lambda}\right) = \begin{cases} f\left(\bar{x}\right) & , \qquad x \in Feasible\ Set \\ \infty & , \qquad x \notin Feasible\ Set \end{cases}$$

למעשה קיבלנו את פונקצית העונשין המצרפית האידיאלית שסימנו בעבר להיות $F_{\infty}\left(\bar{x}\right)$. מכאן, אם נניח כי הפתרון האופטימלי הוא $\bar{x}^*$ אז נקבל כי ערך הפונקציה האופטימלי הוא:

$$f\left(\bar{x}^*\right) = \min_{\bar{x}} \max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} L\left(\bar{x}, \bar{\lambda}\right)$$

לפי משפט ה-Minmax מתקיים:

$$f\left(\bar{x}^*\right) = \min_{\bar{x}} \max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} L\left(\bar{x}, \bar{\lambda}\right) \geq \max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} \min_{\bar{x}} L\left(\bar{x}, \bar{\lambda}\right)$$

נסמן $\eta\left(\bar{\lambda}\right) = \min_{\bar{x}} L\left(\bar{x}, \bar{\lambda}\right)$, ופונקציה זו נקראת Dual function. לסיכום קיבלנו:

$$f\left(\bar{x}^*\right) \geq \max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} \eta\left(\bar{\lambda}\right)$$

הבעיה של למקסם את הפונקציה הדואלית $\eta\left(\bar{\lambda}\right)$, נקראת <u>הבעיה הדואלית</u> של מיזעור הלגרנז'יאן. אי-השיוויון $f\left(\bar{x}^*\right) \geq \max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} \eta\left(\bar{\lambda}\right)$ נקרא <u>משפט הדואליות החלש</u> (בדואליות חזקה נקבל אי-שיוויון ממש).

<u>משפט הדואליות ה"חזקה"</u>:

נניח ובעית האופטימיזציה היא:

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} f\left(\bar{x}\right) \\ &Subject\ To\ (S.T): \qquad g_i\left(\bar{x}\right) \leq 0 \qquad , \qquad i = 1, \ldots, m \end{aligned}$$

אם מתקיים כי $f\left(\bar{x}\right), g_i\left(\bar{x}\right)$ (לכל $i=1,....,m$) הן פונקציות קמורות וגם תנאי מסדר ראשון (Karush-Kuhn-Tucker KKT) מתקיימים, אז מתקיים אי-שיוויון ממש, כלומר:

$$f\left(\bar{x}^*\right) > \max_{\substack{\lambda_i \geq 0 \\ \forall i \in [1,m]}} \eta\left(\bar{\lambda}\right)$$

<u>הוכחה:</u>

נביט על הלגרנז'יאן:

$$L\left(\bar{x},\bar{\lambda}\right) = f\left(\bar{x}\right) + \sum_{i=1}^{m} \lambda_i g_i\left(\bar{x}\right)$$

הלגרנז'יאן הוא פונקציה קמורה כאשר $\lambda_i \geq 0$ לכל $1 \leq i \leq m$.

בנוסף, לפי התנאי ההכרחי מסדר ראשון מתקיים:

$$\nabla_{\bar{x}} L\left(\bar{x}^*,\bar{\lambda}^*\right) = 0$$

אבל עבור פונקציה שהיא קמורה, התנאי ההכרחי מסדר ראשון הוא גם תנאי מספיק להיותה של הנקודה $\bar{x}^*$ להיות הפתרון האופטימלי ולכן מתקיים:

$$\bar{x}^* = \arg\min_{\bar{x} \in \mathbb{R}^n} L\left(\bar{x},\bar{\lambda}^*\right)$$

ולכן מתקיים:

$$L\left(\bar{x},\bar{\lambda}^*\right) \geq L\left(\bar{x}^*,\bar{\lambda}^*\right)$$

וכן לפי התנאי מסדר ראשון, KKT, ראינו כי מתקיים Complimentary Slackness, כלומר:

$$\sum_{i=1}^{m} \lambda_i^* g_i\left(\bar{x}^*\right) = 0$$

בנוסף, כיוון שדרשנו $\lambda_i \geq 0$ לכל $1 \leq i \leq m$, אז בנקודה $\bar{x}^* \in Feasible\ Set$, האילוצים מתקיימים, כלומר $g_i\left(\bar{x}^*\right) \leq 0$ לכל $1 \leq i \leq m$ ולכן למעשה מתקיים:

$$\sum_{i=1}^{m} \lambda_i g_i\left(\bar{x}^*\right) \leq 0$$

ולכן מתקיים:

$$L\left(\bar{x}^*,\bar{\lambda}^*\right) \geq L\left(\bar{x}^*,\bar{\lambda}\right)$$

ומכאן נוכל לראות כי הנקודה $\left(\bar{x}^*,\bar{\lambda}^*\right)$ היא נקודת אוכף של הלגרנז'יאן כי הראנו כי מתקיים:

$$L\left(\bar{x},\bar{\lambda}^*\right) \geq L\left(\bar{x}^*,\bar{\lambda}^*\right) \geq L\left(\bar{x}^*,\bar{\lambda}\right)$$

כעת, כיוון שהנקודה $\left(\bar{x}^*,\bar{\lambda}^*\right)$ היא נקודת אוכף של הלגרנז'יאן, $L\left(\bar{x},\bar{\lambda}\right)$, אז לפי משפט נקודת האוכף מתקיים:

$$f\left(\bar{x}^*\right)=\min_{\bar{x}}\left(\max_{\substack{\lambda_i\geq 0\\ i\in[1,m]}}L\left(\bar{x},\bar{\lambda}\right)\right)=\max_{\substack{\lambda_i\geq 0\\ i\in[1,m]}}\left(\underbrace{\min_{\bar{x}}L\left(\bar{x},\bar{\lambda}\right)}_{\eta\left(\bar{\lambda}\right)}\right)$$

מש"ל.

<u>הערה</u>: (Slayter Condition)

קיים תנאי מספיק נוסף (גם התנאים שהוכחנו במשפט הדואליות החזקה היו תנאים מספיקים). אם קיים $\bar{x}\in Feasible\ Set$ אשר מקיים $g_i\left(\bar{x}\right)<0$ לכל אילוץ <u>לא לינארי</u> g_i (עבור אילוץ לינארי אין שום צורך ש- $\bar{x}\in Feasible\ Set$ יקיים משהו) אז מתקיימת דואליות חזקה.

<u>דוגמאות של בעיות דואליות</u>:

<u>בעיה מספר 1</u>: בעיה ריבועית

נתונה בעית האופטימיזציה הבאה:

$$\begin{aligned}&\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right)=\frac{1}{2}\bar{x}^T x\\ &Subject\,To\ (S.T): \qquad A\bar{x}-\bar{b}\leq 0\end{aligned}$$

הלגרנז'יאן עבור בעיה זו הוא:

$$L\left(\bar{x},\bar{\lambda}\right)=\frac{1}{2}\bar{x}^T\bar{x}+\bar{\lambda}^T\left(A\bar{x}-\bar{b}\right)$$

(ניזכר כי הלגרנז'יאן נתון ע"י: $L\left(\bar{x},\bar{\lambda}\right)\triangleq f\left(\bar{x}\right)+\sum_{i\in I}\lambda_i g_i\left(\bar{x}^*\right)$ כאשר $g_i\left(\bar{x}\right)$ הם האילוצים)

ככלל, הפתרון לבעיה נתון ע"י:

$$\min_{\bar{x}}\left(\max_{\bar{\lambda}}L\left(\bar{x},\bar{\lambda}\right)\right)$$

אך לפי משפט Minmax אנחנו יודעים כי מתקיים:

$$\min_{\bar{x}}\left(\max_{\bar{\lambda}}L\left(\bar{x},\bar{\lambda}\right)\right)\geq\max_{\bar{\lambda}}\left(\min_{\bar{x}}L\left(\bar{x},\bar{\lambda}\right)\right)$$

ולכן אם נפתור את הבעיה $\max_{\bar{\lambda}}\left(\min_{\bar{x}}L\left(\bar{x},\bar{\lambda}\right)\right)$ אז בוודאות נקבל פתרון שאפילו ייתכן שהוא טוב יותר מהפתרון של $\min_{\bar{x}}\left(\max_{\bar{\lambda}}L\left(\bar{x},\bar{\lambda}\right)\right)$. למעשה הבעיה שלנו כעת שקולה למציאת הפונקציה הדואלית, ובשביל זה אנחנו צריכים למזער את הלגרנז'יאן לפי $\bar{x}$. נסמן כי הפתרון האופטימלי למיזעור הלגרנז'יאן מתקבל בנקודה $\bar{x}_{\bar{\lambda}}$ ולכן בנקודה זו מתקיים תנאי הכרחי מסדר ראשון, KKT (הנגזרת/גרדיאנט הראשונה מתאפסת):

$$\nabla_{\bar{x}} L\left(\bar{x}_{\bar{\lambda}}, \bar{\lambda}\right) = \bar{x}_{\bar{\lambda}} + A^T \bar{\lambda} = 0$$

מהשיוויון האחרון נקבל:

$$\bar{x}_{\bar{\lambda}} = -A^T \bar{\lambda}$$

כאשר נציב את $\bar{x}_{\bar{\lambda}}$ בפונקצית המטרה (כאשר נזכיר כי המטרה הראשונה היא למזער את הלגרנזי'אן), כלומר בלגרנזי'אן ואז נישאר עם פונקציה שתלויה רק במשתנה $\bar{\lambda}$ כפי שרצינו. נקבל:

$$\eta\left(\bar{\lambda}\right) \triangleq \min_{\bar{x}} L\left(\bar{x}, \bar{\lambda}\right) = L\left(\bar{x}_{\bar{\lambda}}, \bar{\lambda}\right) = \frac{1}{2}\bar{\lambda}^T AA^T \bar{\lambda} - \bar{\lambda}^T AA^T \bar{\lambda} - \bar{\lambda}^T \bar{b} = -\frac{1}{2}\bar{\lambda}^T AA^T \bar{\lambda} - \bar{\lambda}^T \bar{b}$$

ולכן הבעיה הדואלית היא למקסם:

$$\max_{\bar{\lambda} \ge 0} \eta\left(\bar{\lambda}\right) = \max_{\bar{\lambda} \ge 0}\left(-\frac{1}{2}\bar{\lambda}^T AA^T \bar{\lambda} - \bar{\lambda}^T \bar{b}\right)$$

(כאשר $\bar{\lambda} \ge 0$ משמעו: $\forall i : \lambda_i \ge 0$)

נגזור לפי $\bar{\lambda}$ ונדרוש שהנגזרת תתאפס (זו למעשה פונקציה ריבועית שאנחנו מחפשים את המקסימום שלה ע"י התאפסות הנגזרת/גרדיאנט). נקבל:

$$\nabla_{\bar{\lambda}}\left(-\frac{1}{2}\bar{\lambda}^T AA^T \bar{\lambda} - \bar{\lambda}^T \bar{b}\right) = -\frac{1}{2}\left(AA^T + \left(AA^T\right)^T\right)\bar{\lambda} - \bar{b} = 0$$

ולכן קיבלנו כי נקודת המינימום מתקבלת עבור:

$$\bar{\lambda} = -2\left(AA^T + \left(AA^T\right)^T\right)^{-1}\bar{b} = -2\left(AA^T + AA^T\right)^{-1}\bar{b} = -\left(AA^T\right)^{-1}\bar{b}$$

ולכן קיבלנו כי המינימום (של בעית האופטימיזציה המקורית) מתקבל בנקודה:

$$\bar{x}_{\bar{\lambda}} = -A^T \bar{\lambda} = \left(-A^T\right)\cdot\left(-\left(AA^T\right)^{-1}\bar{b}\right) = A^T\left(AA^T\right)^{-1}\bar{b}$$

אם מטריצה A היא הפיכה אז ניתן להמשיך ולקבל:

$$\bar{x}_{\bar{\lambda}} = A^T\left(AA^T\right)^{-1}\bar{b} = A^T\left(A^T\right)^{-1}A^{-1} = A^{-1}\bar{b}$$

<u>בעיה מספר 2</u>: בעיה לינארית, תכנות לינארי

נתונה בעית האופטימיזציה הבאה:

$$\min_{\bar{x}\in\mathbb{R}^n} f\left(\bar{x}\right) = \bar{c}^T x$$
$$Subject\,To\ (S.T): \qquad -\left(A\bar{x} - \bar{b}\right) \le 0$$

הלגרנזי'אן עבור בעיה זו הוא:

$$L\left(\bar{x},\bar{\lambda}\right)=\bar{c}^T\bar{x}-\bar{\lambda}^T\left(A\bar{x}-\bar{b}\right)$$

(ניזכר כי הלגרנז'יאן נתון ע"י: $L\left(\bar{x},\bar{\lambda}\right)\triangleq f\left(\bar{x}\right)+\sum_{i\in I}\lambda_i g_i\left(\bar{x}^*\right)$ כאשר $g_i\left(\bar{x}\right)$ הם האילוצים)

נרצה למצוא את הפונקציה הדואלית, ובשביל זה אנחנו צריכים למזער את הלגרנז'יאן לפי $\bar{x}$. נסמן כי הפתרון האופטימלי למיזעור הלגרנז'יאן מתקבל בנקודה $\bar{x}_{\bar{\lambda}}$ ולכן בנקודה זו מתקיים תנאי הכרחי מסדר ראשון, KKT, (הנגזרת/גרדיאנט הראשונה מתאפסת):

$$\nabla_{\bar{x}}L\left(\bar{x},\bar{\lambda}\right)=\bar{c}-A^T\bar{\lambda}=0$$

נבחין כי לפי תנאי זה אנחנו רואים כי קיים תחום שלם (שלא תלוי ב- $\bar{x}$) שבו מתקבלת נקודת המינימום של הפונקציה $f\left(\bar{x}\right)$.
מהשיוויון האחרון, אם קיים מינימום ללגרנז'יאן אז ניתן לדרוש:

$$\bar{c}-A^T\bar{\lambda}=0$$

ולכן הבעיה הדואלית היא למקסם:

$$\eta\left(\bar{\lambda}\right)\triangleq\min_{\bar{x}}L\left(\bar{x},\bar{\lambda}\right)=\min_{\bar{x}}\left(\left(\bar{c}-A^T\bar{\lambda}\right)^T\bar{x}+\bar{\lambda}^T b\right)=\begin{cases}\bar{\lambda}^T\bar{b} & ,A^T\bar{\lambda}=\bar{c}\\ -\infty & ,otherwise\end{cases}$$

או בצורה אחרת:

$$\begin{aligned}&\max_{\bar{\lambda}\geq 0}\eta\left(\bar{\lambda}\right)=\max_{\bar{\lambda}\geq 0}\bar{\lambda}^T\bar{b}=\max_{\bar{\lambda}\geq 0}\bar{b}^T\bar{\lambda}\\ &Subject\,To\ (S.T): \qquad A^T\bar{\lambda}=\bar{c}\end{aligned}$$

(כאשר $\bar{\lambda}\geq 0$ משמעו $\forall i:\lambda_i\geq 0$)

כאשר את הבעיה הזו אפשר לפתור באמצעות פונקצית עונשין. כיוון שזו בעית אופטימיזציה שצריך למצוא בה את המקסימום (ולא את המינימום כפי שאנחנו רגילים) אז פונקצית העונשין תהיה שווה למינוס אינסוף כאשר אנחנו לא נמצאים בתחום המותר לפי האילוץ ($A^T\bar{\lambda}=\bar{c}$) וערכה יהיה אפס כאשר אנחנו נמצאים בתחום המותר.

בזאת סיימנו את החלק העיקרי השני בקורס: אופטימיזציה של פונקציות כלליות לא לינאריות. כעת מה שנשאר זה ללמוד שיטות אופטימיזציה מודרניות שהתפתחו במהלך שנות ה-90 ומתפתחות גם בשנות האלפיים. שיטות אלו ישתמשו בתכנות קוני (מהמילה: קונוס).

הרצאה 16 – Conic Programming, תכנות קוני

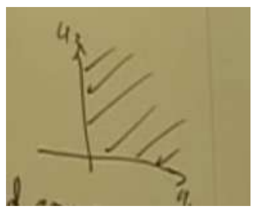
נושא התכנות הקוני הוא הנושא המסכם של הקורס והוא מתייחס לשיטות אופטימיזציה מודרניות. תכנות קוני זו למעשה הכללה של תכנות לינארי. נניח ויש לנו בעית תכנות לינארי $\left(\text{Linear Programming}\right)$ LP :

$$LP: \ \min_{\bar{x}\in\mathbb{R}^m}\langle\bar{c},\bar{x}\rangle=\min_{\bar{x}\in\mathbb{R}^m}\bar{c}^T\bar{x} \qquad ,\bar{c}\in\mathbb{R}^m$$
$$S.T\,(Subject\ To): \ \ A\bar{x}-\bar{b}\geq\bar{0} \qquad ,\bar{b}\in\mathbb{R}^n,\ A_{[n\times m]}$$

ניתן להכליל את ניסוח הבעיה הזו למחלקה רחבה יותר של בעיות שניתנות לתכנות. מה שלמדנו עד עכשיו זה תכנות לא לינארי ואת זה עשינו כאשר החלפנו את פונקצית המטרה מפונקציה לינארית לפונקציה אחרת (פונקצית עונשין מצרפית למשל) כלשהי (לרוב דרשנו שהפונקציה תהיה קמורה וגזירה ברציפות). התכנות הקוני אומר שנשאיר את ניסוח הבעיה כמו שהוא, רק שנבצע הכללה עבור האילוץ בלבד (שהוא מסוג Inequality):

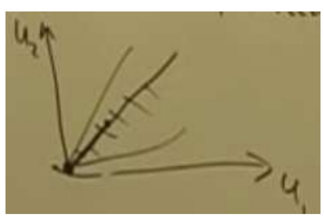
בתכנות לינארי, באי-השיוויון לעיל, הווקטור $A\bar{x}-\bar{b}$ למעשה דורשים ממנו שיקיים: $A\bar{x}-\bar{b}\in\mathbb{R}_n^+$ (כלומר כל רכיב של הווקטור הוא חיובי), כי מתקיים לפי ההגדרה:

$$\mathbb{R}_n^+=\left\{\bar{u}:\ u_i\geq 0\ \ ,\forall 1\leq i\leq n\right\}$$



ולמעשה $\mathbb{R}_n^+$ הוא קבוצה קונית אך אפשר גם להשתמש בקונוסים אחרים. נכליל את התכנות הלינארי, כלומר את הקבוצה הקונית $\mathbb{R}_n^+$ ונגדיר:

K – Convex pointed cone



(הקבוצה הקונית נתחמת ע"י שני הווקטורים ה"חיצוניים" ומכילה את כל הווקטורים ביניהם)

<u>הגדרה:</u>

קבוצה קונית מקיימת את 3 התכונות הבאות. לכל שני איברים בקונוס, $\forall\bar{x},\bar{y}\in K$, מתקיימות:

- $\lambda\bar{x}\in K,\ \ 0\leq\lambda\in\mathbb{R}$
- $-\bar{x}\notin K,\ \ \ \bar{x}\neq\bar{0}$
- $\bar{x}+\bar{y}\in K$ (חיבור וקטורי כפי שלמדנו עד כה באלגברה)

נבחין כי בקבוצה הקונית $\mathbb{R}_n^+$ מתקיימות הטענות הבאות:

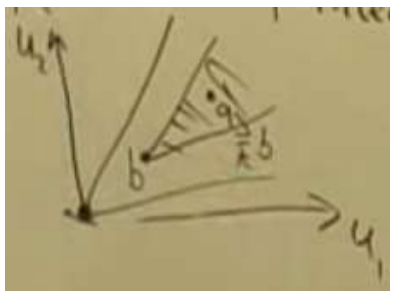
$$\bar{a} \geq 0 \quad \Leftrightarrow \quad \bar{a} \in \mathbb{R}_n^+$$
$$\bar{a} \geq \bar{b} \quad \Leftrightarrow \quad \bar{a} - \bar{b} \in \mathbb{R}_n^+$$

ונוכל להכליל טענות אלו בקבוצה הקונית K באופן הבא:

$$\bar{a} \geq_K 0 \Leftrightarrow \quad \bar{a} \in K$$
$$\bar{a} \geq_K \bar{b} \Leftrightarrow \quad \bar{a} - \bar{b} \in K$$

(כאשר האי-שיוויון $\leq_K, \geq_K$ הוא סימון לאי-שיוויון בקבוצה הקונית K)

בצורה גרפית, מתקיים $\bar{a} \geq_K \bar{b}$ אמ"מ הווקטור $\bar{a}$ נמצא בקונוס ה- K שמוזז כך שקודקודו הוא בווקטור $\bar{b}$ ולא בראשית הצירים):



<u>דוגמאות לקבוצות קוניות</u>:

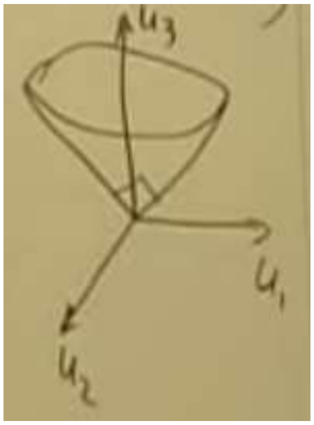
<u>דוגמה 1</u>:

הקבוצה $\mathbb{R}_n^+ = \left\{ \bar{u} : \; u_i \geq 0 \;\; , \forall 1 \leq i \leq n \right\}$.

זוהי למעשה קבוצת הווקטורים האי-שליליים (ולכן כוללת גם את ווקטור האפס) ממימד n. ווקטור אי-שלילי מוגדר להיות ווקטור שכל רכיביו הם אי-שליליים.

<u>דוגמה 2</u>:

הקבוצה L^n : הקבוצה הקונית של לורנץ (Lorentz Cone) וקבוצה זו גם נקראת Ice cream cone כיוון שהיא נראית כמו גביע גלידה. למשל בתלת מימד:



בקבוצה L^n מתקיים:

$$u_n \geq \left\|(u_1, u_2, ..., u_{n-1})\right\|_2$$

כלומר הרכיב האחרון בווקטור הוא בגודל שגדול מהנורמה של כל שאר הרכיבים בווקטור.

<u>דוגמה 3</u>:

הקבוצה S_+^m קבוצת המטריצות האי-שליליות (Positive semi-definite) הסימטריות מסדר $[m \times m]$. נוכיח את 3 התכונות הדרושות עבור קבוצה זו בכדי שתהיה קבוצה קונית:

- אם נכפיל מטריצה אי-שלילית סימטרית בסקלר חיובי התוצאה עדיין תהיה מטריצה אי-שלילית סימטרית ולכן בקבוצה.
- אם נכפיל מטריצה אי-שלילית סימטרית בסקלר שלילי נקבל כי המטריצה שלילית ולכן לא נמצאת בקבוצה.
- אם ניקח שתי מטריצות אי-שליליות סימטריות ונחבר אותן עדיין נקבל מטריצה אי-שלילית סימטרית ולכן בקבוצה.

הוכחנו את 3 התכונות ולכן קבוצה זו היא קבוצה קונית.

<u>בעיות שנפתרות ע"י תכנות קוני:</u>

בעיות שנפתרות ע"י תכנות קוני הן באופן כללי בעיות מהצורה הבאה:

$$LP: \min_{\bar{x}\in\mathbb{R}^m}\langle\bar{c},\bar{x}\rangle = \min_{\bar{x}\in\mathbb{R}^m}\bar{c}^T\bar{x} \qquad ,\bar{c}\in\mathbb{R}^m$$
$$S.T\ (Subject\ To):\ A\bar{x}-\bar{b}\geq_K \bar{0} \qquad ,A_{[n\times m]}\ ,\ \bar{b}\in\mathbb{R}^n,\quad A\bar{x}-\bar{b}\in K=\mathbb{R}_n^+$$

<u>דוגמאות לבעיות שנפתרות ע"י תכנות קוני:</u>

<u>דוגמה 1</u>:

בעיה סטנדרטית של תכנות קוני: (תכנות לינארי)

$$LP: \min_{\bar{x}\in\mathbb{R}^m}\langle\bar{c},\bar{x}\rangle = \min_{\bar{x}\in\mathbb{R}^m}\bar{c}^T\bar{x} \qquad ,\bar{c}\in\mathbb{R}^m$$
$$S.T\ (Subject\ To):\ A\bar{x}-\bar{b}\geq_K \bar{0} \qquad ,A_{[n\times m]}\ ,\ \bar{b}\in\mathbb{R}^n,\quad A\bar{x}-\bar{b}\in K=\mathbb{R}_n^+$$

<u>דוגמה 2</u>:

בעיה תכנות קוני ריבועית: (תכנות ריבועי)

$$LP:\ \min_{\bar{x}\in\mathbb{R}^m}\langle \bar{c},\bar{x}\rangle=\min_{\bar{x}\in\mathbb{R}^m}\bar{c}^T\bar{x} \qquad ,\bar{c}\in\mathbb{R}^m$$

$$S.T\ (Subject\ To):\quad \begin{bmatrix} A\bar{x}-\bar{b} \\ \bar{d}^T\bar{x}-e \end{bmatrix}\succeq_{L^{n+1}}\bar{0} \qquad ,\bar{b}\in\mathbb{R}^n,\ A_{[n\times m]}\ ,\ \bar{d}\in\mathbb{R}^m,\ e\in\mathbb{R},$$

$$\left\|A\bar{x}-\bar{b}\right\|_2\le\bar{d}^T\bar{x}-e$$

<u>דוגמה 3</u>: (דוגמה חשובה!)

בעיות תכנות קוני, אשר הקבוצה הקונית היא S_+^m. תכנות זה נקרא: Semi Definite Programming או בקיצור SDP.

נניח כי בידינו $\bar{x}\in\mathbb{R}^n$ וקבוצת מטריצות אי-שליליות סימטריות: $\{B,A_1,A_2,...,A_n\}\in S_m^+$. נגדיר אופרטור שמופעל על וקטור $\bar{x}\in\mathbb{R}^n$ באופן הבא:

$$A(\bar{x})=\left(\sum_{i=1}^n x_iA_i\right)-B\in S^m$$

(נבחין כי תמונת האופרטור זו מטריצה סימטרית כיוון שכל צירוף לינארי של מטריצות סימטריות הוא מטריצה סימטרית)

נתייחס ל- $A(\bar{x})$ כאל <u>העתקה</u> ממרחב $\mathbb{R}^n$ למרחב המטריצות ה<u>סימטריות</u>, כלומר $A:\mathbb{R}^n\to S^m$. בנוסף נסמן:

$$\mathrm{A}(\bar{x})\triangleq\sum_{i=1}^n x_iA_i \qquad , \qquad \bar{x}\in\mathbb{R}^n$$

(הערה: נבחין כי האות A שונה מהאות A וכן $\mathrm{A}(\bar{x})$ זה גם אופרטור העתקה $\mathrm{A}:\mathbb{R}^n\to S^m$)

ולכן קיבלנו:

$$A(\bar{x})=\mathrm{A}(\bar{x})-B$$

וכעת הבעיה (תכנות SDP) היא:

$$LP:\ \min_{\bar{x}\in\mathbb{R}^n}\bar{c}^T\bar{x} \qquad , \quad \bar{c}\in\mathbb{R}^n$$

$$S.T:\ A(\bar{x})=\mathrm{A}(\bar{x})-B=\left(\sum_{i=1}^n x_iA_i\right)-B\succeq_K 0\ ,\ B\in S_+^m \qquad ,\mathrm{A}(\bar{x}):\mathbb{R}^m\to S_+^m,\ A:\mathbb{R}^m\to S^m$$

$$,\ K=S_+^m \quad ,\{A_i\}_{i=1}^n\in S_+^m$$

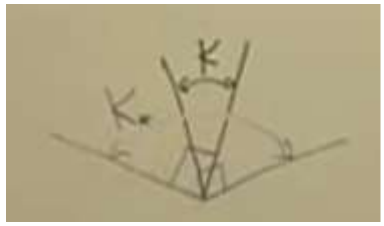
<u>קבוצה קונית דואלית, קונוס דואלי (Dual Cone):</u>

<u>הגדרה:</u>

קבוצה קונית דואלית K_* לקבוצה הקונית K היא:

$$K_*=\left\{\bar{y}:\ \langle\bar{x},\bar{y}\rangle\ge 0\ ,\ \forall\bar{x}\in K\right\}$$

ובצורה גרפית:



(אם הקונוס שלנו הוא הקונוס $K=\mathbb{R}^2$, ניתן לחשוב על הקבוצה הדואלית כעל קבוצה שמכילה את כל הווקטורים שהמכפלה הפנימית שלהם עם כל וקטור בקבוצה הקונית היא גדולה או שווה לאפס, אך המכפלה הפנימית הסטנדרטית היא למעשה כפי שהכרנו אותה עוד בפיסיקה 1: $\langle \bar{x},\bar{y}\rangle = |\bar{x}||\bar{y}|\cos(\theta)$ (כאשר θ זו הזווית בין שני הווקטורים) ולכן הקבוצה הקונית הדואלית מכילה את כל הווקטורים שהם יוצרים זווית שגדולה מ- 90° עם כל וקטור בקבוצה הקונית)

<u>הערה</u>: הקבוצה הקונית הדואלית של הקבוצה הקונית K_* היא הקבוצה הקונית K, כלומר:

$$(K_*)_* = K$$

<u>הגדרה</u>:

קבוצה קונית K שדואלית לעצמה (Self-Dual), מקיימת: $K=K_*$. בצורה גרפית:

<u>דוגמאות לקבוצות קוניות שדואליות לעצמן (תרגיל הוכחה לשיעורי הבית)</u>:

- $\mathbb{R}_m^+$ - קבוצת הווקטורים שכל רכיביהם חיוביים, והווקטורים כולם ממימד m.
- L^m - הקבוצה הקונית של לורנץ (שבה הרכיב האחרון בווקטור גדול מהנורמה של הווקטור שמורכב משאר הרכיבים)
- S_m^+ - קבוצת המטריצות הסימטריות שהן גם אי-שליליות (Positive semi-definite) מסדר $m\times m$.

<u>הוכחה כי הקבוצה הקונית S_m^+ היא דואלית לעצמה</u>:

נסמן $K=S_m^+$. אנחנו צריכים להוכיח את שתי הטענות הבאות:

$$\begin{aligned}
&(1) \qquad if \;\; X\in K \qquad \Rightarrow \qquad X\in K_* \\
&(2) \qquad if \;\; Y\notin K \qquad \Rightarrow \qquad \exists X\in K:\ \langle X,Y\rangle <0
\end{aligned}$$

(ניזכר כי הקבוצה הקונית הדואלית היא: $K_*=\{Y:\ \langle X,Y\rangle \ge 0 \ ,\ \ \forall X\in K\}$)

נוכיח את טענה (1). נניח כי בידינו $X \in K = S_+^m$ ואנחנו נוכיח (לפי ההגדרה של הקבוצה הקונית) כי לכל $Y \in S_m^+$ שנבחר נקבל $\langle X,Y \rangle \geq 0$ ולכן למעשה לפי ההגדרה נוכיח כי $X \in K_*$. נניח כי אכן $X \in K_*$. עבור כל מטריצה $Y \in S_m^+$ המכפלה הפנימית ביניהן מקיימת:

$$\langle X,Y \rangle = Trace\left(X^T Y\right) \underset{\substack{X\ symetric \\ \Rightarrow X=X^T}}{=} Trace\left(XY\right)$$

בנוסף, כל מטריצה סימטרית יכולה להיות מיוצגת ע"י מכפלה בין מטריצה משולשית למטריצה המשוכלפת שלה:

$$X = U^T U, \qquad Y = V^T V$$

ולכן:

$$\langle X,Y \rangle = Trace\left(XY\right) = Trace\left(U^T U V^T V\right) \underset{\substack{Circular\ shift \\ under\ Trace}}{=} Trace\left(\underbrace{VU^T}_{\equiv A^T} \underbrace{UV^T}_{\equiv A}\right) = Trace\left(A^T A\right) = \sum_{i=1}^{n} \left(a_{ii}\right)^2 \geq 0$$

עבור הטענה (2), אם $Y \notin K$ אז קיים וקטור $\bar{u} \in \mathbb{R}^m$ כך שמתקיים $\bar{u}^T Y \bar{u} < 0$ (שזה למעשה אומר כי המטריצה היא שלילית ולכן $Y \notin K = S_+^m$ ואז:

$$0 > \bar{u}^T Y \bar{u} \underset{\substack{\bar{u}^T Y \bar{u} \in \mathbb{R} \\ a \in \mathbb{R} \Rightarrow a = Trace(a)}}{=} Trace\left(\bar{u}^T Y \bar{u}\right) \underset{\substack{Circular\ shift \\ under\ Trace}}{=} Trace\left(\underbrace{\bar{u}\,\bar{u}^T}_{\triangleq X^T\ (matrix)} Y\right) = \langle X,Y \rangle$$

ואם נראה כי $X \in S_m^+$ אז מתקיים למעשה $\langle X,Y \rangle < 0$ ולכן הוכחנו את הטענה. נוכיח כי $\bar{u}\,\bar{u}^T = X \in S_m^+$. נבדוק את אי-שליליות המטריצה $X = \bar{u}\,\bar{u}^T$ לפי ההגדרה:

$$\forall \bar{v} \in \mathbb{R}^n : \qquad \bar{v}^T X \bar{v} = \underbrace{\bar{v}^T \bar{u}}_{a \triangleq \in \mathbb{R}} \ \underbrace{\bar{u}^T \bar{v}}_{\substack{\triangleq a^T \in \mathbb{R} \\ \Rightarrow a^T = a}} = a \cdot a^T = a^2 \geq 0$$

כעת נותר רק להוכיח כי המטריצה X היא סימטרית אך את זה קל לראות לפי ההגדרה של מכפלת וקטור בווקטור המשוכפל (זה מקרה פרטי של Outer Product).

<u>הוכחה כי הקבוצה הקונית $\mathbb{R}_m^+$ היא דואלית לעצמה (קבוצת הווקטורים ממימד m):</u>

נרצה להוכיח שוב את שתי הטענות הבאות:

$$\begin{aligned} &(1) \qquad if\ \ \bar{x} \in K \qquad \Rightarrow \qquad \bar{x} \in K_* \\ &(2) \qquad if\ \ \bar{y} \notin K \qquad \Rightarrow \qquad \exists \bar{x} \in K : \langle \bar{x}, \bar{y} \rangle < 0 \end{aligned}$$

(ניזכר כי הקבוצה הקונית, לפי ההגדרה זו הקבוצה: $K_* = \left\{ \bar{y} : \langle \bar{x}, \bar{y} \rangle \geq 0 \ , \ \forall \bar{x} \in K \right\}$)

נתחיל מטענה (2), נוכיח כי אם קיים $\bar{y} \notin K = \mathbb{R}_m^+$ אז קיים $\bar{x} \in K = \mathbb{R}_m^+$ כך שמתקיים:

$$\langle \bar{x},\bar{y}\rangle<0 \qquad \Leftrightarrow \qquad \bar{x}^T\bar{y}<0 \qquad \Leftrightarrow \qquad \sum_{i=1}^{m}x_iy_i<0$$

אך אם $\bar{y}\notin\mathbb{R}_m^+$, זה אומר כי קיים (לפחות) אינדקס אחד $1\le i\le m$ כך שמתקיים $y_i<0$. אנחנו נבחר $\bar{x}\in\mathbb{R}_m^+$ באופן הבא:

$$\bar{x}=\{x_1,...,x_{i-1},x_i,x_{i+1},...,x_m\}=\{0,...,0,1,0,...,0\}$$

ואז נקבל כי אכן מתקיים $\langle \bar{x},\bar{y}\rangle<0$ כפי שרצינו להראות.

עבור טענה (1), נניח כי $\bar{y}\in K=\mathbb{R}_m^+$ ולכן $y_i\ge 0$ לכל $1\le i\le m$ וכן אם $\bar{x}\in\mathbb{R}_m^+$ אז $x_i\ge 0$ לכל $1\le i\le m$ אז מתקיים:

$$\langle \bar{x},\bar{y}\rangle=\sum_{i=1}^{m}x_iy_i\ge 0$$

<u>דואליות בתכנות קוני מסוג SDP</u>:

כפי שראינו בתכנות לא לינארי (כאשר למדנו על הלגרנז'יאן התעסקנו עם פונקציית מטרה שהיא הייתה פונקצית עונשין מצרפית והיא הייתה לא לינארית כיוון שפונקציות העונשין היו לא לינאריות), לכל בעיה הצלחנו למצוא את הבעיה הדואלית שלה וגם במקרה זה נראה כיצד למצוא את הבעיה הדואלית של כל בעיה בתכנות קוני (הקונוס בבעיה הכללית שנציג הוא קונוס המטריצות הסימטריות האי-שליליות, כלומר S_+^m וזה כאמור קונוס שגם הוכחנו שהוא דואלי לעצמו וכן קראנו לתכנות זה תכנות SDP).
הבעיה הכללית בתכנות קוני SDP, מנוסחת באופן הבא:

$$LP:\ \min_{\bar{x}\in\mathbb{R}^n}\langle \bar{c},\bar{x}\rangle=\min_{\bar{x}\in\mathbb{R}^n}\bar{c}^T\bar{x} \qquad ,\ \bar{c}\in\mathbb{R}^n$$

$$S.T\ (Subject\ To):\ A(\bar{x})=\mathrm{A}(\bar{x})-B=\left(\sum_{i=1}^{n}x_iA_i\right)-B\succeq_K 0 \qquad ,\ B\in S^m$$

$$,\ \mathrm{A}(\bar{x})-B\in K=S_+^m$$

$$,\ \mathrm{A}(\bar{x})\triangleq\left(\sum_{i=1}^{n}x_iA_i\right)\in S^m$$

$$,\ \forall 1\le i\le n:\ A_i\in S^m$$

(נבחין כי הקבוצה S^m זו קבוצת המטריצות הסימטריות מסדר $[m\times m]$ ואילו S_+^m זו קבוצת המטריצות הסימטריות <u>האי-שליליות</u> מסדר $[m\times m]$ ונבחין כי רק מטריצת האילוץ $A(\bar{x})$ צריכה להיות סימטרית חיובית אך שאר המטריצות צריכות להיות רק סימטריות) תחילה, נבחין כי איבר בקבוצה הקונית הדואלית $Y\in K_*=S_+^m$ (הוכחנו כי הקבוצה S_+^m דואלית לעצמה) מקיים את התכונות/עובדות הבאות:

- אם $\bar{x}\in Feasible\ Set$ אז $\langle \mathrm{A}(\bar{x})-B,Y\rangle\ge 0$. <u>הוכחה</u>: נבחין כי אם $\bar{x}\in Feasible\ Set$ כלומר, $\mathrm{A}(\bar{x})-B\succeq_K 0$ אז לפי ההגדרה של קבוצה קונית דואלית, מתקיים $\langle \mathrm{A}(\bar{x})-B,Y\rangle\ge 0$.
- $\langle \mathrm{A}(\bar{x}),Y\rangle\ge_K\langle B,Y\rangle$. הוכחה: ע"י העברת אגף נקבל:

$$0 \le_K \left\langle A\left(\overline{x}\right), Y\right\rangle - \left\langle B, Y\right\rangle = A\left(\overline{x}\right)^T Y - B^T Y = \left(A\left(\overline{x}\right) - B\right)^T Y = \left\langle A\left(\overline{x}\right) - B, Y\right\rangle$$

- $\left\langle \overline{x}, \mathrm{A}^*\left(Y\right)\right\rangle \ge \left\langle \overline{b}, \mathrm{A}^*\left(Y\right)\right\rangle$, כאשר A^* מוגדר להיות אופרטור שמעביר מטריצות לווקטורים (נגדיר בהמשך).

המטרה שלנו זה למצוא איבר Y (כאשר מתקיים $\mathrm{A}^*\left(Y\right) = \overline{y}$) כך שמתקיים $\mathrm{A}^*\left(Y\right) = \overline{c}$ ואז נוכל לומר כי לפי התכונה השלישית לעיל יתקיים:

$$\left\langle \overline{x}, \overline{c}\right\rangle \ge \left\langle \overline{b}, \overline{y}\right\rangle$$

ובכך למעשה נמצא ערך גבול תחתון עבור הבעיה שאנחנו מנסים לפתור שכן מתקיים: $\left\langle \overline{x}, \overline{c}\right\rangle = \overline{c}^T \overline{x}$. בנוסף, נרצה שערך תחתון זה יהיה כמו שיותר "צמוד" (מלמטה) לערך האמיתי של הפתרון, כלומר אנחנו רוצים למקסם את הביטוי $\left\langle \overline{b}, \overline{y}\right\rangle$, כלומר לפתור:

$$\begin{aligned} &\max_{\mathrm{A}\left(\overline{y}\right)\in K_*} \left\langle \overline{b}, \overline{y}\right\rangle = \overline{b}^T \overline{y} && , \quad \mathrm{A}\left(\overline{b}\right) \in K_* = S_+^m \\ &S.T\ \left(Subject\ To\right):\ \ \mathrm{A}^*\left(Y\right) = \overline{c} && , \quad \mathrm{A}\left(\overline{c}\right) \in K_* = S_+^m \end{aligned}$$

וזו למעשה <u>הבעיה הדואלית של תכנות קוני SDP</u>. זו נקראת הבעיה הדואלית, כי במקום למצוא את ה<u>מינימום</u> של פונקצית המטרה שנדרשנו, אנחנו מנסים למצוא את ה<u>מקסימום</u> של פונקצית מטרה אחרת.

<u>משפט הדואליות החלשה:</u>

עבור הבעיה:

$$\begin{aligned} &LP:\ \min_{\overline{x}\in\mathbb{R}^n} \left\langle \overline{c}, \overline{x}\right\rangle && , \overline{c} \in \mathbb{R}^n \\ &S.T\,(Subject\ To):\ \ A\left(\overline{x}\right) = \mathrm{A}\left(\overline{x}\right) - B \succeq_K 0 && , K = S_+^m \end{aligned}$$

ואם נניח כי $Y \in K_*$ וכן $\mathrm{A}^*\left(Y\right) = \overline{c}$ אז מתקיים:

$$\left\langle \overline{x}, \overline{c}\right\rangle \ge \left\langle \overline{b}, \overline{y}\right\rangle$$

כאשר $\overline{x} \in Feasible\ Set$ וכן $\overline{y} \in Dual\ \ Feasible\ Set$.(כאשר נזכור כי הוכחנו כי הקבוצה $\mathbb{R}_n^+$ דואלית לעצמה)

<u>משפט הדואליות החזקה:</u>

כי שראינו, בעית התכנות הקוני והבעיה הדואלית לה הן:

$$\left(\text{Primary Problem } = P\right): \quad \begin{cases} LP:\ \min_{\overline{x}\in\mathbb{R}^n} \left\langle \overline{c}, \overline{x}\right\rangle & , \overline{c} \in \mathbb{R}^n \\ S.T\,(Subject\ To):\ \ \mathrm{A}\left(\overline{x}\right) - B \succeq_K 0 & , K = S_+^m \end{cases}$$

$$\left(\text{Dual Problem } = D\right): \quad \begin{cases} \max_{\mathrm{A}\left(\overline{y}\right)\in K_*} \left\langle \overline{b}, \overline{y}\right\rangle & , \overline{b} \in \mathbb{R}^n, \mathrm{A}\left(\overline{y}\right) \in K_* = S_+^m \\ S.T\ \left(Subject\ To\right):\ \ \mathrm{A}^*\left(Y\right) = \overline{c} & , \overline{c} \in \mathbb{R}^n \end{cases}$$

אם בבעיה (P), מצאנו פתרון $\bar{x}^* \in \mathbb{R}^n$ אשר מקיים $\mathrm{A}(\bar{x}^*) - B >_K 0$ **כלומר** נמצא בקבוצה הקונית ממש, כלומר $\bar{x}^* \in Strictly\ Feasible\ Set$ אז לבעיה הדואלית (D) יש פיתרון והפתרון האופטימלי מקיים שהפער הדואלי (duality gap) הוא אפס:

$$duality\ \ gap = \langle \bar{c}, \bar{x}^* \rangle - \langle \bar{b}, \bar{y}^* \rangle = 0$$

או אם בבעיה (D), מצאנו פתרון $\bar{y}^* \in \mathbb{R}^n$ אשר $\mathrm{A}^*(Y)$ נמצא בקבוצה הקונית ממש, כלומר $\bar{y}^* \in Strictly\ Feasible\ Set$ אז לבעיה (P) יש פיתרון והפתרון האופטימלי מקיים גם הוא את השיוויון לעיל.

<u>הערה</u>:

במשפט הדואליות בתכנות לינארי, לא דרשנו את התנאי של Strict Feasiblty. משפט הדואליות החזקה האחרון (לעיל) מזכיר במידה מסויימת את Slayter Condition עבור דואליות חזקה בתכנות לא לינארי. התנאי בדבר Strict Feasiblty הוא תנאי יחסית חזק שצריך לקיים בשביל הדואליות החזקה אז לרוב תנאי זה מתקיים.

<u>הערה</u>: Complimentary Slackness

כמו בתכנות הלינארי שדרשנו שכופלי לגרנז' יקיימו: $\left(\bar{\lambda}^*\right)^T \bar{g}\left(\bar{x}^*\right) = 0$ אז במקרה של תכנות קוני מתקיים:

$$\left\langle Y^*, \mathrm{A}\left(\bar{x}^*\right) - B \right\rangle = 0$$

או באופן שקול:

$$\left\langle Y^*, \mathrm{A}\left(\bar{x}^*\right) - B \right\rangle - \left\langle Y^*, B \right\rangle = 0$$

אך זה ניתן גם לרישום אחר (באמצעות אופרטור האדג'וינט $\mathrm{A}^*(\cdot)$):

$$\left\langle \mathrm{A}^*\left(Y^*\right), \bar{x}^* \right\rangle - \left\langle \bar{y}^*, \bar{b} \right\rangle = 0$$

ולאחר שימוש באילוץ $\mathrm{A}^*(Y) = \bar{c}$ נקבל:

$$\left\langle \bar{c}, \bar{x}^* \right\rangle - \left\langle \bar{y}^*, \bar{b} \right\rangle = 0$$

שזהו בדיוק הפער הדואלי (duality gap).

<u>דוגמה</u>: Semi Definite Programming (SDP)

תזכורת: נניח $\bar{x} \in \mathbb{R}^n$, וכן קבוצת מטריצות: $\{B, A_1, A_2, ..., A_n\} \in S^m$. האופרטור הוא:

$$A(\bar{x}) = \left(\sum_{i=1}^{n} x_i A_i\right) - B \triangleq \mathrm{A}(\bar{x}) - B$$

ובעית האופטימיזציה היא:

$$(P):\begin{cases} LP: \ \min\limits_{\bar{x}\in\mathbb{R}^n} \bar{c}^T\bar{x} & ,\bar{c}\in\mathbb{R}^n \\ S.T\,(Subject\ To): \ A(\bar{x}) = \mathrm{A}(\bar{x}) - B \succeq_K 0 & ,K = S_+^m \end{cases}$$

והבעיה הדואלית:

$$(D):\begin{cases} \max\limits_{Y\succeq_{K_*}0} \langle B,Y\rangle \equiv Trace\left(B^TY\right) & ,B\in K_* = S_+^m \\ S.T\ (Subject\ To): \ \mathrm{A}^*(Y) = \bar{c} & ,\bar{c}\in\mathbb{R}^n \end{cases}$$

(כאשר נבחין כי S_+^m היא קבוצה דואלית לעצמה ולכן אם $Y\in K$ אז גם $Y\in K_*$)

נמצא איך מחשבים את האילוץ $\mathrm{A}^*(Y)$ כלומר את הגדרת אופרטור האדג'וינט . נבחין כי מתקיים:

$$\left\langle \bar{x},\mathrm{A}^*(Y)\right\rangle = \left\langle \mathrm{A}(\bar{x}),Y\right\rangle \equiv \left\langle \left(\sum_{i=1}^n x_iA_i\right),Y\right\rangle = \sum_{i=1}^n \left(x_i\left\langle A_i,Y\right\rangle\right) = \left\langle \bar{x},\begin{pmatrix}\langle A_1,Y\rangle \\ \vdots \\ \langle A_n,Y\rangle\end{pmatrix}\right\rangle$$

כלומר מתקיים:

$$\left\langle \bar{x},\mathrm{A}^*(Y)\right\rangle = \left\langle \bar{x},\begin{pmatrix}\langle A_1,Y\rangle \\ \vdots \\ \langle A_n,Y\rangle\end{pmatrix}\right\rangle$$

וזה אומר כי ניתן לחשב את הביטוי $\mathrm{A}^*(Y)$ באופן הבא:

$$\mathrm{A}^*(Y) = \begin{pmatrix}\langle A_1,Y\rangle \\ \vdots \\ \langle A_n,Y\rangle\end{pmatrix} = \begin{pmatrix} Trace\left(A_1^TY\right) \\ \vdots \\ Trace\left(A_n^TY\right)\end{pmatrix}$$

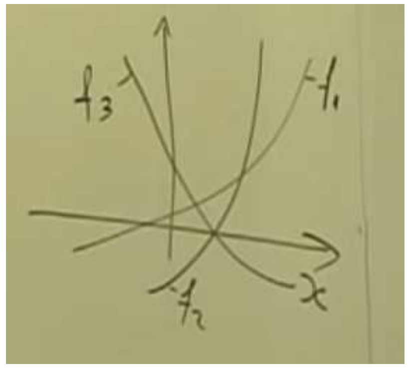
וזו הגדרת אופרטור האדג'וינט (אופרטור אדג'וינט) כאשר נזכור כי $\{A_1,A_2,...,A_n\}\in S_+^m$

<u>כיצד בעית Minmax קשורה לתכנות SDP (Semi Definite Programming)</u>:

נראה יישום של תכנות Semi Definite באמצעות דוגמה של בעית Minmax דיסקרטית. נפתור בעיות במימד אחד.

<u>בעית Minmax ראשונה</u>:

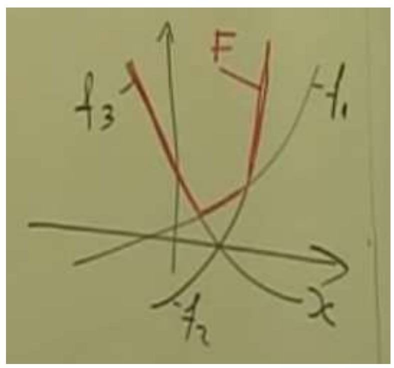
נניח את המצב הבא:



ונגדיר:

$$F\left(x\right) \triangleq \max_{i=\{1,\ldots,m\}} f_i\left(x\right)$$

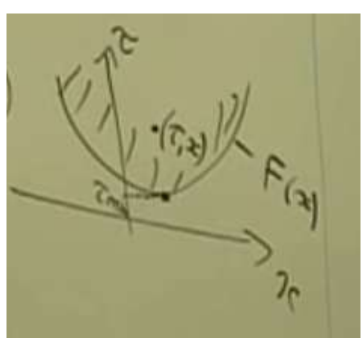
ואם נרצה לשרטט את $F\left(x\right)$ נקבל (באדום):



ובעית האופטימיזציה היא:

$$\min_{x\in\mathbb{R}} F\left(x\right)$$

זו נראת כמו בעית אופטימיזציה ללא אילוצים אך נשים לב אליו הוא שהפונקציה $F\left(x\right)$ היא לא גזירה בנקודות "חיבור", ויש לה "שפיצים" בהם הפונקציה לא "חלקה". נביט על הרישום הבא (כאשר למעשה קראנו לציר האנכי בשם $\tau \equiv F\left(x\right)$):



נרצה לפתור את בעית האופטימיזציה עם אילוצים:

$$\min_{\tau,x} \tau$$
$$S.T\ (Subject\ To):\ (x,\tau)\in epigraph(F)$$

ובאופן שקול ניתן גם לכתוב:

$$\min_{\tau,x} \tau$$
$$S.T\ (Subject\ To):\ \tau \ge F(x)$$

או באופן שקול נוסף:

$$\min_{\tau,x} \tau$$
$$S.T\ (Subject\ To):\ \tau \ge f_i(x)\quad ,\qquad i=1,\dots,m$$

כלומר הפכנו בעית אופטימיזציה ללא אילוצים שפונקצית המטרה שלה לא "חלקה" לבעית אופטימיזציה עם אילוצים אשר כל אילוץ בה הוא "חלק" (אבל עדיין פונקצית המטרה לא "חלקה" – אז למה מה שעשינו הוא עדיף?)

<u>בעית Minmax שניה</u>:

נגדיר פעם נוספת (כמו בבעית Minmax הראשונה לעיל) את הפונקציה $F(x)$ אך בצורה קצת שונה:

$$F(x) \triangleq \max_{i=\{1,\dots,m\}} |f_i(x)|$$

ונרצה לפתור את בעית האופטימיזציה:

$$\min_{x\in\mathbb{R}} F(x)$$

נרצה לבטא שוב את הבעיה לעיל, ע"י שני משתנים x,τ . נוכל באופן דומה למה שראינו בבעיה הראשונה לרשום:

$$\min_{\tau,x} \tau$$
$$S.T\ (Subject\ To):\ \tau \ge |f_i(x)|\quad ,\qquad i=1,\dots,m$$

אך זה ניסוח לא טוב כיוון שהאילוץ שלנו הוא לא "חלק" כיוון שפונקצית ערך מוחלט לא גזירה בראשית.

נוכל לנסח את הבעיה בצורה קצת שונה אם נבחין כי:

$$|f_i(x)| = \max\{-f_i(x), f_i(x)\}$$

ואז נקבל את בעית האופטימיזציה עם אילוצים הבאה:

$$\min_{\tau,x} \tau$$
$$\begin{aligned} S.T\ (Subject\ To):\quad & \tau \ge f_i(x) \quad ,\qquad i=1,\dots,m \\ & \tau \ge -f_i(x) \quad ,\qquad i=1,\dots,m \end{aligned}$$

או באופן קומפקטי:

$$\min_{\tau,x} \tau$$
$$S.T\ (Subject\ To): \qquad -\tau \le f_i(x) \le \tau\ , \qquad i=1,...,m$$

קירוב צ'בישב (Chebyshev Approximation):

נניח ובידינו פונקציה של משתנה יחיד $h(t)$ ונרצה לקרב את הפונקציה באמצעות קבוצה של פונקציות נתונות.

$$h(t) \approx \sum_i \alpha_i \varphi_i(t)$$

כאשר קבוצת הפונקציות $\{\varphi_i(t)\}_i$ היא קבוצה כלשהי של פונקציות (ייתכן בסיס פוריה או כל קבוצה אחרת). אנחנו נתייחס כאילו קבוצה הפונקציות $\{\varphi_i(t)\}_i$ היא סופית. כלומר:

$$h(t) \approx \sum_{i=1}^{n} \alpha_i \varphi_i(t)$$

ואנחנו נרצה למזער את הביטוי:

$$\left| h(t) - \sum_{i=1}^{n} \alpha_i \varphi_i(t) \right|$$

נגדיר: $\psi(t) \triangleq \sum_{i=1}^{n} \alpha_i \varphi_i(t)$ ולכן אנחנו רוצים למזער את הביטוי:

$$\left| h(t) - \psi(t) \right|$$

אך באיזה מובן אנחנו רוצים למזער את הביטוי? אנחנו נרצה שבאינטרבל כלשהו T נקבל הפרש מקסימלי שהוא המינימלי שניתן לקבל בין $h(t)$ לבין $\psi(t)$ כאשר הקירוב שלנו $\psi(t)$ הוא גם פונקציה של הפרמטרים α_i שאנחנו בוחרים. באופן פורמלי אנחנו רוצים לפתור את הבעיה:

$$\min_{\alpha \in \mathbb{R}^n} \left(\max_{t \in T} \left| h(t) - \sum_{i=1}^{n} \alpha_i \varphi_i(t) \right| \right)$$

נבחין כי בבעיה זו, הפרמטרים שעליהם אנחנו עושים מינימזציה הם רציפים, ואנחנו נעבור כעת למישור בדיד, כלומר המשתנה t יוכל לקבל ערכים בדידים בלבד: $t = \{1, 2, ..., T\}$. ולכן בעית האופטימיזציה הופכת להיות:

$$\min_{\alpha \in \mathbb{R}^n} \max_{t=\{1,..,T\}} \left| h_t - \sum_{i=1}^{n} \alpha_i \varphi_{t,i} \right|$$

כאשר φ היא למעשה מטריצה, שהשורה ה- i מכילה את הערכים של הפונקציה φ_i בנקודות הבדידות t . כעת נבחין כי הבעיה הזו מתקרבת בניסוח שלה לבעיה שראינו בדוגמאות של Minmax. אם נסמן: $f_t(\bar{\alpha}) \triangleq h_t - \sum_{i=1}^{n} \alpha_i \varphi_{t,i}$, אז נבחין כי $f_t(\bar{\alpha})$ היא פונקציה לינארית ביחס ל- $\bar{\alpha}$. בדוגמאות של בעית ה-Minmax הגענו בסופו של דבר לניסוח הבעיה:

$$\min_{\tau,x} \tau \quad ,\tau \in \mathbb{R}$$
$$S.T\ (Subject\ To): \qquad -\tau \le f_i(x) \le \tau\ , \qquad i=1,\dots,m$$

ולכן גם במקרה זה נוכל לנסח את הבעיה באופן דומה:

$$\min_{\tau,x} \tau \quad ,\tau \in \mathbb{R}$$
$$S.T\ (Subject\ To): \qquad -\tau \le f_t(\bar{\alpha}) \le \tau \qquad , \qquad t=1,\dots,T$$

או באופן מפורש בעית האופטימיזציה שאנחנו צריכים לפתור היא:

$$\min_{\tau,x} \tau \quad ,\tau \in \mathbb{R}$$
$$S.T\ (Subject\ To): \qquad -\tau \le h_t - \sum_{i=1}^{n} \alpha_i \varphi_{t,i} \le \tau \quad , \qquad t=1,\dots,T$$

וזו בעית תכנות לינארי (שזה מקרה פרטי של תכנות קוני) וזה ניסוח בעית האופטימיזציה אשר הפתרון שלה הוא למעשה קירוב צ'בישב במישור הבדיד.

נוכל לרשום את כל מה שרשמנו על קירוב צ'בישב, בכתיב ווקטורי באופן הבא:

$$\bar{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_T \end{pmatrix} \quad , \quad \vartheta = \begin{pmatrix} \varphi_1(1) & \cdots & \varphi_1(T) \\ \vdots & \ddots & \vdots \\ \varphi_n(1) & \cdots & \varphi_n(T) \end{pmatrix} \quad , \quad \bar{\tau} = \begin{pmatrix} \tau \\ \vdots \\ \tau \end{pmatrix}$$

ונקבל את הבעיה:

$$\min_{\tau,x} \tau \quad ,\tau \in \mathbb{R}$$
$$S.T\ (Subject\ To): \qquad -\bar{\tau} \le \bar{h} - \vartheta\bar{\alpha} \le \bar{\tau}$$

אפשר לנסח את הבעיה לעיל בצורה כללית יותר, בצורה של תכנות לינארי כללי:

$$\min_{\bar{x}} \langle \bar{c}, \bar{x} \rangle = \min_{\bar{x}} \bar{c}^T \bar{x}$$
$$S.T\ (Subject\ To): \qquad A\bar{x} \ge \bar{b}$$

כאשר אפשר למצוא את המטריצה A ואת הווקטורים $\bar{c}, \bar{b}$ ולקבל בדיוק את בעית האופטימיזציה שניסחנו עבור קירוב צ'בישב.

<u>קירוב צ'בישב מרוכב:</u>

לעיתים מתעורר צורך לקרב פונקציה מרוכבת כלשהי (אשר הארגומנט שלה הוא ממשי) באמצעות קירוב צ'בישב (כלומר ע"י קבוצה סופית של פונקציות כלשהן אך הפעם הן מרוכבות).נתבונן בבעיה הבאה:

$$\min_{\bar{\alpha}} \max_{t=1,\dots,T} \left| h[t] - \sum_{i=1}^{n} \alpha_i \varphi_{ti} \right| \qquad ,t \in \mathbb{R}, \quad h[t], \varphi_{ti}, \alpha_i \in \mathbb{C}$$

ונסמן למען הנוחות את הפונקציה המרוכבת הבאה: $z_t\left(\bar{\alpha}\right)\triangleq h[t]-\sum_{i=1}^{n}\alpha_i\varphi_{ti}$ ולכן הבעיה היא:

$$\min_{\bar{\alpha}}\max_{t=1,\ldots,T}\left|z_t\left(\bar{\alpha}\right)\right|$$

כעת, בכדי להימנע מהתעסקות עם מספרים מרוכבים, נתאר כל ערך של הפונקציה המרוכבת $z_t\left(\bar{\alpha}\right)$ כווקטור בעל שתי קואורדינטות, רכיב ממשי ומדומה. נגדיר:

$$\bar{u}_t\left(\bar{\alpha}\right)=\begin{pmatrix}\operatorname{Re}\left[z_t\left(\bar{\alpha}\right)\right]\\ \operatorname{Im}\left[z_t\left(\bar{\alpha}\right)\right]\end{pmatrix}$$

ולכן הבעיה בניסוח הנוכחי היא:

$$\min_{\bar{\alpha}}\max_{t=1,\ldots,T}\left|z_t\left(\bar{\alpha}\right)\right|=\min_{\bar{\alpha}}\max_{t=1,\ldots,T}\left\|\bar{u}_t\left(\bar{\alpha}\right)\right\|_2$$

או באופן שקול:

$$\min_{\bar{\alpha},\tau}\tau$$
$$S.T\ \left(Subject\,To\right):\ \tau\geq\left\|\bar{u}_t\left(\bar{\alpha}\right)\right\|_2\qquad,\qquad t=1,\ldots,T$$

וזו בעיה תכנות קוני ריבועית, כאשר האילוץ הוא אילוץ קוני מסדר שני כי אפשר לחשוב על אי-השיוויון באילוץ כעל הביטוי הבא:

$$\begin{pmatrix}\bar{u}_t\left(\bar{\alpha}\right)\\ \tau\end{pmatrix}=\begin{pmatrix}\operatorname{Re}\left[z_t\left(\bar{\alpha}\right)\right]\\ \operatorname{Im}\left[z_t\left(\bar{\alpha}\right)\right]\\ \tau\end{pmatrix}\in L^3$$

כאשר L^3 זה הקונוס של לורנץ וזה נכון כיוון שלמדנו שבקונוס של לורנץ הרכיב האחרון (במקרה שלנו τ) צריך להיות גדול מהנורמה של כל הרכיבים האחרים (במקרה שלנו $\bar{u}_t\left(\bar{\alpha}\right)$, כאשר $\bar{u}_t\left(\bar{\alpha}\right)$ הוא למעשה שני רכיבים).

אנחנו נרצה לנסח את הבעיה האחרונה:

$$\min_{\bar{\alpha},\tau}\tau$$
$$S.T\ \left(Subject\,To\right):\ \tau\geq\left\|\bar{u}_t\left(\bar{\alpha}\right)\right\|_2\qquad,\qquad t=1,\ldots,T$$

בצורה נוחה יותר, בכך שנצטרך למזער פונקציה לינארית שתלויה בווקטור שמורכב מהמשתנים $\bar{\alpha},\tau$ כך שהווקטור הנ"ל שייך לקונוס כלשהו שנגדיר. נגדיר:

$$\bar{v}_t\triangleq\begin{pmatrix}\bar{u}_t\left(\bar{\alpha}\right)\\ \tau\end{pmatrix}\triangleq\begin{pmatrix}\operatorname{Re}\left[z_t\left(\bar{\alpha}\right)\right]\\ \operatorname{Im}\left[z_t\left(\bar{\alpha}\right)\right]\\ \tau\end{pmatrix}\in L^3$$

ולכן נוכל לומר כי מתקיים:

$$\bar{v}_t \in \mathbb{R}^{3T}, \qquad \bar{v}_t(\bar{\alpha}) = \begin{pmatrix} v_1(\bar{\alpha}) \\ v_2(\bar{\alpha}) \\ \vdots \\ v_T(\bar{\alpha}) \end{pmatrix} = \begin{pmatrix} \bar{u}_1(\bar{\alpha}) \\ \tau \\ \bar{u}_2(\bar{\alpha}) \\ \tau \\ \vdots \\ \bar{u}_T(\bar{\alpha}) \\ \tau \end{pmatrix}$$

אנחנו רוצים לקבל בעיה מהסוג הבא:

$$\begin{aligned} &\min_{\bar{\alpha},\tau} \tau \\ &S.T\ \ (Subject\,To):\ \bar{v}(\bar{\alpha}) \in K \end{aligned}$$

ולכן נגדיר את הקונוס הבא:

$$K = \underbrace{L^3 \times L^3 \times \cdots \times L^3}_{T\ \ times}$$

ובכך למעשה הפכנו את הבעיה של תכנות קוני ריבועי לבעית תכנות קוני לינארית.

כל הניתוח והניסוח של הבעיה לעיל היה תחת ההנחה כי רכיבי הווקטור $\bar{\alpha}$ הם <u>מרוכבים</u>, כלומר $\alpha_i \in \mathbb{C}$. אם ברצוננו לנסח בעית אופטימיזציה תוך שימוש <u>במספרים ממשיים</u>, אז ניתן להגדיר:

$$\bar{x} = \begin{pmatrix} \tau \\ \mathrm{Re}\left[u_1(\bar{\alpha})\right] \\ \mathrm{Im}\left[u_1(\bar{\alpha})\right] \\ \mathrm{Re}\left[u_2(\bar{\alpha})\right] \\ \mathrm{Im}\left[u_2(\bar{\alpha})\right] \\ \vdots \\ \mathrm{Re}\left[u_n(\bar{\alpha})\right] \\ \mathrm{Im}\left[u_n(\bar{\alpha})\right] \end{pmatrix}$$

וכעת אפשר לנסח את בעית האופטימיזציה בתור בעית תכנות קוני ריבועית (תוך שימוש בווקטור $\bar{x}$ שהגדרנו זה עתה):

$$\begin{aligned} &\min_{\bar{x}} \bar{c}^T \bar{x} \\ &S.T\ (Subject\ To): \qquad A\bar{x} - \bar{b} \geq_K 0 \end{aligned}$$

כאשר נשאיר לקורא למצוא את המטריצה A והווקטורים $\bar{c}, \bar{b}$ המתאימים לבעיה של קירוב צ'בישב.

הרצאה 17 – Conversion of different problems to SDP (Semi Definite Programming)

הבסיס שעליו מתבססות על השיטות להפיכת בעיות אופטימיזציה שונות לבעית אופטימיזציה מסוג SDP (Semi Definite Programming) הוא אחד והוא למעשה הלמה של Schur.

<u>למה</u>: (Schur Complement)

נניח כי מטריצה A היא מטריצת <u>בלוקים</u> (כל איבר במטריצה הוא מטריצה בפני עצמו) סימטרית מהצורה:

$$A=\begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix}\ , \qquad R\succ 0$$

אז מתקיים:

- המטריצה A היא אי-שלילית $\Leftrightarrow$ המטריצה $P-Q^TR^{-1}Q$ אי-שלילית.
- המטריצה A היא חיובית $\Leftrightarrow$ המטריצה $P-Q^TR^{-1}Q$ חיובית.

<u>הוכחה</u>: (Schur Complement)

נבדוק את החיוביות של A באמצעות ההגדרה ונזכיר כי מטריצה W היא חיובית אם לכל $\bar{x}$ מתקיים $\bar{x}^TW\bar{x}>0$.

במקרה שלנו אנחנו צריכים לדרוש:

$$\forall \bar{u},\bar{v}:\ \begin{pmatrix} \bar{u}^T & \bar{v}^T \end{pmatrix}\begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix}\begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix}\geq 0 \quad , \qquad \bar{x}\triangleq\begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix}$$

נבחין כי אם אנחנו דורשים כי אי-השיוויון יתקיים לכל $\bar{u},\bar{v}$ אז בהכרח הוא יתקיים גם עבור $\inf \bar{v}$ ולכן נוכל לומר כי אם אי-השיוויון לעיל מתקיים אז גם מתקיים:

$$\forall \bar{u}: \quad \inf_{\bar{v}}\begin{pmatrix} \bar{u}^T & \bar{v}^T \end{pmatrix}\begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix}\begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix}\geq 0$$

אם נפתח את הסוגריים ונבצע את המכפלה בצורה מפורשת נקבל פונקציה ריבועית ולכן אנחנו למעשה מתבקשים לפתור בעית אופטימיזציה לפי ווקטור $\bar{v}$, ואנחנו יודעים לנסח את התנאים ההכרחיים לפתרון בעיות מהסוג הזה, למשל שהגרדיאנט ביחס לווקטור $\bar{v}$ צריך להתאפס בנקודה שהיא הפתרון וכו'. לאחר הניסוח אנחנו נראה כי הדרישה שאי-השיוויון לעיל מתקיים הוא שקול (אמ"מ) לדרישה שיתקיים (כאשר נתון לנו כי $R\succ 0$):

$$\bar{u}^T\left(P-Q^TR^{-1}Q\right)\bar{u}\geq 0$$

ולפי ההגדרה של מטריצה אי-שלילית, המטריצה $P-Q^TR^{-1}Q$ היא אי-שלילית אמ"מ אי-השיוויון לעיל מתקיים ולכן הוכחנו את הנדרש. נכתוב את הפתרון בתורה מפורשת (פתיחת הסוגריים וכו'):

$$\begin{pmatrix} \bar{u}^T & \bar{v}^T \end{pmatrix}\begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix}\begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix}=\bar{u}^TP\bar{u}+\bar{v}^TR\bar{v}+\bar{u}^TQ^T\bar{u}+\bar{v}^TQ\bar{u}=\bar{u}^TP\bar{u}+\bar{v}^TR\bar{v}+2\bar{u}^TQ^T\bar{u}$$

כעת נמצא את המינימום לפי $\bar{v}$ ולשם כך נמצא את הגרדיאנט לפי $\bar{v}$:

$$\nabla_{\bar{v}}\left(\bar{u}^T P\bar{u}+\bar{v}^T R\bar{v}+\bar{u}^T Q^T\bar{u}+\bar{v}^T Q\bar{u}=\bar{u}^T P\bar{u}+\bar{v}^T R\bar{v}+2\bar{u}^T Q^T\bar{u}\right)=2Q\bar{u}+2R\bar{v}=0$$
$$\Rightarrow v=-R^{-1}Q$$

ולכן נציב את הנקודה $v=-R^{-1}Q$ בחזרה לביטוי $\begin{pmatrix}\bar{u}^T & \bar{v}^T\end{pmatrix}\begin{pmatrix}P & Q^T\\ Q & R\end{pmatrix}\begin{pmatrix}\bar{u}\\ \bar{v}\end{pmatrix}$ ונקבל לאחר פישוט:

$$\bar{u}^T\left(P-Q^T R^{-1}Q\right)\bar{u}\ge 0$$

וזה נכון אמ"מ $P-Q^T R^{-1}Q\succeq 0$. מש"ל.

<u>מזעור של הערך העצמי המקסימלי (Minimize Maximal Eigenvalue):</u>

בעית התכונת הראשונה שנדגים כיצד מבצעים לה המרה לבעית תכנות SDP היא תהיה בעית המזעור של הערך העצמי המקסימלי. הבעיה מנוסחת כדלקמן:

נניח כי $\bar{x}\in\mathbb{R}^n$ וכן נסמן:

$$A(\bar{x})\triangleq \mathrm{A}(\bar{x})-\bar{b}\triangleq\sum_{i=1}^{n}x_iA_i-B \qquad , \qquad A_1,A_2,...,A_n,B\in S^m$$

(כאשר S^m זו קבוצת המטריצות הסימטריות מסדר $[m\times m]$)

בנוסף נסמן את האופרטור $\lambda_{\max}(\cdot)$ להיות אופרטור שמחזיר את הערך העצמי המקסימלי של הארגומנט של האופרטור (שהוא כמובן מטריצה). בעית האופטימיזציה היא:

$$\min_{\bar{x}}\lambda_{\max}\left(A(\bar{x})\right)$$

בהרבה שימושים הנדסיים, פעמים רבות המטרה היא להקטין את הערך העצמי של המטריצה ולכן זו בעיה מאד שימושית.

כיצד ניתן לפתור זאת? נניח כי הערכים העצמים של מטריצה A הם λ_i ונמקמם בתוך מטריצה אלכסונית שכל האיברים באלכסון שלה הם למעשה הע"ע. נוכל לרשום את הבעיה $\min_{\bar{x}}\lambda_{\max}\left(A(\bar{x})\right)$ בתור בעית אופטימיזציה של סקלר כלשהו (שהוא למעשה הערך העצמי המקסימלי) ולכן הבעיה תהיה:

$$\min_{t,\bar{x}} t$$
$$S.T\ (Subject\ To):\qquad t\ge\lambda_{\max}\left(A(\bar{x})\right)$$

את האילוץ לעיל ניתן לרשום גם בצורה אחרת. נוכל לרשום את זה בתור n אילוצים:

$$t-\lambda_i\ge 0 \qquad , \qquad \forall i=1,2,...,n$$

ואם נרשום בצורה מטריצית נקבל (דורש הוכחה):

$$tI-A(\bar{x})\succeq 0$$

כלומר האילוץ הוא למעשה שהמטריצה $tI-A(\bar{x})$ תהיה אי-שלילית. ואי-השיוויון לעיל שקול (דורש הוכחה) לדרישה:

$$\begin{pmatrix} t-\lambda_1 & & 0 \\ & \ddots & \\ 0 & & t-\lambda_n \end{pmatrix} \succeq 0$$

ולכן הבעיה המקורית שלנו (מזעור הערך העצמי המקסימלי) בניסוח SDP היא:

$$\min_{t,\bar{x}} t$$
$$S.T\ (Subject\ To): \qquad tI - A(\bar{x}) \succeq 0$$

בקרוב נלמד כי ישנן שיטות יעילות לחישוב ופתרון הבעיה הזו בניסוח SDP.

<u>קירוב לינארי של מטריצה (Linear Matrix Aprox.):</u>

נושא שקשור לבעית מזעור של הע"ע המקסימלי היא בעית הקירוב הלינארי של מטריצה, ובעית הקירוב הלינארי של מטריצה היא בעיה נפוצה מאד בקרב מהנדסים בתחומים שונים. בעית הקירוב הלינארי של מטריצה נתונה היא:

ברצוננו לקרב את המטריצה A_0 באמצעות קומבינציה לינארית של מטריצות בסיס (מטריצות כלשהן שנבחר), למשל מטריצות פשוטות כלשהן עם תכונות שאנחנו מעוניינים בהן $\{A_1,\dots,A_n\}$. כלומר נקבל:

$$A_0 \approx \sum_{i=1}^{n} x_i A_i \qquad ,\forall i=1,..,n: x_i \in \mathbb{R}$$

ברצוננו לקרב את המטריצה A_0 ע"י צירוף לינארי של מטריצות הבסיס שלנו $\sum_{i=1}^{n} x_i A_i$ במובן של שגיאה מינימלית, כאשר השגיאה מוגדרת להיות:

$$Error\left(A(\bar{x})\right) \triangleq \left(\sum_{i=1}^{n} x_i A_i\right) - A_0$$

ואנחנו נרצה למזער את הנורמה של <u>מטריצת</u> השגיאה $Error\left(A(\bar{x})\right)$. ניתן להגדיר נורמות שונות עבור מטריצות ואנחנו נרצה למעשה למזער את האופרטור נורמה שפועל על המטריצה. אופרטור הנורמה של מטריצה הוא מוגדר ע"י:

$$Norm(A) \triangleq \|A\| \triangleq \max_{\|\bar{u}\|_2 \le 1} \|A\bar{u}\|_2$$

כמו כן, ידוע מאלגברה לינארית כי הנורמה של מטריצה מקיימת:

$$\|A\| = \sigma_{\max}(A) = \sqrt{\lambda_{\max}\left(A^T A\right)}$$

כאשר $\sigma_{\max}(A)$ זה הערך הסינגולרי המקסימלי של המטריצה A. בסך הכול, בעית הקירוב הלינארי של מטריצה A_0 כעת ניתנת לרישום ע"י הבעיה:

$$\min_{\bar{x}} \lambda_{\max}\left(\left(A(\bar{x})\right)^T A(\bar{x})\right)$$

אך את הבעיה $\min_{\bar{x}} \lambda_{\max}\left(\left(A(\bar{x})\right)^T A(\bar{x})\right)$ אנחנו יודעים לפתור, כי זו בעיה ששקולה לבעיה:

$$\begin{aligned} &\min_{t,\bar{x}} t \quad , t \in \mathbb{R} \\ &S.T\ (Subject\ To): \quad t^2 I - \left(A(\bar{x})\right)^T A(\bar{x}) \succeq 0 \\ &\qquad\qquad\qquad\qquad\quad t \geq 0 \end{aligned}$$

כאשר האילוץ t^2 הוא לשם נוחות. נבחין כי הבעיה שרשמנו לעיל היא <u>אינה</u> בעית SDP רגילה כי האילוץ הוא לא לינארי (כמו שהיינו רוצים) אך ניתן לרשום בעיה זו באופן שונה (תוך שימוש בלמה Schur Complement):

$$\begin{aligned} &\min_{t,\bar{x}} t \quad , t \in \mathbb{R} \\ &S.T\ (Subject\ To): \quad \begin{pmatrix} tI & A^T(\bar{x}) \\ A(\bar{x}) & tI \end{pmatrix} \succeq 0 \\ &\qquad\qquad\qquad\qquad\quad t > 0 \end{aligned}$$

(כאשר $\begin{pmatrix} tI & A^T(\bar{x}) \\ A(\bar{x}) & tI \end{pmatrix}$ זו מטריצת בלוקים)

לסיכום, מצאנו איך לבטא את בעית קירוב מטריצה כלשהי ע"י צירוף לינארי באמצעות בעית SDP.

<u>תכנות לינארי באמצעות תכנות SDP (LP to SDP):</u>

בחלק זה נלמד איך לבטא מספר בעיות אופטימיזציה לינאריות ידועות באמצעות בעיות SDP שקולות. נושא זה הוא בעיקר חשוב לידע התיאורטי שכן פתרון בעיות אופטימיזציה לינאריות הוא יעיל יותר מפתרון בעיות SDP שקולות אך חשוב לראות את הקשר בין שני סוגי הבעיות. למעשה בעית אופטימיזציה לינארית, היא מקרה פרטי של בעית SDP.

נניח ונתונה לנו בעית האופטימיזציה הלינארית הבאה:

$$\begin{aligned} &\min_{\bar{x}\in\mathbb{R}^n} \bar{c}^T \bar{x} \qquad\qquad\qquad\qquad , \bar{c} \in \mathbb{R}^n \\ &S.T\ (Subject\ To): \quad \bar{a}_i^T \bar{x} - \bar{b}_i \geq 0 \quad , i = 1,\dots,m, \quad \bar{a}_i, \bar{b}_i \in \mathbb{R}^n \end{aligned}$$

ובעית ה-SDP השקולה שלה היא:

$$\begin{aligned} &\min_{\bar{x}\in\mathbb{R}^n} \bar{c}^T \bar{x} \qquad\qquad , \bar{c} \in \mathbb{R}^n \\ &S.T\ (Subject\ To): \quad \begin{pmatrix} \bar{a}_1^T \bar{x} - \bar{b}_1 & & 0 \\ & \ddots & \\ 0 & & \bar{a}_n^T \bar{x} - \bar{b}_n \end{pmatrix} \succeq 0 \end{aligned}$$

(קל לראות שזה שקול כיוון שהמטריצה $\begin{pmatrix} \bar{a}_1^T\bar{x}-\bar{b}_1 & & 0 \\ & \ddots & \\ 0 & & \bar{a}_n^T\bar{x}-\bar{b}_n \end{pmatrix}$ תהיה אי-שלילית אמ"מ כל הערכים העצמיים שלה הם חיובים, אך אם המטריצה היא אלכסונית (כמו במקרה זה) אז הערכים העצמיים נמצאים על האלכסון הראשי שלה ולכן למעשה אנחנו דורשים שכל אחד מהערכים העצמיים הוא אי-שלילי ונבחין כי כל אחד מהערכים העצמיים הוא למעשה האילוץ בבעיה המקורית ולכן זו אכן בעיה שקולה)

<u>תכנות בעיות תכנות קוני ריבעיות (CQP) באמצעות תכנות SDP (CQP via SDP):</u>

בעיות תכנות קוני ריבועיות, בדומה לבעיות תכנות לינארי, יכולות להיות מבוטאות ע"י בעיות אופטימיזציה מסוג SDP. נתבונן בבעית CQP כללית:

$$\min_{\bar{x}\in\mathbb{R}^n} \bar{c}^T\bar{x} \qquad , \bar{c}\in\mathbb{R}^n$$

$$S.T\ \ (Subject\ To): \qquad \Big\| \underbrace{A\bar{x}-\bar{b}}_{\bar{u}(\bar{x})} \Big\|_2 \le \underbrace{\bar{e}^T\bar{x}-d}_{v(\bar{x})\in\mathbb{R}} \quad , \begin{pmatrix} \bar{u}(\bar{x}) \\ v(\bar{x}) \end{pmatrix} \in L^n \quad , v\in\mathbb{R}$$

$$, d\in\mathbb{R}, \ \bar{e}\in\mathbb{R}^n$$

ניתנת לתיאור ע"י בעית SDP שקולה באופן הבא (תוך שימוש בלמה Schur Complement):

$$\min_{\bar{x}} \bar{c}^T\bar{x} \qquad , \bar{x}\in\mathbb{R}^n$$

$$S.T\ \ (Subject\ To): \qquad \begin{pmatrix} (v) & (u_1 \ \cdots \ u_n) \\ \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} & \begin{pmatrix} v & & 0 \\ & \ddots & \\ 0 & & v \end{pmatrix} \end{pmatrix} \succeq 0 \qquad , \qquad v>0$$

נבחין כי תוך שימוש בלמה Schur Complement ניתן לכתוב את האילוץ לעיל בתור האילוץ:

$$\bar{v}-\bar{u}\begin{pmatrix} v & & 0 \\ & \ddots & \\ 0 & & v \end{pmatrix}\bar{u} \ge 0 \qquad , \qquad \bar{v}=\begin{pmatrix} v \\ \vdots \\ v \end{pmatrix} \ , v\in\mathbb{R}$$

וזה שקול (לאחר מניפולציות) לאי-השיוויון:

$$v-v^{-1}\bar{u}^T\bar{u} \ge 0$$

ואם נכפול <u>בסקלר</u> v נקבל:

$$v^2-\bar{u}^T\bar{u} \ge 0$$

וזה אכן שקול לאילוץ המקורי.

<u>שיטת המחסום עבור בעיות תכנות קוני (Barrier method for QP):</u>

בחלק זה נלמד איך לפתור בעיות תכנות קוני כלליות באמצעות שיטת המחסום. כבר למדנו את שיטת המחסום בבעיות תכנות לא לינארי והרעיון הכללי זהה לרעיון שראינו בעבר. נניח ויש לנו בעית תכנות קוני מהצורה הבאה:

$$\begin{aligned}&\min_{\bar{x}\in\mathbb{R}^n} \bar{c}^T\bar{x} \qquad , \bar{c}\in\mathbb{R}^n\\ &S.T\ (Subject\ To): \qquad A(\bar{x}) \triangleq \mathrm{A}(\bar{x}) - B \geq_K 0\end{aligned}$$

(כאשר נזכיר: $\mathrm{A}(\bar{x}) \triangleq \sum_{i=1}^{n} A_i\bar{x}_i$ וכן $\{A_1, A_2, ..., A_n\}$ מטריצות)

בנוסף נניח כי קיימת נקודה $\bar{x}$ בתוך תחום האילוץ (Feasible Area) ולא רק על גבולות התחום (כלומר ניתן למצוא סביבה שלמה של הנקודה שכל הסביבה נמצאת גם היא בתוך התחום) אשר מקיימת (אי-שיוויון חזק):

$$\mathrm{A}(\bar{x}) - B >_K 0$$

ולכן נוכל להגדיר פונקצית מחסום אשר יהיו לה ערכים קטנים (מתונים) יחסית בתוך הקונוס ובאיזור גבולות הקונוס ערכי הפונקציה ילכו וישאפו לאינסוף. נגדיר את פונקצית המחסום בצורה הבאה:

$$\begin{aligned}&Barrier\ \ function: \qquad \vartheta : K \to \mathbb{R}\\ &\qquad\qquad\qquad\qquad\qquad \lim_{\substack{u\to\partial K\\ u\in K}} \vartheta(u) = \infty\end{aligned}$$

כאשר ∂K זה השפה של הקונוס K.

<u>פונקצית המחסום המצרפית (Barrier aggregate):</u>

נגדיר את פונקצית המחסום המצרפית, בדומה לאיך שהגדרנו את פונקצית העונשין המצרפית (סכום פונקצית המטרה ועוד סכום פונקצית העונשין), בצורה הבאה:

$$F_\mu(\bar{x}) = \bar{c}^T\bar{x} + \frac{1}{\mu}\vartheta\big(\mathrm{A}(\bar{x}) - B\big) \qquad , \qquad \mu\in\mathbb{R}$$

נבחין שעבור פרמטר μ גדול, נקבל כי עבור איבר $\mathrm{A}(\bar{x}) - B$ בתוך התחום ממש הביטוי $\frac{1}{\mu}\vartheta\big(\mathrm{A}(\bar{x}) - B\big)$ יהיה מאד קטן, אך בגבולות הקונוס אנחנו עדיין נקבל כי הביטוי שואף לאינסוף ולכן ניתן לרשום:

$$F_\infty(\bar{x}) = \begin{cases}\bar{c}^T\bar{x} & ,\mathrm{A}(\bar{x}) - B >_K 0\\ \infty & ,\mathrm{A}(\bar{x}) - B \leq_K 0\end{cases}$$

ברור כי אם נמצא את המינימום של פונקציה המחסום המצרפית אז זה יהיה גם הפתרון של הבעיה:

$$\left.\begin{aligned}&\min_{\bar{x}\in\mathbb{R}^n} F_\infty(\bar{x}) = \min_{\bar{x}\in\mathbb{R}^n} \bar{c}^T\bar{x} \ \ , \bar{c}\in\mathbb{R}^n\\ &S.T\ (Subject\ To): \qquad \mathrm{A}(\bar{x}) - B \geq_K 0\end{aligned}\right\} \Rightarrow \qquad \min_{\bar{x}} F_\infty(\bar{x}) = f^*$$

והאלגוריתם הוא להתחיל לפתור את בעית האופטימיזציה עם פרמטר μ קטן יחסית, ובכל איטרציה להכפיל אותו פי כמה (למשל פי 2 עד 10) עד אשר הוא גדל להיות בסדר גודל של $\sim 10^5$.

<u>דוגמאות של פונקציות מחסום מוכרות:</u>

קונוס	פונקצית מחסום $\vartheta(\bar{u})$
$\mathbb{R}_+^m$	$-\sum_{i=1}^{m}\log(u_i)$
L^m (קונוס לורנץ)	$-\log\left(u_m^2-\sum_{i=1}^{m-1}u_i^2\right)$
S_+^m (קונוס המטריצות הסימטריות האי-שליליות)	$-\log(\det(U))$

<u>פונקצית המחסום</u>: $-\log(\det(U))$

נניח ויש לנו מטריצה A וניתן לבטא אותה באמצעות מטריצה $S\Lambda S^{-1}$ כאשר S מטריצה שעמודותיה הם הווקטורים העצמיים של מטריצה A וכן Λ זו מטריצה אלכסונית שעל האלכסון שלה נמצאים כל הערכים העצמיים של מטריצה A אז ניתן לכתוב:

$$\begin{aligned}\det(A)=\det\left(S^{-1}\Lambda S\right)=\det\left(S^{-1}\right)\det(\Lambda)\det(S)=\\ =(\det(S))^{-1}\det(\Lambda)\det(S)=\det(\Lambda)(\det(S))^{-1}\det(S)=\\ =\det(\Lambda)=\prod_i\lambda_i\end{aligned}$$

ולכן נקבל:

$$-\log(\det(U))=-\log\left(\sum_{i=1}^{m}\lambda_i\right)=-\sum_{i=1}^{m}\log(\lambda_i)$$

<u>פונקציה מטריצית (Matrix functions): (זו תזכורת, שכן כבר ראינו את הנושא בהרצאות הראשונות)</u>

כאשר רוצים לפתור בעית אופטימיזציה, למשל באמצעות שיטת המחסום, אנחנו למעשה צריכים לחשב את הגרדיאנט של פונקציה המחסום ולעיתים פונקצית המחסום מוגדרת ע"י מטריצה ולכן אנחנו צריכים להיזכר באנליזה מטריצית: כיצד גוזרים מטריצה וכיצד מפעילים אופרטורים שונים על מטריצה.

נניח ויש לנו פונקציה של משתנה יחיד (ממשי) בעל פיתוח פולינומי (למשל טיילור):

$$\varphi(t)=\sum_{i=0}^{\infty}c_it^i$$

אז נגדיר פונקציה מטריצית באופן הבא (גם ע"י פיתוח פולינומי):

$$\varphi(A)=\sum_{i=0}^{\infty}c_iA^i$$

כאשר A^i אומר להכפיל את המטריצה A בעצמה, i פעמים.

נבחין כי אם ניתן לבטא את המטריצה A באמצעות פירוק למטריצת הערכים העצמיים, $A=S^{-1}\Lambda S$ (שהיא מטריצה אלכסונית) אז החישוב של הפונקציה $\varphi(A)=\sum_{i=0}^{\infty}c_iA^i$ שהגדרנו, נהיה פשוט הרבה יותר, כי נזכור כי להכפיל מטריצות אלכסוניות אחת בשניה זה לכפול את הערכים על האלכסון הראשי, אחד בשני.

<u>פירוק לערכים עצמים / מציאת הערכים העצמיים של מטריצה (Eigenvalue decomposition):</u>

נזכיר תחילה כי ערך עצמי λ_i של מטריצה A הוא ערך אשר עבורו קיים ווקטור $\bar{s}_i\in\mathbb{R}^n$ כך שמתקיים:

$$A\bar{s}_i=\lambda_i\bar{s}_i$$

נניח כי למטריצה A הריבועית בגודל $[m\times m]$ יש m ערכים עצמיים ווקטורים עצמיים (ולפי אלגברה לינארית אנחנו יודעים כי גם כל הווקטורים העצמיים הם בלתי תלויים לינארית אחד בשני). נגדיר את המטריצה S להיות מטריצת עמודות, שבה כל עמודה היא ווקטור עצמי של מטריצה A. לכן נוכל לרשום (דורש הוכחה):

$$AS=S\Lambda$$

כיוון שכל הווקטורים העצמיים של A הם בלתי תלויים, כלומר כל העמודות של מטריצה S הן בלתי תלויות, אז זה אומר כי מטריצה S היא מדרגה מלאה ולכן קיימת לה הופכית. ולכן נכפול את השיוויון האחרון מימין במטריצה S^{-1} ונקבל:

$$A=S\Lambda S^{-1}$$

ניזכר כי הגדרנו $\varphi(A)=\sum_{i=0}^{\infty}c_iA^i$ ולכן אנחנו צריכים לחשב את הכפולות של מטריצה A בעצמה. נבחין כי מתקיים:

$$A^2=\left(S\Lambda S^{-1}\right)^2=S\Lambda S^{-1}S\Lambda S^{-1}=S\Lambda\left(S^{-1}S\right)\Lambda S^{-1}=S\Lambda^2S^{-1}=S\begin{pmatrix}\lambda_1^2 & & 0\\ & \ddots & \\ 0 & & \lambda_n^2\end{pmatrix}S^{-1}$$

וניתן לראות באופן דומה (ע"י אינדוקציה) כי מתקיים:

$$A^i=S\Lambda^iS^{-1}=S\begin{pmatrix}\lambda_1^i & & 0\\ & \ddots & \\ 0 & & \lambda_n^i\end{pmatrix}S^{-1}$$

ולכן נקבל:

$$\varphi(A)=\sum_{i=0}^{\infty}c_iA^i=\sum_{i=0}^{\infty}c_iS\Lambda^iS^{-1}=S\left(\sum_{i=0}^{\infty}c_i\Lambda^i\right)S^{-1}=S\varphi(\Lambda)S^{-1}=S\begin{pmatrix}\varphi(\lambda_1) & & 0\\ & \ddots & \\ 0 & & \varphi(\lambda_m)\end{pmatrix}S^{-1}$$

בנוסף, ניתן לראות כי אם מטריצה A היא סימטרית, מתקבל $S^{-1}=S^T$ ואז נקבל:

$$\varphi(A)=S\varphi(\Lambda)S^T=S\begin{pmatrix}\varphi(\lambda_1) & & 0\\ & \ddots & \\ 0 & & \varphi(\lambda_m)\end{pmatrix}S^T$$

תכונה נוספת של הפונקציה המטריצית שהגדרנו: $\varphi(A)=\sum_{i=0}^{\infty}c_iA^i$, מתקיים:

$$Trace(\varphi(A))=Trace(S\varphi(\Lambda)S^{-1})\underset{\substack{cyclic\ shift\\ under\ \ Trace}}{=}Trace(S^{-1}S\varphi(\Lambda))=Trace(\varphi(\Lambda))=\sum_{i=1}^{m}\varphi(\lambda_i)$$

<u>גרדיאנט של עכבה של פונקציה מטריצית:</u>

נרצה לדעת מהו הביטוי $\nabla_A Trace(\varphi(A))$ באופן מפורש כאשר $\varphi:[m\times m]\to[m\times m]$ (פונקציה <u>מטריצית</u>!). מתקיים:

$$\nabla_A Trace(\varphi(A))=(\varphi'(A))^T$$

כאשר $\varphi(t):[m\times m]\to\mathbb{R}$ היא פונקציה סקלרית.

לא נוכיח את הטענה האחרונה אך נדגים את השימושיות והנכונות של הביטוי על דוגמה כללית יחסית.

נניח פונקציה <u>סקלרית</u> $\varphi(t)=t^3$ ולכן הנגזרת שלה היא $\varphi'(t)=3t^2$. בנוסף, נסמן: $f(A)\triangleq Trace(\varphi(A))$ (ונבחין כי $f:[m\times m]\to\mathbb{R}$ כלומר היא פונקציה סקלרית!) ולכן:

$$f(A)=Trace(\varphi(A))=Trace(A\cdot A\cdot A)$$

הדיפרנציאל של $f(A)$ הוא:

$$df(A)=Trace(dA\cdot A\cdot A)+Trace(A\cdot dA\cdot A)+Trace(A\cdot A\cdot dA)$$

אך לפי תכונת הציקליות של העכבה מתקיים:

$$df(A)=Trace(dA\cdot A\cdot A)+Trace(A\cdot dA\cdot A)+Trace(A\cdot A\cdot dA)=3\cdot Trace(dA\cdot A^2)$$

בנוסף לפי ההגדרה של מכפלה פנימית בין מטריצות: $\langle A,B\rangle\triangleq Trace(A^TB)$ נקבל:

$$df(A)=3\cdot Trace(dA\cdot A^2)=\left\langle dA,(3A^2)^T\right\rangle \quad\Leftrightarrow\quad df(A)=\langle dA,\nabla f\rangle$$

ולכן מצאנו כי מתקיים:

$$\nabla f=(3A^2)^T=3(A^2)^T=3(A^2)^T=(\varphi'(A))^T$$

ולכן "הוכחנו" את השיוויון לעיל, לפחות שהוא נכון עבור הדוגמה $\varphi(t)=t^3$.

<u>חישוב הגרדיאנט של פונקצית מחסום:</u>

תוך שימוש בטענה שראינו: $\nabla_A f(A) = \nabla_A Trace(\varphi(A)) = (\varphi'(A))^T$ נרצה לחשב את הגרדיאנט של הפונקציה:

$$f(A) = \log(\det(A)) \quad , \qquad A \succ 0$$

כאשר f היא פונקציה סקלרית. מתקיים:

$$f(A) = \log(\det(A)) = \log\left(\prod_{i=1}^{m} \lambda_i\right) = \sum_{i=1}^{m} \log(\lambda_i) = Trace(\log(A))$$

ולכן הפונקציה הסקלרית שנבחר היא $\varphi(t) = \log(t)$ והנגזרת שלה היא: $\varphi'(t) = \frac{1}{t} = t^{-1}$ ולכן:

$$\nabla_A f(A) = (\varphi'(A))^T = (A^{-1})^T = (A^T)^{-1} \equiv A^{-T}$$

(ואם A סימטרית אז נזכור כי מתקיים: $A = A^T$)

לסיכום: אם פונקצית המחסום שלנו היא: $\varphi(A) = -\log(\det(A))$ אז $\nabla_A \varphi(A) = -A^{-1}$.

בתכנות SDP, אנחנו לרוב נוהגים לכתוב את הארגומנט של פונקציה המחסום בצורה: $A(\bar{x}) \triangleq \mathrm{A}(\bar{x}) - B$ ולכן אם נרצה למצוא את הגרדיאנט של $\varphi(A(\bar{x}))$ ביחס לווקטור $\bar{x}$ נרצה להשתמש בהגדרה הדיפרנציאלית:

ומתקיים: $d\varphi(A(\bar{x})) = \left\langle \nabla_A \varphi(A(\bar{x})), dA(\bar{x}) \right\rangle$

$$d(A(\bar{x})) = \left\langle \nabla_A \varphi(A(\bar{x})), dA(\bar{x}) \right\rangle = \left\langle \nabla_A (A(\bar{x})), d\mathrm{A}\bar{x} \right\rangle = \left\langle A^* (\nabla_A \varphi(A(\bar{x}))), d\bar{x} \right\rangle$$

ולכן מצאנו למעשה:

$$\nabla_{\bar{x}} \varphi(\mathrm{A}\bar{x} - B) = A^* (\nabla_A \varphi(A(\bar{x})))$$

כעת לאחר שראינו כיצד מחשבים את הגרדיאנט של פונקצית המחסום, ניתן למעשה להתחיל לפתור באמצעות האלגוריתם שאנחנו מכירים לתכנות בעיות אופטימיזציה, למשל שיטת ניוטון, למרות שעבור שיטת ניוטון צריך לחשב את ההסיאן, וזה מעבר למסגרת הקורס הזה. ניתן להשתמש גם באלגוריתם BFGS שלמדנו שלא דורש את חישוב ההסיאן.