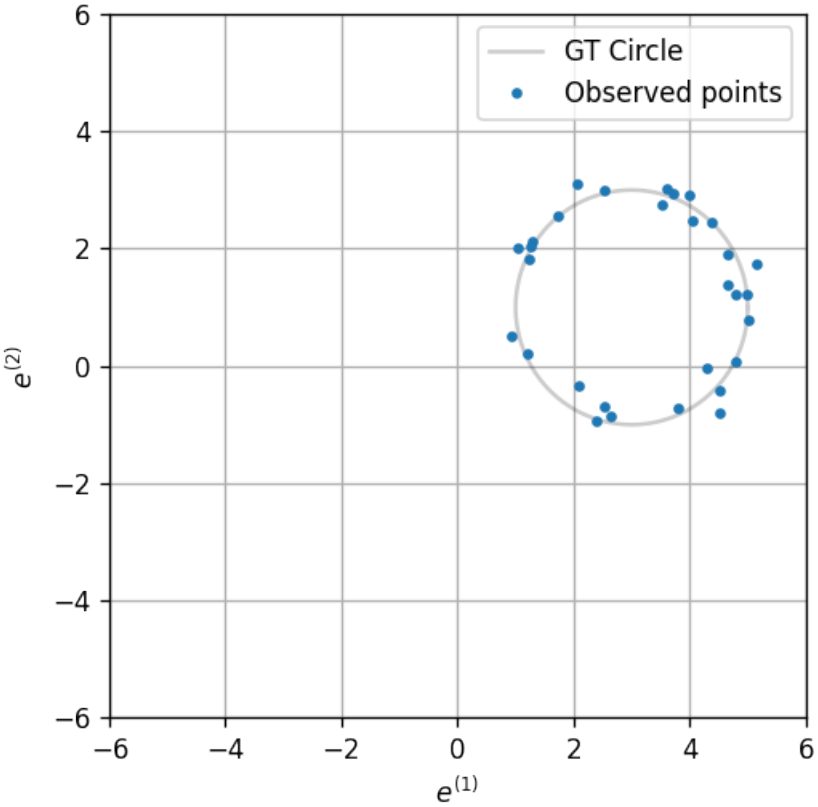
תרגול 1 - בעיות אופטימיזציה וגזירה וקטורית

הקדמה

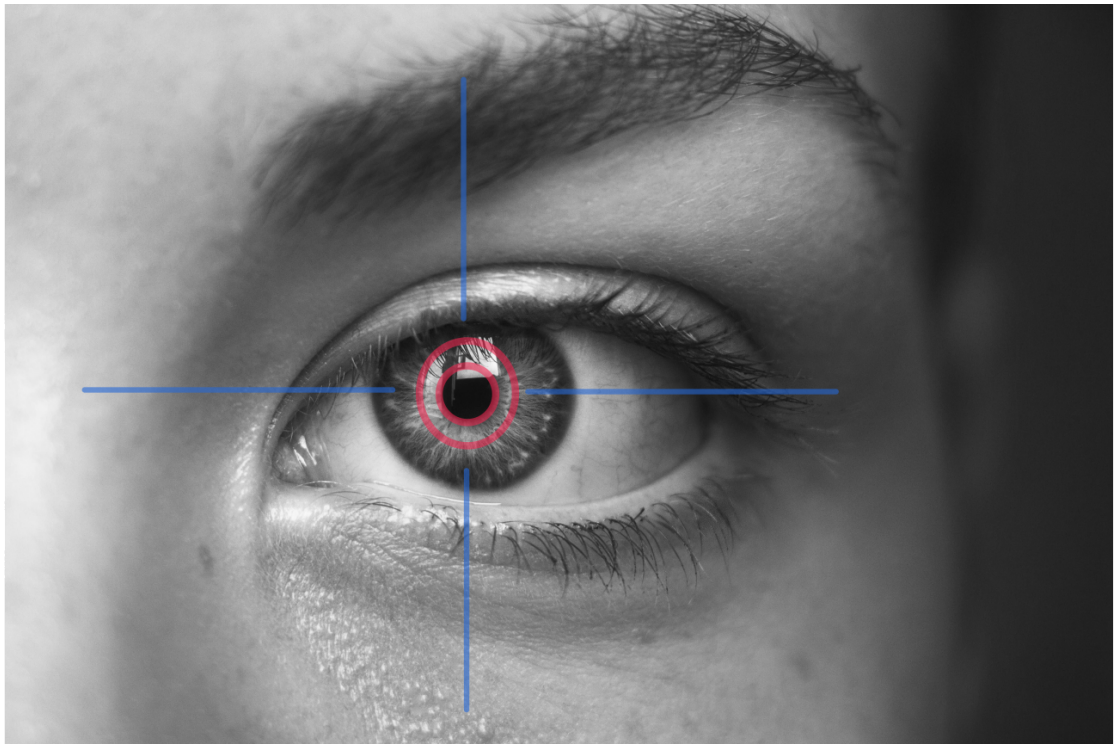
מרבית הבעיות והשיטות בתחום של מערכות לומדות עוסקות בבעיה של התאמת מודל מתמטי כך שיתאר בצורה מיטבית תופעה או תהליך מסויים על סמך אוסף נתון של תצפיות / מדידות. בהרצא הראשונה נציג את התחום באופן מפורט יותר, אך לעת אתה הגדרה זו תספק אותנו.

לצורך המחשה נסתכל על הדוגמא הבאה. נניח ונתונים לנו מספר מדידות של נקודות שיושבות על מעגל בעל מרכז ורדיוס לא ידועים. בנוסף נניח ותהליך המדידה עצמו רועש ואנו מקבלים גירסאות מורעשות של הנקודות כפי שמודגם בשרטוט הבא:



מקובל להשתמש בשם בשם **truth ground**, או בקיצור **GT**, כדי להתייחס למודל המקורי (הלא ידוע).

כעת נניח ואנו מעוניינים לשחזר את הפרמטרים של המעגל המקורי על פי הדגימות שבידינו. הבעיה של התאמת בעיה כדוגמאת זו, של מעגל לאוסף נקודות, מופיעה באלגוריתמי eye-tracking אשר מנסים לעקוב אחרי המיקום של האישון על מנת להבין מהו הכיוון שאליו אדם מביט.



נציין שהבעיה הזו לא מאד מייצגת ונחשבת ליחסית פשוטה בהשוואה לבעיות הטיפוסיות שאותם מנסים לפתור בתחום של מערכות לומדות , אך עם זאת היא תשמש כדוגמא טובה לעקרונות שבהם נעסוק בתרגול הנוכחי.

פערים מתמטיים

לפני שנוכל לצלול לחומר העיקרי של הקורס עלינו להתחיל בהשלמה ורענון של הבסיס המתימטי אשר ישמש אותנו לתורך ניסוח הפתרון של הבעיות בהם נעסוק בקורס. ספציפית אנו נעשה שימוש ב:

- אלגברה לינארית: על מנת לתאר את המידע שהמודלים שאיתם נעבוד.
- הסתברות: על מנת לתאר את האופן שבו נוצרים המדידות ובכדי לתאר את מידת הסבירות שבה מודל מסווים מתאים לתצפיות.
- תורת האופטימיזציה: על מנת למצוא את המודלים אשר מתאימים באופן מיטבי לדגימות שבידינו.

בתרגול הנוכחי נתעסק בבעיות אופטימיזציה סקלריות ווקטוריות ובתרגול הבא נוסיף את לכך את ההיבט ההיסתברותי.

נוטציות

בקרוס זה נצמד לנוטציות המתמטיות המופיעות בספר Deep Learning (מאת Goodfellow, I. & Bengio Y. Courville). ניתן למצוא את הפירוט המלא בקישור הבא. (למתקדמים: ניתן למצוא את פקודות ה$LaTeX$ הרלוונטיות כאן)

אלגברה לינארית

בהתאם להגדרות אלו אנו נשתמש בסימונים הבאים בהקשר של אלגברה לינארית:

- x - אותיות סטנדרטיות (italic lower case) לועזיות או יווניות - סקלרים.

- x - אותיות מודגשות - וקטורי **עמודה**
- $x^\top$ - וקטורי **שורה**
- x_i - האיבר הi בוקטור x.
- $\langle x, y\rangle (= x^\top y = \sum_i x_i y_i)$ - המכפלה הוקטורית הסטנדרטית בין x ל y.
- $\|x\|_2 (= \sqrt{\langle x, x\rangle})$- הנורמה הסטנדרטית (נורמת $l2$(של הוקטור x.
- $\|x\|_l (= \sqrt[l]{\sum_i x_i^l})$ - נורמת l של x

- A - אותיות לועזיות גדולות מודגשות (bold capittal) - מטריצה
- $A^\top$ - המטריצה Transposed A (המטריצה המשוחלפת).
- $A_{i,j}$ - האיבר הj שורה הi של A.
- $A_{i,:}$ - השורה הi של A.
- $A_{:,i}$ - העמודה הi של A.

Sets (קבוצות)

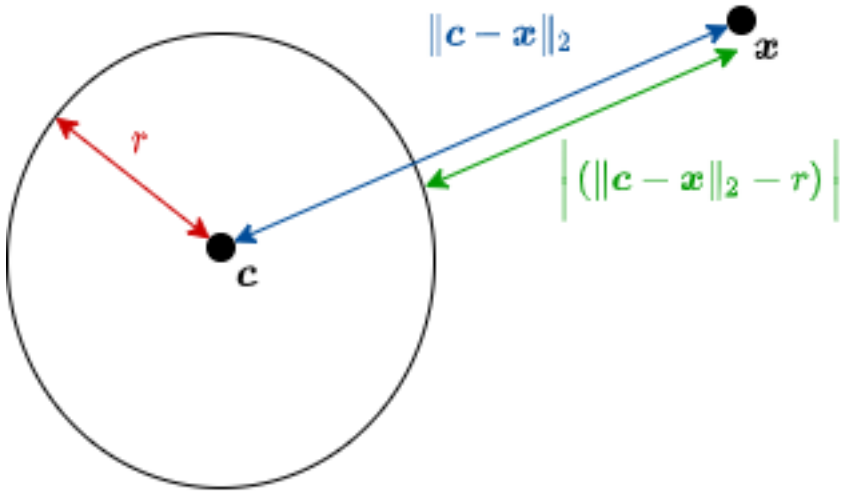
את סדרת התצפיות אנו נסמן כקבוצה (או סדרה) בעזרת הנוטציות הבאות:

- $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ - סדרה של n וקטורים.

בעיית האופטימיזציה

נחזור לדוגמא של התאמת המעגל. דרך אחת לגשת לבעיה מסוג זה הינה לחפש מבין כל המעגלים האפשריים את המעגל אשר המרחק הריבועי הממוצע (MSE - squered mean error) של הדגימות ממנו היא המינימאלית (בתרגול הבא אנו ניתן הצדקה מתימטית לשימוש במדד שגיאה זה). נרשום זאת באופן מתמטי.

נרשום תחילה את המרחק של נקודה בודדת x ממעגל בעל מרכז ב c ורדיוס r. מרחק זה שווה להפרש בין המרחק בין x ל c והרדיוס, כפי שמופיע בשרטוט הבא:



נסמן את הריבוע של מרחק זה ב e:

$$e = (\|x - c\|_2 - r)^2$$

אם כן בעבור n נקודות, $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$, המרחק הריבועי הממוצע (MSE) של הנקודות מהמעגל הינו:

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(\|x^{(i)} - c\|_2 - r)^2$$

מצאנו אם כן למעשה מדד טיב אשר מאפשר לנו לתת "ציון" לכל מעגל כתלות ב c ו r. נסמן את הפונקציה הזו כ:

$$f(c,r) = \frac{1}{n}\sum_{i=1}^{n}(\|x^{(i)} - c\|_2 - r)^2$$

כעת, אם ברצונינו למצוא עלינו למצוא את המעגל ה**אופטימאלי**, עלינו למצוא את הפרמטרים c ו r אשר מניבים את הערך הנמוך ביותר של f. מקובל לסמן את הפרמטרים האופטימאליים בעזרת $*$ באופן הבא: c^* וr^*.

בעיות מסוג זה הם בדיוק סוגי הבעיות שתורת האופטימיזציה באה לפתור. לבנתיים נשים בצד את הפונקציה הזו ונחזור אליה לקראת סוף התרגול.

המקרה הפשוט

בעיות אופטימיזציה עוסקות במציאת הארגומנט θ שבעבורו פונקציה נתונה $f(\theta)$ מחזירה את הערך המינימאלי או המקסימאלי שלה. לרוב מקובל לנסח בעיות אופטימיזציה כבעיות **minimization** (מזעור), כאשר ניתן כמובן לרשום כל בעיית **maximization** כבעיית minimization של $\tilde{f}(\theta) = -f(\theta)$. כמו כן, את פונקציה $f(\theta)$ שאותה מנסים למזער (למקסם) נהוג לכנות ה**objective** או פונקציית המטרה. באופן פורמלי, בעיות אופטימיזציה נרשמות באופן הבא:

$$\theta^* = \arg\min_{\theta} \quad f(\theta)$$

פונקציות מרובות משתנים

בעיות אופטימיזציה כמובן לא מוגבלות רק לפונקציות של משתנה יחיד וכמו בדוגמא של המעגל, ניתן להסתכל גם על הבעיה של מציאת סט הערכים האופטימאליים $\theta_1^*, \theta_2^*, \dots, \theta_d^*$ שממזערים פונקציה של מספר משתנים $f(\theta_1, \theta_2, \dots, \theta_d)$:

$$\theta_1^*, \theta_2^*, \dots, \theta_d^* = \arg\min_{\theta_1, \theta_2, \dots, \theta_d} \quad f(\theta_1, \theta_2, \dots, \theta_d)$$

במקרים רבים נוח יותר לאגד את כל הארגומנטים של f לוקטור אחד $\theta = (\theta_1, \theta_2, \dots, \theta_d)^\top$ ולרשום את בעיית האופטימזציה כ:

$$\theta^* = \arg\min_{\theta} \quad f(\theta)$$

שיטות לפתרון בעיות אופטימיזציה

בעיות אופטימיזציה הם למעשה חלק מהתהליך של חקירת פונקציה אשר נלמד בחדו"א 1 ובתיכון, שם מצאנו נקודות הקיצון על ידי גזירה והשוואה ל-0. שיטה זו טובה למקרים פשוטים בהם המשוואה $\nabla f(\theta) = 0$ ניתנת לפתרון באופן אנליטי, אך במקרים רבים זהו אינו המצב ויש צורך להשתמש בשיטות נומריות על מנת לפתור את הבעיה.

התחום של תורת האופטימיזציה הוא רחב והוא מציע מגוון רחב של שיטות כאלה המתאימות לפונקציות מטרה שונות. בפועל בקורס זה לא נצלול לעומקו של תחום זה, ולמעשה נפתור בעיות מסוג זה בעזרת אחת משלושת השיטות הבאות::

- **גזירה והשוואה ל0**: מתאים למקרים בהם ניתן לפתור את $\nabla f(\theta) = 0$ באופן אנליטי.
- **force Brute** (מעבר על כל האופציות): מתאים למקרים בהם הארגומנט של f לקבל רק אחד מסט ערכים סופי וקטן, לדוגמא $\theta \in \{0, 1, \dots, 10\}$.
- **descent Gradient** (שיטת הגרדיאנט): זוהי אחת משיטות האופטימיזציה הבסיסיות ביותר. בשיטה זו מחפשים מינימום לוקאלי על ידי התקדמות בכיוון ההפוך מהגרדיאנט. בתרגול זה אנחנו נדגים כיצד להשתמש בה.

בעיית אופטימזציה עם אילוצים

נשים לב שבדוגמא של התאמת המעגל, ישנו אילוץ מסויים על הפרמטרים של המעגל, רדיוס המעגל חייב להיות מספר חיובי. מגבלות מסוג זה מכונות אילוצים והם מופיעים בבעיות אופטימיזציה רבות. באופן כללי, בעיות אופטימיזציה יכולות להכיל איצלוצים משני סוגים:

- אילוצי אי-שוויון מהצורה $g(\theta) \geq 0$. לדוגמא, האילוץ שכל הערכים של θ יהיו קטנים מ1, יופיע בתור:

$$g_i(\theta) = 1 - \theta_i \geq 0 \qquad i = 1, \dots, d$$

- אילוצי שוויון מהצורה $h(\theta) = 0$. לדוגמא האילוץ שהנורמה של θ תהיה שווה ל1, יופיע בתור:

$$h(\theta) = \|\theta\|_2 - 1 = 0$$

אם כן הצורה הכללית ביותר של בעיית אופטימיזציה הינה:

$$\begin{aligned} \theta^* = & \arg\min_{\theta} \quad f(\theta) \\ & \text{subject to} \quad g_i(\theta) \leq 0, \qquad i = 1, \dots, m \\ & \qquad\qquad\quad h_j(\theta) = 0, \qquad j = 1, \dots, p \end{aligned}$$

בתרגיל הבא נראה כיצד ניתן להתמודד עם אילוצים פשוטים ובהמשך הקורס נתייחס גם למקרים בהם האילוצים נעשים מסובכים יותר.

תרגיל 1.1 - בעיית אופטימיזציה עם אילוצים

נרצה למצוא את הערך המקסימאלי של הפונקציה

$$f(\theta_1, \theta_2) = f(\theta) = e^{-(3\theta_1^2 + 3\theta_2^2 - 18\theta_1 - 24\theta_2 + 34)}$$

תחת האילוץ ש θ נמצא בתוך או על השפה של מעגל היחידה (במישור של $\theta_1 - \theta_2$.)

רשמו את בעיית האופטימיזציה ומצאו את פתרונה.

פתרון 1.1

רישום בעיית האופטימיזציה הבעיה הנתונה הינה בעיית אופטימיזציה עם אילוץ אי-שיוון אחד. נרשום אותה באופן פורמלי:

$$\begin{aligned} & \arg\min_{\theta = (\theta_1, \theta_2)^\top} \quad - e^{-(3\theta_1^2 + 3\theta_2^2 - 18\theta_1 - 24\theta_2 + 34)} \\ & \text{subject to} \quad 1 - (\theta_1^2 + \theta_2^2) \leq 0 \end{aligned}$$

החלפת פונקציית המטרה ראשית נשים לב כי ניתן לפשט את הבעיה על ידי השימוש בעובדה שe^x היא פונקציה מונוטונית עולה ולכן נוכל לבצע את ההחלפה הבאה מבלי לשנות את תוצאת בעיית האופטימיזציה:

$$\begin{aligned} & \arg\min_{\theta_1, \theta_2} \quad - e^{-(3\theta_1^2 + 3\theta_2^2 - 18\theta_1 - 24\theta_2 + 34)} \\ = & \arg\min_{\theta_1, \theta_2} \quad 3\theta_1^2 + 3\theta_2^2 - 18\theta_1 - 24\theta_2 + 34 \end{aligned}$$

באופן דומה נוכל גם להיפתר מהתוספת של הקבוע $+34$) וגם לחלק את פונקציית המטרה 3, ולקבל את בעיית האופטימיזציה הבאה:

$$= \underset{\theta_1,\theta_2}{\arg\min} \quad \theta_1^2 + \theta_2^2 - 6\theta_1 - 8\theta_2$$

טיפול באילוץ אי-השוויון במקרים כגון זה, בהם מספר אילוצי אי-השוויון קטן (במקרה זה יש אילוץ בודד) נוכל לפרק כל אילוץ שוויון לשני מקרים פשוטים יותר.

1. המקרה בו המינימום נמצא על השפה של האילוץ (בתרגיל זה, זאת אומרת שהמינימום נמצאת ממש על מעגל היחידה). במקרה זה האילוץ הופך לאילוץ שוויון.
2. המקרה זו המינימום לא נמצא על השפה של האילוץ. במקרה זה נחפש נקודות מינימום לוקאליות תוך התעלמות מהאילוץ ואחר כך נבדוק מי מהם מקיימת את האילוץ (אם בכלל יש נקודה כזו).

הפתרון יהיה הפתרון הנמוך יותר מבין השתיים. (כאשר יש יותר מאילוץ אי-שוויון יחיד צריך לפרק את כל אילוצי אי השוויון לשני מקרים. זאת אומרת, שכמות המקרים שיש לבדוק הינה 2 בחזקת מספר אילוצי אי-השוויון).

החיפוש בתוך מעגל היחידה נתחיל במציאת המינימום בחלקו הפנימי של העיגול. תחילה נתעלמהאילוץ נחפש את כל נקודות המינימום (לוקאליות או גלובליות) של הבעיה. אחר כך נפסול את אלו שלא מקיימות את האילוץ. בעיית האופטימיזציה ללא האילוץ הינה:

$$\underset{\theta_1,\theta_2}{\arg\min} \quad \theta_1^2 + \theta_2^2 - 6\theta_1 - 8\theta_2$$

בעיה זו ניתנת לפתרון בקלות על ידי גזירה והשוואה ל0:

$$\begin{cases} \frac{d}{d\theta_1}\theta_1^2 + \theta_2^2 - 6\theta_1 - 8\theta_2 = 0 \\ \frac{d}{d\theta_2}\theta_1^2 + \theta_2^2 - 6\theta_1 - 8\theta_2 = 0 \end{cases}$$
$$\Leftrightarrow (\theta_1, \theta_2) = (3, 4)$$

נקודה זו אומנם חשודה כנקודת קיצון אך היא לא מקיימת את האילוץ ולכן היא אינה יכול להיות פתרון לבעיה. נוכל להסיק אם כן שאין נקודות מינימום בתוך המעגל ולכן נקודת המינימום תהיה חייבת להימצא על השפה.

החיפוש על מעגל היחידה על השפה אילוץ האי-השוויון הופך לשוויון:

$$\underset{\theta=(\theta_1,\theta_2)^\top}{\arg\min} \quad \theta_1^2 + \theta_2^2 - 6\theta_1 - 8\theta_2$$
$$\text{subject to} \quad 1 - (\theta_1^2 + \theta_2^2) = 0$$

בתרגול הבא נלמד שיטה מסודרת לפתרון בעיות אופטימיזציה עם אילוצי שוויון כגון זו, אך לבנתיים נתאר כאן פתרון חליפי אשר משתמש בניסוח מחד של הבעיה (השיטה שמוצגת כאן לא רלוונטית להבנת החומר מופיעה פה רק לשם השלמות).

הדרך בה נימצא את המינימום על המעגל הינה על ידי החלפת הבעיה הנתונה בבעיית אופטימיזציה של משתנה יחיד ללא אילוצים. אנו נשתמש באילוץ (התנאי שהנקודה תמצא על מעגל היחידה) על מנת לבטא את θ_2 בעזרת θ_1 ורישום מחדש של פונקציית המטרה כפונקציה של θ_1 בלבד.

מתוך האילוץ נקבל ש:

$$1 - (\theta_1^2 + \theta_2^2) = 0$$
$$\Leftrightarrow \theta_2 = \pm\sqrt{1 - \theta_1^2}$$

כאשר עלינו לבדוק את שני המקרים כאשר θ_2 חיובי (החלק העליון של מעגל היחידה) וכשאר הוא שלילי (החלק התחתון). לאחר ההחלפה נקבל את שני בעיות האופטימיזציה הבאות (המקרה החיובי והשלילי):

$$\begin{aligned}
&\arg\min_{\theta_1} \quad \theta_1^2 + (1-\theta_1^2) - 6\theta_1 \pm 8\sqrt{1-\theta_1^2} \\
=&\arg\min_{\theta_1} \quad 1 - 6\theta_1 \pm 8\sqrt{1-\theta_1^2} \\
=&\arg\min_{\theta_1} \quad -3\theta_1 \pm 4\sqrt{1-\theta_1^2}
\end{aligned}$$

נפתור את את הבעיות האלה על יד גזירה והשוואה ל0:

$$\begin{aligned}
&\frac{d}{d\theta_1} - 3\theta_1 \pm 4\sqrt{1-\theta_1^2} = 0 \\
\Leftrightarrow &-3 \pm 4\frac{\theta_1}{\sqrt{1-\theta_1^2}} = 0 \\
\Leftrightarrow &\pm\tfrac{4}{3}\theta_1 = \sqrt{1-\theta_1^2} \\
\Rightarrow &\tfrac{16}{9}\theta_1^2 = 1-\theta_1^2 \\
\Leftrightarrow &\theta_1^2 = \tfrac{9}{25} \\
\Leftrightarrow &\theta_1 = \pm\tfrac{3}{5} \\
\Leftrightarrow &(\theta_1,\theta_2) = (\pm\tfrac{3}{5}, \pm\tfrac{4}{5})
\end{aligned}$$

על ידי בדיקה של הערכים שמניבים ארבעת הנקודות האלה מוצאים כי המינימום הגלובלי של פונקציית המטרה מתקבל במקרה של $(\theta_1,\theta_2) = (\frac{3}{5}, \frac{4}{5})$.

סקלרים, וקטורים מטריצות ונגזרות

במהלך הקורס אנו נתקל פעמים רבות בצורך לחשב נגזרות המערבות וקטורים ומטריצות. נזכיר / נסביר בקצרה של כיצד נגזרת אלו מחושבות. נתחיל במקרה המוכר של הגרדיאנט, בו אנו מבצעים גזירה של פונקציה סקלרית לפי וקטור. לאחר מכאן נראה כיצד הגדרה זו מורחבת גם למקרים נוספים בהם הפונקציה לא בהכרח סקלרית והגזירה היא לא בהכרח לפי וקטור.

בעבור פונקציה $f(x)$ אשר מקבלת וקטור x באורך d ומחזירה סקלר, פעולת הגרדיאנט מוגדרת באופן הבא:

$$\nabla_x f(x) = \frac{d}{dx} f(x) = \begin{bmatrix} \frac{d}{dx_1} f(x) \\ \frac{d}{dx_2} f(x) \\ \vdots \\ \frac{d}{dx_d} f(x) \end{bmatrix}$$

לדוגמא:

$$\frac{d}{dx}(a^\top x) = \frac{d}{dx}(\sum_{i=1}^{d} a_i x_i) = \begin{bmatrix} \frac{d}{dx_1}(\sum a_i x_i) \\ \frac{d}{dx_2}(\sum a_i x_i) \\ \vdots \\ \frac{d}{dx_d}(\sum a_i x_i) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} = a$$

נסתכל כעת על מקרה מעט יותר מורכב בו אנו רוצים לגזור פונקציה וקטורית $f(x)$ אשר מקבלת וקטור x באורך d ומחזירה וקטור באורך n:

$$\frac{d}{dx}f(x) = \begin{bmatrix} | & | & & | \\ \frac{d}{dx_1}f(x) & \frac{d}{dx_2}f(x) & \cdots & \frac{d}{dx_d}f(x) \\ | & | & & | \end{bmatrix}$$

שימו לב שהתוצאה של הגזירה הינה מטריצה בגודל $n \times d$. באופן כללי, הגדול תוצאת פעולת הגזירה יהיה תמיד הגודל של האובייקט שאותו גוזרים בתוספת הגודל של האיבר שלפיו אנו מבצעים את הגזירה. דוגמאות:

- תוצאת הגזירה של מטריצה בגודל $n \times m$ לפי וקטור באורך d תהיה טנזור בגודל $n \times m \times d$.
- תוצאת הגזירה של סקלאר לפי מטריצה בגודל $n \times m$ תהיה מטריצה בגודל $n \times m$.
- תוצאת הגזירה של מטריצה בגודל $n \times m$ לפי מטריצה בגודל $o \times p$ תהיה טנזור בגודל $n \times m \times o \times p$.

למורת חשוב להבין כיצד נגזרות אלו מודרות ומחושבות, בפועל אנחנו כמעט ולא ניתקל בצורך לחשב אותם על פי ההגדרה. במרבית המקרים, הנגזרות בהם ניתקל יבואו מתוך קבוצה מצומצמת של נגזרות מוכרות (או הרכבה שלהם עם פונקציות אחרות) ונוכל להשתמש בתוצאות מוכרות ולחסוך עבודה מיותרת. ניתן למצוא רשימה של נגזרות בדף הנוסחאות של הקורס.

תרגיל 1.2 - תרגיל בנגזרות

חשבו את הנגזרות הבאות:

1. $\frac{d}{dx}\|x\|_2^2$.
2. $\frac{d}{dx}\|x\|_2$. הנחיה: השתמש בכלל השרשרת
3. $\frac{d}{dx}(x^\top Ax)$.
4. $\frac{d}{dA}(x^\top Ax)$.

פתרון 1.2

סעיף 1

$$\frac{d}{dx}\|x\|_2^2 = \frac{d}{dx}(x^\top x) = \frac{d}{dx}(\sum_{i=1}^{d} x_i^2) = \begin{bmatrix} \frac{d}{dx_1}(\sum x_i^2) \\ \frac{d}{dx_2}(\sum x_i^2) \\ \vdots \\ \frac{d}{dx_d}(\sum x_i^2) \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_d \end{bmatrix} = 2x$$

סעיף 2 נסמן $h(x) = \|x\|_2^2$ ונשים לב ש $\|x\|_2 = \sqrt{h(x)}$. כמו כן נשתמש בעובדה שאת הנגזרת של $h(x)$ כבר חישבנו בסעיף הקודם:

$$\frac{d}{dx}\|x\|_2 = \frac{d}{dx}\sqrt{h(x)} = \frac{1}{2\sqrt{h(x)}} \cdot \frac{d}{dx}h(x) = \frac{x}{\sqrt{h(x)}}$$

סעיף 3 נגזור על פי הגדרה:

$$\frac{d}{dx}(x^\top Ax) = \frac{d}{dx}(\sum_{i,j} a_{i,j}x_ix_j) = \begin{bmatrix} \frac{d}{dx_1}(\sum_{i,j} a_{i,j}x_ix_j) \\ \frac{d}{dx_2}(\sum_{i,j} a_{i,j}x_ix_j) \\ \vdots \\ \frac{d}{dx_d}()\sum_{i,j} a_{i,j}x_ix_j \end{bmatrix} = \begin{bmatrix} \sum_i a_{i,1}x_i + \sum_j a_{1,j}x_j \\ \sum_i a_{i,2}x_i + \sum_j a_{2,j}x_j \\ \vdots \\ \sum_i a_{i,d}x_i + \sum_j a_{d,j}x_j \end{bmatrix}$$

$$= \begin{bmatrix} A_{:,1}^\top x + A_{1,:}x \\ A_{:,2}^\top x + A_{2,:}x \\ \vdots \\ A_{:,d}^\top x + A_{d,:}x \end{bmatrix} = A^\top x + Ax = (A^\top + A)x$$

סעיף 4 נגזור על פי הגדרה:

$$\frac{d}{dA}(x^\top Ax) = \frac{d}{dA}(\sum_{i,j} a_{i,j}x_ix_j)$$

$$= \begin{bmatrix} \frac{d}{da_{1,1}}(\sum_{i,j} a_{i,j}x_ix_j) & \frac{d}{da_{1,2}}(\sum_{i,j} a_{i,j}x_ix_j) & \cdots & \frac{d}{da_{1,m}}(\sum_{i,j} a_{i,j}x_ix_j) \\ \frac{d}{da_{2,1}}(\sum_{i,j} a_{i,j}x_ix_j) & \frac{d}{da_{2,2}}(\sum_{i,j} a_{i,j}x_ix_j) & \cdots & \frac{d}{da_{2,m}}(\sum_{i,j} a_{i,j}x_ix_j) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d}{da_{n,1}}(\sum_{i,j} a_{i,j}x_ix_j) & \frac{d}{da_{n,2}}(\sum_{i,j} a_{i,j}x_ix_j) & \cdots & \frac{d}{da_{n,m}}(\sum_{i,j} a_{i,j}x_ix_j) \end{bmatrix}$$

$$= \begin{bmatrix} x_1x_1 & x_1x_2 & \cdots & x_1x_m \\ x_2x_1 & x_2x_2 & \cdots & x_2x_m \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \cdots & x_nx_m \end{bmatrix} = xx^\top$$

תרגיל 1.3 - בחזרה לבעיית המעגל

נחזור לפונקציית ה"ציון" של מידת ההתאמה של מעגל לנקודות מהדוגמא בתחילת התרגול.

$$f(c,r) = \frac{1}{n}\sum_{i=1}^{n}(\|x^{(i)} - c\|_2 - r)^2$$

חשבו את הנגזרות של פונקציה זו: $\nabla_c f$ ו $\frac{d}{dr}f$. (לצורך החישוב של $\nabla_c f$ השתמשו בשיטה דומה לזו שהופיע בסעיף 2 של השאלה הקודמת)

פתרון 1.3

נתחיל בגזירה לפי r:

$$\frac{d}{dr}f(c,r) = \frac{2}{n}\sum_{i=1}^{n}(r - \|x^{(i)} - c\|_2)r$$

על פי ההנחיה, לצורך הגזירה לפי c נסמן את פונקציית העזר $h(x,c) = \|x - c\|_2^2$ בעזרתה נוכל לרשום את f כ:

$$f(c,r) = \frac{1}{n}\sum_{i=1}^{n}(\sqrt{h(x^{(i)},c)} - r)^2$$

נחשב תחילה את הנגזרת של $h(x,c)$:

$$\begin{aligned}\frac{d}{dc}h(x,c) &= \frac{d}{dc}\|x-c\|_2^2 \\ &= \frac{d}{dc}(x-c)^\top(x-c) \\ &= \frac{d}{dc}(\|x\|_2^2 - 2c^\top x + \|c\|_2^2) \\ &= 2(c-x)\end{aligned}$$

נשתמש כעת בתוצאה זו על מנת לחשב את הנגזרת הכוללת:

$$\begin{aligned}\frac{d}{dc}f(c,r) &= \frac{2}{n}\sum_{i=1}^{n}(\sqrt{h(x^{(i)},c)}-r))\cdot\left(\frac{d}{dh(x^{(i)})}\sqrt{h(x^{(i)},c)}\right)\cdot\frac{d}{dx}h(x^{(i)}) \\ &= \frac{2}{n}\sum_{i=1}^{n}(\sqrt{h(x^{(i)},c)}-r))\frac{1}{\sqrt{h(x^{(i)},c)}}(c-x^{(i)}) \\ &= \frac{2}{n}\sum_{i=1}^{n}(r-\|x^{(i)}-c\|_2)\frac{x^{(i)}-c}{\|x^{(i)}-c\|_2}\end{aligned}$$

descent Gradient (שיטת הגרדיאנט)

כפי שציינו קודם, במקרים רבים לא נוכל פתור את בעיות האופטימיזציה על ידי גזירה והשוואה ל0 ונאלץ לעשות שימוש בשיטות נומריות. נתאר כאן בקצרה את שיטת ה**descent gradient** (שיטת הגרדיאנט) שהיא אחת השיטות הבסיסיות ביותר אשר מנסה לפתור את בעיות מסוג זה. אנו עוד נדון בהמשך הקורס בהרחבה בתכונות ובבעיות הקיימות בשיטה זו, אך בשלב זה נסתפק בלתאר את אופן פעולתה.

הרעיון מאחורי שיטה זו הינו להתחיל בנקודה אקראית כל שהיא במרחב ולהתחיל לזוז בצעדים קטנים לכיוון שבו פונקציית המטרה קטנה באופן המהיר ביותר. הכיוון הזה הוא כמובן הכיוון ההפוך לגרדיאנט של הפונקציה. זהו אלגוריתם חמדן (greedy) אשר מנסה בכל צעד לשפר במעט את מצבו ביחס לשלב הקודם. אלגוריתמים מסוג זה מתכנסים לרוב למינימום לוקאלי ולא למינימום הגלובלי של הפונקציה. האלגוריתם זה רחוק מלתת מענה מושלם לבעיה, אך במקרים רבים הוא מצליח לספק פתרון סביר.

האלגוריתם

- מאתחלים את $\theta^{(0)}$ בנקודה אקראית כל שהיא
- חוזרים על צעד העדכון הבא עד להתכנסות:

$$\theta^{(t+1)} = \theta^{(t)} - \eta\nabla_\theta f(\theta^{(t)})$$

הפרמטר η אשר קובע את גודל הצעדים אשר נעשה בתהליך ההתכנסות. (את הדיון על קריטריון ההתכנסות ועל הבחירת של η נשאיר לשלב מאוחר יותר)

תרגיל 1.4 - fitting circle for descent Gradient

רשומו את צעד העדכון של אלגורתם הdescent gradient בעבור המקרה של התאמת העיגול.

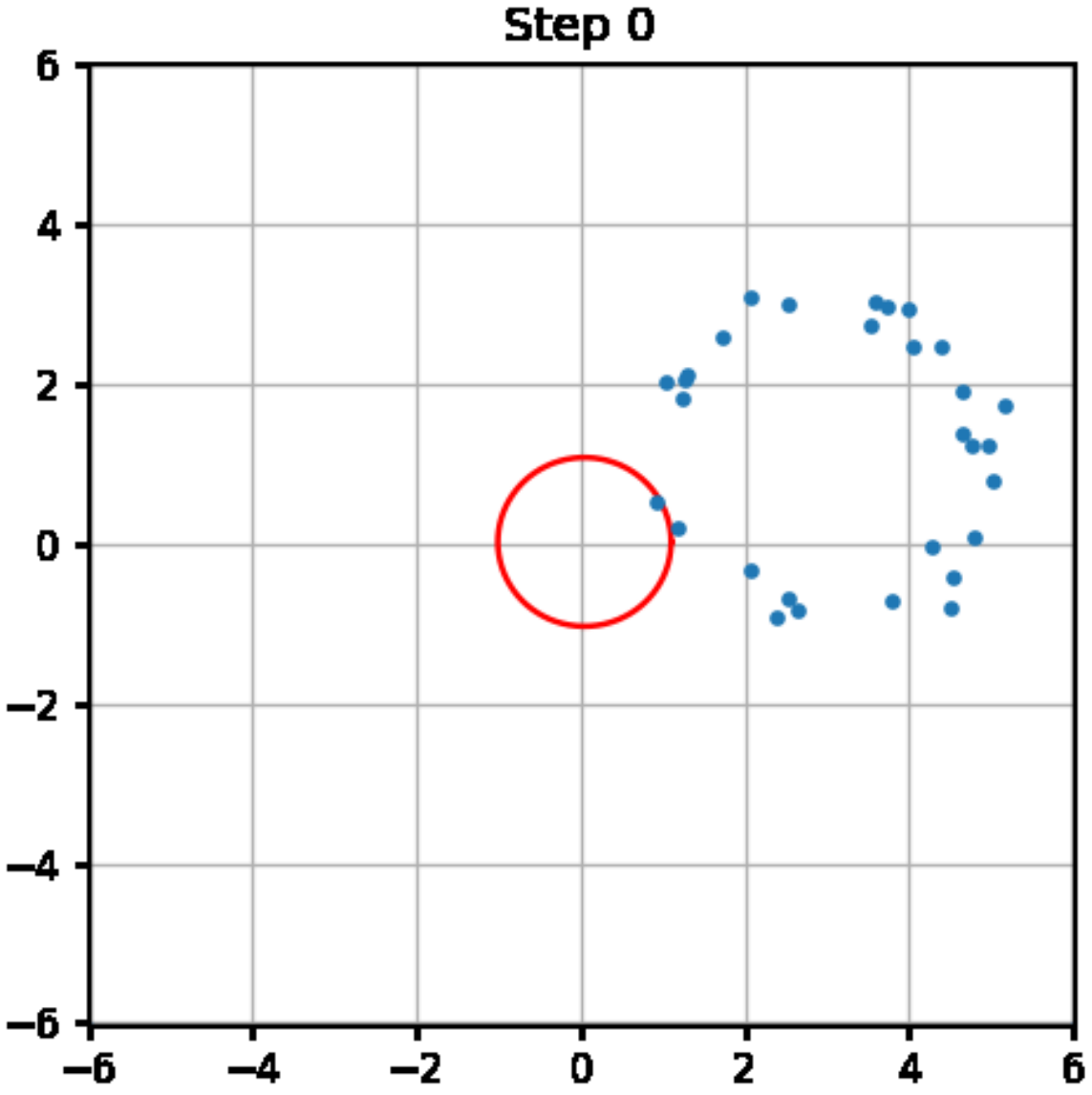
פתרון 1.4

על פי הנגזרת שחיבנו בתרגיל 1.3 נסיק כי צעד העדכון הוא:

$$r^{(t+1)} = r^{(t)} - \frac{2\eta}{n}\sum_{i=1}^{n}(r^{(t)} - \|x^{(i)} - c^{(t)}\|_2)r^{(t)}$$
$$c^{(t+1)} = c^{(t)} - \frac{2\eta}{n}\sum_{i=1}^{n}(r^{(t)} - \|x^{(i)} - c^{(t)}\|_2)\frac{x^{(i)} - c^{(t)}}{\|x^{(i)} - c^{(t)}\|_2}$$

הרצה של האלגוריתם

למטה מוצג תהליך ההתכנסות של אלגוריתם הגרדיאנט בעבור $\eta = 0.01$ ו 500 צעדים כאשר מתחילים את התהליך ממעגל היחידה:


Step 0
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6