שיטות חישוביות באופטימיזציה 046197
תרגיל מחשב
תאריך הגשה 07/06/2021

דגשים כלליים

- יש להגיש תיקיית zip ובה קבצי הקוד וקובץ pdf המכיל את התשובות לשאלות ואת הגרפים שהינכם מתבקשים להפיק.
- יש **לנמק** כל תשובה ולהסביר את הגרפים המתקבלים. תשובה/גרף ללא הסבר יקבלו ניקוד חלקי.
- הינכם רשאים לכתוב ב־Matlab או ב־Python.
- בכל מקום בו הינכם מתבקשים לשרטט על גרף **פונקצית שגיאה**, עליכם להציג את הגרף בסקלה לוגריתמית.

בתרגיל זה נפתור בעית Least Squares לשערוך פרמטרים.
נתון הפולינום מסדר n:

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

המקדמים $a_0, a_1, a_2, \ldots, a_n$ הינם מקדמים לא ידועים שברצוננו לשערך. לשם כך, מבצעים m מדידות ($m > n$):

$$y_i = f(x_i) + w_i, \quad i = 1, 2, \ldots, m.$$

נקודות המדידה, x_i, ידועות, אך למדידה מתווסף רעש גאוסי, $w_i \sim \mathcal{N}(0, \sigma^2)$. הרעש במדידה ה־$i$ בלתי תלוי ברעש במדידות האחרות. נסמן:

$$\mathbf{a} = \begin{bmatrix} a_0 & a_1 & \cdots & a_n \end{bmatrix}^T \in \mathbb{R}^{n+1}$$
$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^T \in \mathbb{R}^{m}$$

מוצע לשערך את $\mathbf{a}$ ע"י פתרון בעיית האופטימיזציה

$$\min_{\mathbf{a}} \left\{ \frac{1}{2m} \sum_{i=1}^{m} (y_i - f(x_i))^2 \right\}.$$

1. מצאו מטריצה $X \in \mathbb{R}^{m \times (n+1)}$ שעבורה הבעיה (P) שקולה לבעיה הבאה:

$$(P) \qquad \min_{\mathbf{a}} \left\{ h(\mathbf{a}) = \frac{1}{2m} \|\mathbf{y} - X\mathbf{a}\|_2^2 \right\}.$$

2. הראו כי הבעיה (P) קמורה, ומצאו את הפתרון שלה באופן אנליטי (הניחו כי דרגת המטריצה X היא $n+1$).

נניח מעתה כי $n=3$ וכי $a_0=0, a_1=1, a_2=1/2, a_3=-2$.

צרו וקטור $\mathbf{x}\in\mathbb{R}^m$ של $m=10,000$ נקודות מדידה הנדגמות באקראי מתוך פילוג אחיד על פני הקטע $[-1,1]$. חשבו את f על פני נקודות המדידה וצרו את הוקטור $\mathbf{y}$ על ידי הוספת רעש גאוסי בעל שונות $\sigma^2=0.5$ לכל אחת מהמדידות, כפי שתואר מעלה.

חלק 1

3. כתבו פונקציה המקבלת כקלט את הוקטורים $\mathbf{x}$ ו־$\mathbf{y}$, ומחזירה כפלט את וקטור המקדמים המשוערך $\hat{\mathbf{a}}$, כלומר, על הפונקציה לממש את פתרון הבעיה (P). נסמן: $h(\hat{\mathbf{a}})=h^*$.

4. הריצו את הפונקציה הנ"ל עבור הוקטורים $\mathbf{x}$ ו־$\mathbf{y}$ שיצרתם. שרטטו על גרף אחד בקטע $[-1,1]$ את:
   - הפולינום $f(x)$
   - וקטור המדידות המורעשות, $\mathbf{y}$ (באמצעות scatter)
   - הפולינום המשוערך $\hat{f}(x)$, שמקדמיו הם $\hat{\mathbf{a}}$

כעת, הניחו כי הפתרון האופטימלי לבעיה (P) מתקבל בתחום $C=\left\{\mathbf{a}\in\mathbb{R}^{n+1}:\|\mathbf{a}\|_2\le r\right\}$, עבור $r=4$.

הערה: שימו לב כי וקטור המקדמים ה"אמיתי" (זה שבעזרתו מיוצרות המדידות) אכן ב־C.

אם כן, נתעניין כעת בבעיה המאולצת הבאה:

$$(P')\qquad \begin{array}{ll} \min\limits_{\mathbf{a}} & h(\mathbf{a})=\frac{1}{2m}\|\mathbf{y}-X\mathbf{a}\|_2^2 \\ \text{s.t.} & \|\mathbf{a}\|_2\le r. \end{array}$$

עבור בעיה זו, מתקיים:

$$D\equiv \operatorname{diam} C=2r.$$

כמו כן, ניתן להראות שפונקציית המטרה רציפה ליפשיץ בתחום C עם קבוע הנתון על ידי

$$G=2r\cdot\lambda_{\max}\left(\frac{X^TX}{m}\right)+\left\|\frac{X^T\mathbf{y}}{m}\right\|_2$$

כאשר $\lambda_{\max}(\frac{X^TX}{m})$ הוא הערך העצמי המקסימלי של המטריצה $\frac{1}{m}X^TX$.

חלק 2

5. כתבו פונקציה המקבלת כקלט את הוקטורים $\mathbf{x}$ ו־$\mathbf{y}$ (וייתכן פרמטרים נוספים) וממשת את אלגוריתם הגרדיאנט המוטל עבור הבעיה (P'):
   - הגרילו את נקודת ההתחלה מפילוג נורמלי, $\mathbf{a}_0\sim\mathcal{N}(\mathbf{0},I)$.
   - קבעו תנאי עצירה מהצורה $\|\nabla h(\mathbf{a})\|_2\le\epsilon$.
   - הגבילו את מספר הצעדים הכולל של האלגוריתם.

   הדרכה: בדומה למה שראינו בכיתה, ההיטל על הקבוצה C הנ"ל נתון על ידי

$$P_C(\mathbf{a})=\begin{cases} r\cdot\frac{\mathbf{a}}{\|\mathbf{a}\|_2}, & \mathbf{a}\notin C \\ \mathbf{a}, & \mathbf{a}\in C \end{cases}$$

6. הריצו את הפונקציה הנ"ל עבור שתי בחירות גודל הצעד הבאות:

- גודל צעד דועך:

$$\eta_t = \frac{D}{G\sqrt{t}}, \quad t = 1, 2, \ldots$$

- גודל צעד אדפטיבי (AdaGrad):

$$\eta_t = \frac{D}{\sqrt{2\sum_{\tau=1}^{t} \|\mathbf{g}_\tau\|_2^2}}, \quad t = 1, 2, \ldots$$

שרטטו על גרף אחד, בסקלה לוגריתמית, את פונקצית השגיאה כתלות במספר האיטרציות של האלגוריתם, כלומר, את $e_t = h(\mathbf{a}_t) - h^*$, עבור שתי בחירות גודל הצעד. הסבירו.

למעשה, פונקצית המטרה $h(\mathbf{a})$ היא חלקה.

7. הוכיחו כי קבוע החלקות של פונקצית המטרה $h(\mathbf{a})$ נתון על ידי

$$L = \lambda_{\max}\left(\frac{X^T X}{m}\right).$$

8. הריצו את הפונקציה מסעיף 5 עבור שלושת הבחירות הבאות של גודל צעד קבוע: $\eta = {}^{1}/_{10L}, {}^{1}/_{L}, {}^{10}/_{L}$. שרטטו על גרף אחד, בסקלה לוגריתמית, את פונקצית השגיאה כתלות במספר האיטרציות של האלגוריתם עבור הבחירות הנ"ל. הסבירו.

9. השוו בין AdaGrad לבין בחירת גודל צעד קבוע $\eta = {}^{1}/_{L}$ (הציגו על גרף אחד את פונקצית השגיאה עבור שניהם). הסבירו כיצד התוצאות מתיישבות עם החסמים התיאורטיים שלמדנו.

חלק 3

10. כתבו פונקציה המקבלת כקלט את הוקטורים $\mathbf{x}$ ו־$\mathbf{y}$ ופרמטר $b \in \{1, \ldots, m\}$ (וייתכן פרמטרים נוספים) וממשת את אלגוריתם הגרדיאנט הסטוכסטי (המוטל) בגרסת ה־mini-batch עבור הבעיה (P'), כלומר, וקטור הגרדיאנט משוערך באמצעות דגימה אחידה של b מתוך m המדידות.
כלל העדכון של האלגוריתם הינו מהצורה:

$$\mathbf{a}_{t+1} = P_C(\mathbf{a}_t - \eta_{t+1}\mathbf{g}_t),$$

עבור בחירת גודל צעד דועך:

$$\eta_t = \frac{D}{G\sqrt{t}}, \quad t = 1, 2, \ldots$$

11. הריצו את הפונקציה הנ"ל עבור הבחירות הבאות של גודל mini-batch: $b = 1, 10, 100, 10000$.

- שרטטו על גרף אחד את פונקצית השגיאה כתלות במספר האיטרציות של האלגוריתם עבור כל אחת מהבחירות הנ"ל.
- שרטטו על גרף אחד את פונקצית השגיאה כתלות **בזמן** שעבר מתחילת האלגוריתם עבור כל אחת מהבחירות הנ"ל.

חוו דעתכם על התוצאות. התייחסו להבדלים בין הבחירות השונות של גודל mini-batch.