שיטות חישוביות באופטימיזציה 046197
תרגיל בית מספר 3
תאריך הגשה 01/06/2021

תרגיל 1
ראינו כי עבור פונקציה קמורה ו־L־חלקה, אלגוריתם הגרדיאנט עם גודל צעד קבוע $\eta \le 1/L$ מתכנס למינימום הגלובלי של הפונקציה בקצב של $O(1/t)$. בשאלה זו נוכיח תוצאת התכנסות של אלגוריתם הגרדיאנט עבור פונקציה שאינה בהכרח קמורה.
תהי f פונקציה L־חלקה וחסומה מלמטה על ידי f^*. היזכרו בכלל העדכון עבור אלגוריתם הגרדיאנט עם גודל צעד קבוע:

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \eta \nabla f(\mathbf{x}_{t-1}), \quad t \ge 1.$$

נניח כי $\eta \le 1/L$.

1. הוכיחו כי
$$f(\mathbf{x}_t) \le f(\mathbf{x}_{t-1}) - \eta \left(1 - \frac{\eta L}{2}\right) \|\nabla f(\mathbf{x}_{t-1})\|_2^2 .$$

2. היעזרו בכך ש־$\eta \le 1/L$ והראו כי מתקיים
$$\|\nabla f(\mathbf{x}_{t-1})\|_2^2 \le \frac{2}{\eta}(f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t)).$$

3. הוכיחו כי
$$\sum_{t=0}^{T} \|\nabla f(\mathbf{x}_t)\|_2^2 \le \frac{2}{\eta}(f(\mathbf{x}_0) - f^*)$$

4. חסמו מלמטה את הסכום מהסעיף הקודם על מנת לקבל
$$\min_{\tau=0,\ldots,t} \|\nabla f(\mathbf{x}_\tau)\|_2 \le \sqrt{\frac{2}{\eta(t+1)}(f(\mathbf{x}_0) - f^*)}.$$

שימו לב כי הראנו התכנסות בקצב של $O(1/\sqrt{t})$ (או $O(1/\epsilon^2)$) לנקודה שהיא ϵ**־תת־סטציונרית**, כלומר, נקודה $\mathbf{x}$ המקיימת $\|\nabla f(\mathbf{x})\|_2 \le \epsilon$.

תרגיל 2
תהי $f : \mathbb{R}^n \to \mathbb{R}$ פונקציה קמורה ו־L־חלקה. נניח כי מריצים את אלגוריתם הגרדיאנט עם סדרת גדלי צעד $\{\eta_t\}_{t\ge0}$, כלומר, מבצעים עדכון מהצורה

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t), \quad t \ge 0.$$

נסמן את הממזער של f על ידי $\mathbf{x}^*$. כמו כן, נסמן:

$$h_t \triangleq f(\mathbf{x}_t) - f(\mathbf{x}^*); \quad d_t \triangleq \|\mathbf{x}_t - \mathbf{x}^*\|.$$

1. הוכיחו כי לכל $t \geq 0$ מתקיים

$$d_{t+1}^2 \leq d_t^2 - 2\eta_t h_t + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|_2^2$$

רמז: היעזרו בקמירות של f על מנת לחסום מלמטה את $\nabla f(\mathbf{x}_t)^T(\mathbf{x}_t - \mathbf{x}^*)$.

נבחר את גודל הצעד בזמן t להיות

$$\eta_t = \frac{h_t}{\|\nabla f(\mathbf{x}_t)\|_2^2}.$$

שימו לב שבחירת גודל צעד זו ממזערת את החסם על d_{t+1}^2 שהוכחנו בסעיף הקודם (זוהי פרבולה ב־η_t).

2. הוכיחו כי לכל $t \geq 0$ מתקיים

$$d_{t+1}^2 - d_t^2 \leq -\frac{h_t}{2L}$$

הדרכה: היעזרו בלמה 1 המופיעה בתרגיל 2 בתרגול מספר 7:
תהי $f : \mathbb{R}^n \to \mathbb{R}$ פונקציה L־חלקה בעלת מינימום גלובלי המתקבל בנקודה $\mathbf{x}^*$. אזי, לכל $\mathbf{x} \in \mathbb{R}^n$ מתקיים

$$\|\nabla f(\mathbf{x})\|_2^2 \leq 2L(f(\mathbf{x}) - f(\mathbf{x}^*)).$$

3. הוכיחו כי לכל $T \geq 1$ מתקיים

$$\sum_{t=0}^{T-1} h_t \leq 2Ld_0^2$$

והסיקו כי

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{2Ld_0^2}{T} = \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{T}$$

כאשר $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{x}_t$.

הראנו, אם כן, כי עבור בחירת גודל הצעד הנ"ל, אלגוריתם הגרדיאנט עבור פונקציה L־חלקה מתכנס בקצב של $O(1/t)$. כבר ראינו שאלגוריתם הגרדיאנט מתכנס בקצב זה עבור משפחת הפונקציות החלקות, אך עבור בחירת גודל צעד קבוע $1/L$. לעיתים, קשה לחשב את קבוע החלקות של פונקציה ולכן, אם במקרה אנו יודעים את $f(\mathbf{x}^*) \equiv f^*$ (ולכן את h_t בכל צעד), אז ניתן להשתמש בגודל הצעד הנ"ל (שאינו מצריך את ידיעת L) ולקבל את אותו קצב התכנסות.

תרגיל 3

1. נתונה בעית האופטימיזציה החד־מימדית הבאה

$$\max_{0 \leq x \leq S} \left\{ 2\sqrt{S-x} + \frac{x}{\sqrt{S}} \right\}.$$

הוכיחו כי הפתרון האופטימלי של הבעיה הוא $2\sqrt{S}$ והסיקו כי לכל $0 \le x \le S$ מתקיים

$$2\sqrt{S-x} + \frac{x}{\sqrt{S}} \le 2\sqrt{S}$$

הדרכה: הראו כי הנגזרת של פונקצית המטרה היא שלילית בתחום, והסיקו כי הפתרון מתקבל על השפה.

2. הוכיחו את למה 5 המופיעה בתרגיל 3 בתרגול מספר 6:
לכל סדרת מספרים אי־שליליים $a_1, \ldots, a_n$ מתקיים

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \le 2\sqrt{\sum_{i=1}^{n} a_i}.$$

רמז: העזרו בסעיף הקודם והוכיחו באינדוקציה.

תרגיל 4

נתעניין בבעיה הבאה

$$\min_{\mathbf{x} \in C} f(\mathbf{x})$$

עבור f קמורה ו־C קמורה, קומפקטית ובעלת קוטר D.
ראינו בכיתה שעבור אלגוריתם הגרדיאנט המוטל עם בחירת גודל צעד AdaGrad:

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

מתקבלת התוצאה הבאה:

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \le \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right) \le \frac{1}{T} \cdot \sqrt{2D^2 \sum_{t=1}^{T} \|\mathbf{g}_t\|_2^2}$$

עבור f רציפה ליפשיץ עם קבוע G, הדבר שקול לקצב התכנסות של $O(GD/\sqrt{T})$.
בשאלה זו נראה כי אם בנוסף f היא L־חלקה, אז מתקבל קצב התכנסות של $O(LD^2/T)$.

1. הוכיחו את למה 1 המופיעה בתרגיל 2 בתרגול מספר 7:
תהי $f : \mathbb{R}^n \to \mathbb{R}$ פונקציה L־חלקה בעלת מינימום גלובלי המתקבל בנקודה $\mathbf{x}^*$. אזי, לכל $\mathbf{x} \in \mathbb{R}^n$ מתקיים

$$\|\nabla f(\mathbf{x})\|_2^2 \le 2L(f(\mathbf{x}) - f(\mathbf{x}^*)).$$

נניח כי המינימום הגלובלי של f מתקבל ב־C, ולכן ניתן להשתמש בלמה הנ"ל לכל $\mathbf{x} \in C$.

2. הוכיחו כי עבור f שהיא L־חלקה, אלגוריתם הגרדיאנט המוטל עם בחירת גודל צעד כנ"ל מקיים

$$\sum_{t=1}^{T}\left(f(\mathbf{x}_t) - f(\mathbf{x}^*)\right) \le 4LD^2$$

והסיקו כי

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \le \frac{4LD^2}{T}.$$

תרגיל 5

בשאלה זו נתעניין בבעיה

$$\min_{\mathbf{x}} \; F(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}),$$

כאשר $\{f_i\}_{i=1}^m$ הן L־חלקות.
ראינו בכיתה (תרגיל 2 בתרגול מספר 7) שתחת הנחת realizability, קצב ההתכנסות של אלגוריתם הגרדיאנט הסטוכסטי הינו $O(1/t)$ (לעומת $O(1/\sqrt{t})$ עבור המקרה כללי).
בתרגיל זה נראה כי אם בנוסף $F(\mathbf{x})$ קמורה במובן החזק עם קבוע μ, אז תחת ההנחה הנ"ל, קצב ההתכנסות של האלגוריתם הינו $O(e^{-\gamma^2 t})$ (לעומת $O(1/t)$ עבור המקרה הכללי).
נניח, אם כן, משערך גרדיאנט

$$\mathbf{g} = \nabla f_i(\mathbf{x}), \quad i \sim \textsc{Uniform}(\{1,\ldots,m\}).$$

וכלל עדכון

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{g}_t$$

בנוסף, נניח גודל צעד $\eta = \mu/L^2 = \gamma/L$.
נסמן: $h_t = F(\mathbf{x}_t) - F(\mathbf{x}^*)$.

1. הוכיחו כי $F(\mathbf{x})$ היא L־חלקה.
2. הוכיחו כי

$$h_{t+1} - h_t \le -\eta \nabla F(\mathbf{x}_t)^T \mathbf{g}_t + \frac{L}{2}\eta^2 \left\|\mathbf{g}_t\right\|_2^2 .$$

3. הוכיחו כי

$$\mathbb{E}\big[h_{t+1} - h_t\big] \le -\gamma^2 \mathbb{E}\big[h_t\big].$$

הדרכה: העזרו באי השוויון הבא שראינו בכיתה:

$$\mathbb{E}\big[\,\|\mathbf{g}_t\|_2^2 \,|\mathbf{x}_t\big] \le 2L(F(\mathbf{x}_t) - F(\mathbf{x}^*))$$

וכן בלמה הבאה (שראינו בכיתה גם כן):

$$\|\nabla F(\mathbf{x})\|_2^2 \ge 2\mu(F(\mathbf{x}) - F(\mathbf{x}^*)), \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

4. הוכיחו כי

$$\mathbb{E}\big[h_t\big] \le e^{-\gamma^2 t} \cdot \mathbb{E}\big[h_0\big]$$

כלומר,

$$\mathbb{E}\big[F(\mathbf{x}_t) - F(\mathbf{x}^*)\big] \le e^{-\gamma^2 t}(F(\mathbf{x}_0) - F(\mathbf{x}^*)).$$