למידה חישובית - תרגיל 3

(1) א) יהי $f:\mathbb{R}^n\to\mathbb{R}$ קמורה, $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^n$ ונגדיר $g(x)=f(Ax+b)$

נוכיח כי g קמורה. יהי $\lambda\in[0,1]$ $w_1,w_2\in\mathbb{R}^n$

$$g(\lambda w_1+(1-\lambda)w_2)\overset{\text{הגדרה}}{=}f\big(A(\lambda w_1+(1-\lambda)w_2)+b\big)=$$

$$=f\big(A(\lambda w_1+(1-\lambda)w_2)+\lambda b+(1-\lambda)b\big)=$$ השתמשתי בכך ש $b=\lambda b+(1-\lambda)b$

$$=f\big(\lambda(Aw_1+b)+(1-\lambda)(Aw_2+b)\big)\le$$

$$=f(Aw_1+b)+(1-\lambda)f(Aw_2+b)\le\lambda\cdot$$ אי שוויון ינסן

$$\lambda\cdot g(w_1)+(1-\lambda)g(w_2)\overset{\text{הגדרה}}{=}$$ לפי

לכן g קמורה לפי ההגדרה של פונקציה קמורה.

ב) ~~[illegible]~~ יהי $f_1,\dots,f_m$ פונקציות קמורות $f_i:\mathbb{R}^d\to\mathbb{R}$. נגדיר $g(x)=\max_i f_i(x)$

נוכיח כי g קמורה. יהי $\lambda\in[0,1]$ $w_1,w_2\in\mathbb{R}^d$

מסעיף קודם אנו יודעים כי לכל i: $f_i(\lambda w_1+(1-\lambda)w_2)\le\lambda f_i(w_1)+(1-\lambda)f_i(w_2)$

נוכל לבטא מיקסימום לשני הצדדים: $\max_i f_i(\lambda w_1+(1-\lambda)w_2)\le\max_i\{\lambda f_i(w_1)+(1-\lambda)f_i(w_2)\}$

$$\Rightarrow\max_i f_i(\lambda w_1+(1-\lambda)w_2)\le\max_i\lambda f_i(w_1)+\max_i(1-\lambda)f_i(w_2)$$

$$\Rightarrow g(\lambda w_1+(1-\lambda)w_2)\le\lambda g(w_1)+(1-\lambda)g(w_2)$$

ולכן g היא קמורה.

ג) $\ell_{\log}:\mathbb{R}\to\mathbb{R}$ $\ell_{\log}(z)=\log_2(1+e^{-z})$ נראה כי היא קמורה.

נשתמש בהגדרה אלטרנטיבית ש g קמורה כאשר $g''(x)>0$ לכל $x\in A$.

$$\ell'_{\log}(z)=\frac{-e^{-z}}{\ln 2\,(1+e^{-z})}$$

$$\ell''_{\log}(z)=\frac{e^{-z}\big(\ln 2(1+e^{-z})\big)+e^{-z}\big(-\ln 2\,e^{-z}\big)}{\big(\ln 2(1+e^{-z})\big)^2}=$$

$$=\frac{e^{-z}\big(\ln 2+\cancel{\ln 2\,e^{-z}}-\cancel{\ln 2\,e^{-z}}\big)}{\big(\ln 2(1+e^{-z})\big)^2}=\frac{\overbrace{\ln 2}^{>0}\;\overbrace{e^{-z}}^{>0}}{\underbrace{\big(\ln 2(\overbrace{1+e^{-z}}^{>1})\big)^2}_{\ge 0}}>0$$

ולכן $\ell_{\log}$ היא קמורה.

המשך

(1) ב) יהי $f: \mathbb{R}^2 \to \mathbb{R}$ ע"י $f(w) = \log(1+e^{w\cdot x})$.

לפי הגדרה $f(w) = \log_2(1+e^{-w\cdot x}) = \sum_{i=1}^{n} \log_2(1+e^{-w x_i})$

נגדיר פונקציות $f_1, \dots, f_n$ ע"י $f_i = \log_2(1+e^{-w x_i})$

כלומר $f(w) = \sum_i f_i(w)$. הראנו כי $\log$ קמורה ולכן f_i קמורות.

יהי $\lambda \in [0,1]$ ו x, y נקודות כך כי:

$$f(w) = \sum_{i=1}^{n} f_i(w) \le n \cdot \max_i f_i(w) \le n \cdot \max_i f_i(\lambda x + (1-\lambda) y) \le$$

כי קמורה לפי סעיף ב'

$$\le n \max\{\lambda f_i(x)\} + n \cdot \max_i \{(1-\lambda) f_i(y)\}$$

קמורה לפי סעיף ב

ולכן f קמורה מ"ח ל w.

למידה חישובית

(2) לפי הרמז תחילה נראה ש hinge loss חוסם מלמעלה את ה 0-1 loss.

ההגדרה של 0-1 loss מהתרגול:

$$\min_{w\in W} f(w) = \sum_{i=1}^{n} \Delta_{20}(\mathrm{sign}(g(x_i, w)), y_i) = \sum_{i=1}^{n} \ell_{20}(y_i \cdot g(x_i, w))$$

$$\ell_{20} = \begin{cases} 1 & r \le 0 \\ 0 & r \ge 0 \end{cases}$$

* מסקנה: עבור $r \ge 0$ $z_0(r) = 0$

$$\ell_{hinge}(r) = \max\{0, 1-r\}$$

אם $r \ge 1$ נקבל: $\ell_{hinge}(r) = 0 \ge 0 = z_0(r)$

אם $0 \le r < 1$ נקבל: $\ell_{hinge}(r) = 1 - r > 0 = z_0(r)$

אם $r < 0$ נקבל: $\ell_{hinge}(r) = 1 + |r| > 1 \ge z_0(r)$

סה"כ הוכחנו כי לכל $r \in \mathbb{R}$ $\ell_{hinge}(r) \ge z_0(r)$

נתון כי ניתן להפריד ש $y_i \cdot w^* \cdot x_i \ge 0$ לכל i

כאשר $w^* \in \mathbb{R}^d$ ו $y_i = \mathrm{sign}(w^* \cdot x_i)$

נרצה להראות ש $w^*_{hinge} = w^*$ ומה נסיק שאכן w^*_{hinge}

משיג את אותו אופטימליות ל z_0 loss.

כיוון שאנו יודעים שה נתונים ניתנים להפרדה לינארית אזי סבבו w^*_{hinge}

$$\sum_{i=1}^{n} \ell_{hinge}(y_i\, w^*_{hinge} \cdot x_i) = 0$$

וכיוון ש $\ell_{hinge}(r) \ge z_0(r)$ נקבל שלכל $1 \le i \le n$ $\ell_{hinge}(y_i\, w^*_{hinge} \cdot x_i) = 0$

ולכן $y_i\, w^*_{hinge} \cdot x_i \ge 1 > 0$ ולכן $\Delta_{20}(y_i, w^*_{hinge} x_i) = 0$.

בנוסף לכך:

$$y_i = 1 \overset{\text{נתון לכל}}{\Rightarrow} y_i w^*_{hinge} \cdot x_i \ge 1 \Rightarrow w^*_{hinge} x_i \ge 1 \Rightarrow \mathrm{sign}(w^*_{hinge} \cdot x_i) = 1$$

$$y_i = -1 \overset{\text{נתון לכל}}{\Rightarrow} y_i w^*_{hinge} x_i \ge 1 \Rightarrow w^*_{hinge} x_i \le 1 \Rightarrow \mathrm{sign}(w^*_{hinge} \cdot x_i) = -1$$

כלומר אכן w^*_{hinge} הוא האופטימלי. $\left(\mathrm{sign}(w^*_{hinge} \cdot x_i) = y_i \ \forall i\right)$

למידה חישובית המשך

3) א) אנו יודעים כי K קבוצה קמורה ולכן לפי ההגדרה לכל $\lambda \in [0,1]$ ו $w_1, w_2 \in K$

מתקיים $\lambda w_1 + (1-\lambda) w_2 \in K$ בפרט עבור $w_2 = x$, $w_1 = z$ (ב K לפי ההגדרה)

נקבל כי $x - \lambda(x-z) = \lambda z + (1-\lambda) x \in K$

ולכן כיוון ש $x = \arg\min_{x \in K} \|x-y\|_2$ נקבל כי $\|x - \lambda(x-z) - y\|_2 \le \|x-y\|_2$

$$\|x-y\|^2 \le \|x - \lambda(x-z) - y\|^2 = \|(x-y) - \lambda(x-z)\|^2 =$$

$$= \|(x-y)\|^2 - 2\lambda \langle x-y, x-z\rangle + \lambda^2 \|x-z\|^2 \Rightarrow$$

$$\Rightarrow \lambda^2 \|x-z\|^2 \ge 2\lambda \langle x-y, x-z\rangle \overset{\text{הנחה יומם}}{\Rightarrow}$$

$$\Rightarrow \langle x-y, x-z\rangle \le \lambda \|x-z\|^2$$

התנאי מתקיים לכל λ ולכן בפרט גם עבור $\lambda = 0$ (נקבל אי-שוויון

λ כרצונו) ולכן $\langle x-y, x-z\rangle \le 0$.

$$\|y-z\|^2 - \|x-z\|^2 = \|y-z+x-x\|^2 - \|x-z\|^2 =$$

$$= \|(x-z) - (x-y)\|^2 - \|x-z\|^2 =$$

$$= \underbrace{\|x-z\|^2}_{\ge 0} - \underbrace{2\langle x-z, x-y\rangle}_{\le 0} + \underbrace{\|x-z\|^2}_{\ge 0} \ge 0$$

ולכן $\|y-z\|^2 \ge \|x-z\|^2 \Leftarrow \|y-z\| \ge \|x-z\|$

ב) יהי $\varepsilon > 0$

ונראה כי convergence theorem ל GD עדיין מתקיים, כלומר $F(\bar{w}) - F(w^*) \le \varepsilon$

כאשר $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$, $w^* \in \arg\min_w F(w)$.

נשים לב כי ① $\nabla f(x_t) = \frac{x_t - y_{t+1}}{\eta_t}$ (הנחה יומם מהגדרה)

ונתנצל בנוסחה כמו בהרצאה שלפיה וקטורים A, B, $\langle A, B\rangle = \frac{1}{2}\left(\|A\|^2 + \|B\|^2 - \|A-B\|^2\right)$ ②

$$f(\bar{x}) - f(x^*) \overset{\text{מהרצאה}}{\le} \frac{1}{T}\sum_{t=1}^{T} \nabla f(x_t)(x_t - x^*) \le$$

$$\overset{\text{שימוש ①}}{=} \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\eta_t}(x_t - y_{t+1})(x_t - x^*) \le$$

$$\overset{\text{שימוש ②}}{=} \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2\eta_t}\left(\|x_t - y_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x^* - y_{t+1}\|^2\right) \le$$

$$\overset{\text{שינוי סדר}}{=} \frac{1}{T}\sum_{t=1}^{n} \frac{1}{2\eta_t}\left(\|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right) + \frac{1}{T}\sum_{t=1}^{n} \frac{1}{2\eta_t}\left(\|x_t - y_{t+1}\|^2\right) \le$$

המשך

③ ג) $\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2\eta_t}\left(\|x_t-x^*\|^2-\|x_{t+1}-x^*\|^2\right)+\frac{1}{T}\sum_{t=1}^{T}\frac{\eta}{2}\left(\|x_t-y_{t+1}\|^2\right)\le$

שים לב כי מתקיים:

$$\sum_{t=1}^{T}\left(\|x_t-x^*\|^2-\|x_{t+1}-x^*\|^2\right)=$$

$$=\|x_1-x^*\|^2\underbrace{-\|y_2-x^*\|^2+\|x_2-x^*\|^2}_{\le 0}-\dots\underbrace{-\|y_T-x^*\|^2+\|x_T-x^*\|^2}_{\le 0}-\|x_{T+1}-x^*\|^2\le$$

לפי מרחק נידף

$$\le\|x_1-x^*\|^2=\|x^*\|^2\overset{x_1=0}{=}\|x_1-x^*\|^2-\underbrace{\|x_{T+1}-x^*\|^2}_{\ge 0}\le\|x_1-x^*\|^2$$

מ ④ שים לב כי $\|x_t-y_{t+1}\|^2\le\eta^2\cdot\|\nabla f(x_t)\|^2$

נניח כמו כן $\|\nabla f(x_t)\|_2<G$ ו $\|x^*\|<B$

נבחר $\eta=\frac{\varepsilon}{G^2}$, $T=\frac{\varepsilon^2\cdot 2}{B\cdot G^2}$ כאשר $\varepsilon>0$ נתון:

ונקבל: $f(\bar{x})-f(x^*)\le\frac{B\cdot T}{2\eta}+\frac{\eta\cdot G^2}{2}=\varepsilon$.

למידה חישובית

(4) יהי $f: \mathbb{R}^n \to \mathbb{R}$ β-smooth ואינטגרבילית. שתנאי כאינפורמציה הלא סטוכסטי.

נניח כי $\eta < \frac{\beta}{2}$.

$$f(x_{t+1}) \le f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{\beta}{2} \|x_{t+1} - x_t\|^2 =$$

מהגדרת צעד מעבר

$$= f(x_t) + \langle \nabla f(x_t), (x_t - \eta \nabla f(x_t)) - x_t \rangle + \frac{\beta}{2} \|\eta \nabla f(x_t)\|^2 =$$

$$= f(x_t) + \langle \nabla f(x_t), -\eta \nabla f(x_t) \rangle + \frac{\beta \eta^2}{2} \|\nabla f(x_t)\|^2 =$$

$$= f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{\beta \eta^2}{2} \|\nabla f(x_t)\|^2 =$$

$$= f(x_t) + \|\nabla f(x_t)\|^2 - \left(\eta - \frac{\beta \eta^2}{2}\right) \Rightarrow$$

$$\Rightarrow$$

$$\Rightarrow \|\nabla f(x_t)\|^2 \le \left(\eta - \frac{\beta \eta^2}{2}\right)^{-1} \cdot \left(f(x_t) - f(x_{t+1})\right) \Rightarrow$$

$$\Rightarrow \sum_{t=1}^{m} \|\nabla f(x_t)\|^2 \le \sum_{t=1}^{m} \left(\eta - \frac{\beta \eta}{2}\right)^{-1} \left(f(x_t) - f(x_{t+1})\right) =$$

טור טלסקופי

$$= \underbrace{\left(\eta - \frac{\beta \eta}{2}\right)^{-1}}_{\text{קבוע, מ-0 כי } \beta < \frac{\eta}{2}} \left(\underbrace{f(x_1)}_{\text{קבוע}} - \underbrace{f(x_{m+1})}_{\text{מהנתון } \ge 0}\right)$$

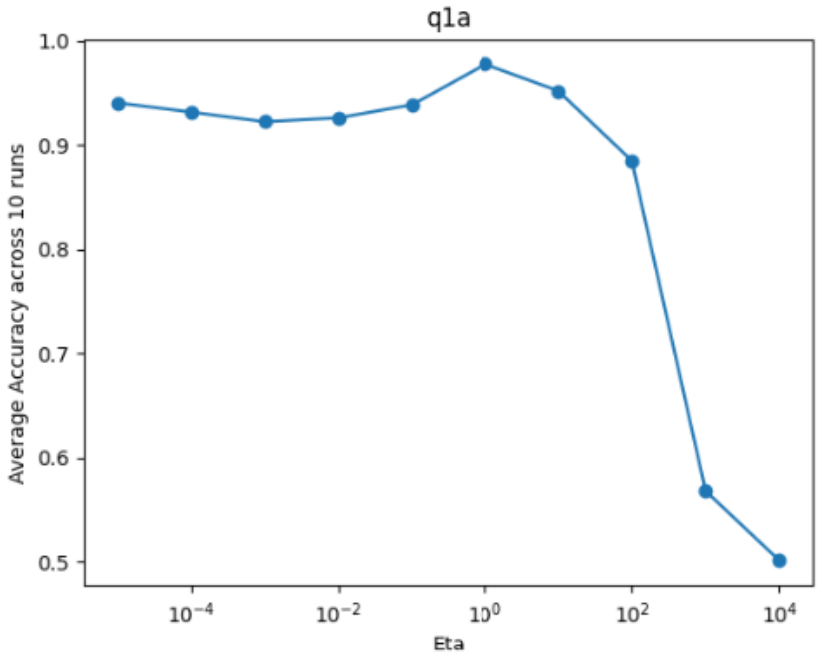
לכן עבור $m \to \infty$ אז $f(x_{m+1})$ יהיה קבוע ולכן סכום הטור יהיה קבוע

אז $f(m) \to \infty$ ולכן סכום הטור ישואף ל $-\infty$. בכל מקרה $\sum_{t=1}^{\infty} \|\nabla f(x_t)\|^2 < \infty$.

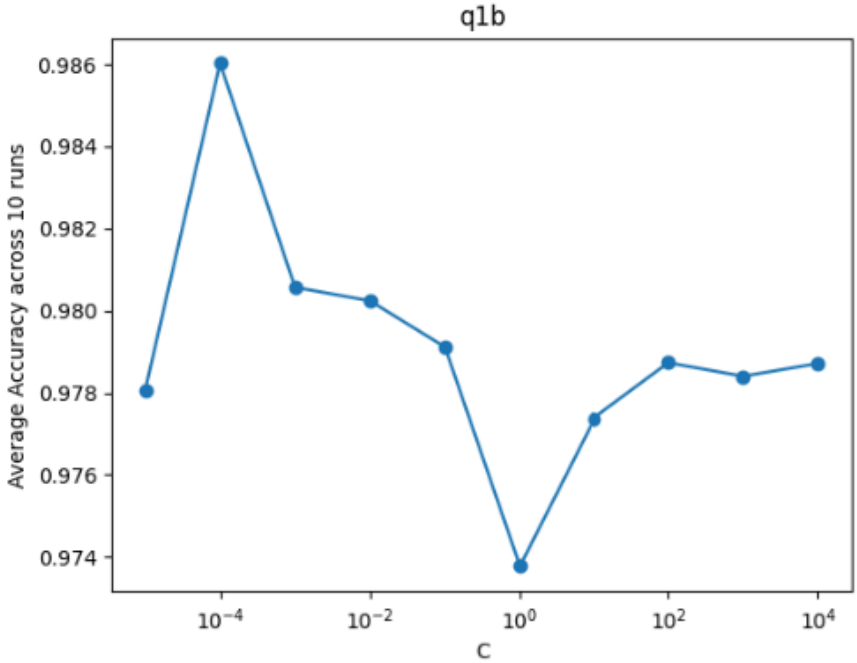
חלק מעשי

שאלה 1

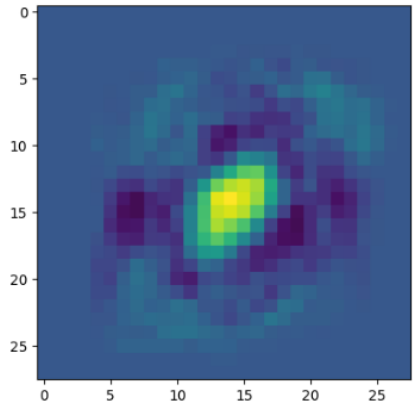
א. מקבלים כי ה eta הכי טוב הוא $10^0 = 1$



ב. מקבלים כי ה C הכי טוב הוא $10^{-4} = 0.0001$



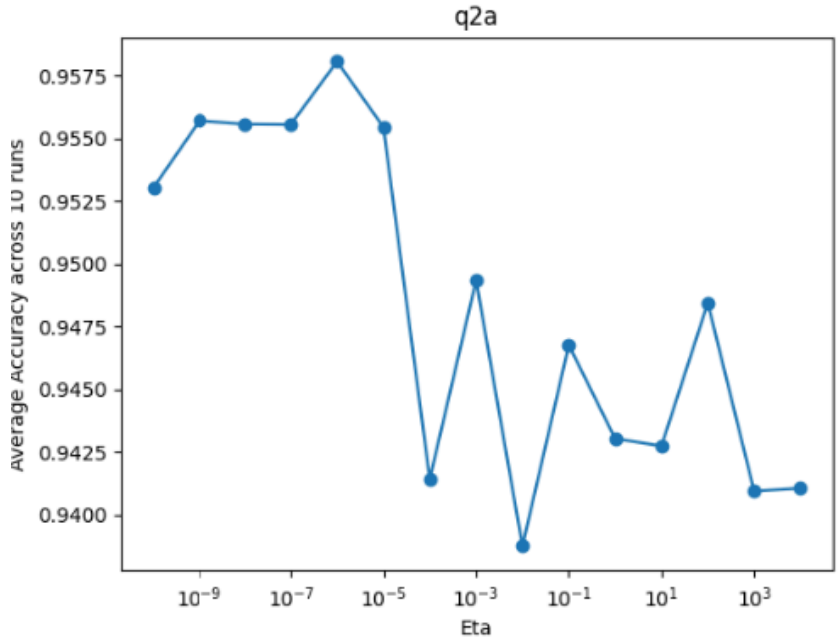
ג. התמונה מנרמלת את וקטור w, האזורים הכהים יותר מציינים סיכוי גבוה יותר לקבל 0 ואילו הבהירים יותר (במרכז) מציינים סיכוי גבוה יותר לקבל 8, ככל שהצבע בהיר המשקל גדול יותר. לכן ההבדל העיקר ביניהם הוא באמצע.



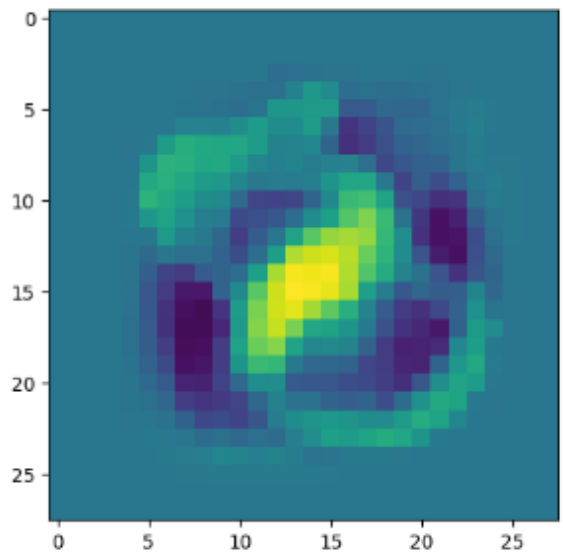
ד. מקבלים כי הדיוק הטוב ביותר הוא 0.9912998976458547.

שאלה 2

א. מקבלים כי ה eta הכי טוב הוא 10^{-6}



ב. מקבלים כי הדיוק הטוב ביותר הוא 0.9616171954964176.



ג. מתחילים עם w=0 לכן בהתחלה הנורמה היא 0. ה eta כיוון שהיא מנורמלת לפי t היא קטנה עם הזמן ולכן קצב הגידול של w קטן עם הזמן. לכן בהתחלה אנחנו רחוקים מהw האופטימלי אבל עם התקדמות באיטרציות של ה SGD ובגלל הרנדומליות נצפה להתקרב למינימום.