תרגיל מחשב

שיטות חישוביות באופטימיזציה – 046197

גרימוביץ יקטרינה 314092826

katy grimovitch
[Email address]

מטרת התרגיל

בתרגיל זה נפתור בעית Least Squares לשערוך פרמטרים

נתון פולינום מהצורה : $f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n$

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

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

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

$$\mathbf{a} = [a_0 \quad a_1 \ \ldots \quad a_n]^T \quad \in \quad R^{n+1}$$

$$\mathbf{y} = [y_0 \quad y_1 \ \ldots \quad y_m]^T \quad \in \quad R^{m}$$

מוצע לשערך את **a** ע"י פתרון בעיית האופטימיזציה:

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

סעיף 1

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

$$(P) \qquad \min_a \left\{ h(a) = \frac{1}{2m} \|y_i - \mathrm{Xa}\|_2^2 \right\}$$

$$\|y_i - Xa\|_2^2 = \sum_{i=1}^{m} \left(y_i - f(x_i)\right)^2 \quad \Rightarrow \quad Xa(i) = f(x_i) = \sum_{i=1}^{n} a_n X_i^n$$

$$\Rightarrow X_{i,j} = (x_i)^{j-1} \;\; for \;\; 1 \le i \le m, 1 \le j \le n+1$$

כלומר ניתן לכתוב כי

$$X = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^n \\ 1 & \ddots & & & \vdots \\ \vdots & & & & \vdots \\ 1 & x_m^1 & x_m^2 & \cdots & x_m^n \end{pmatrix}$$

סעיף 2

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

בעיית אופטימיזציה נתונה: $\min_a \left\{ h(a) = \frac{1}{2m} \|y - \mathrm{Xa}\|_2^2 \right\}$ (P) הינה בעיית אופטימיזציה ללא אילוצים על X לכל התחום לכן התחום הוא קמור (כל $\mathbb{R}^n$).

פונקציית מטרה הינה: $h(a) = \frac{1}{2m} \|y - \mathrm{Xa}\|_2^2$: לפי ההגדרה של נורמה, פונקציה $\|\cdots\|_2^2$ היא פונקציה קמורה. בנוסף, $y_i - \mathrm{Xa}$ הינה פונקציה אפינית . לכן, $h(a)$: בתנאי כי $m > 0$ היא הרכבה של פונקציה קמורה וטרנספורמציה אפינית, לכן כמו שראינו בתרגולים והרצאות , : $h(a)$ קמורה.

לכן, היא בעית אופטימיזציה (P) קמורה. נמצא עבורה פתרון אנליטי:

$$\frac{\partial h(a)}{\partial a} = \frac{\partial}{\partial a}\left(\frac{1}{2m}\|y - Xa\|_2^2\right) = \frac{1}{2m}\left(-2X^T(y - Xa)\right) = 0$$

$$X^T X a = X^T y$$

$$\bar{a}_{analitic} = (X^T X)^{-1} X^T y$$

הוגדר בבעיה כי $n = 3 \;, \; a_0 = 0 \,, \; a_2 = \frac{1}{2} \,, a_3 = -2$

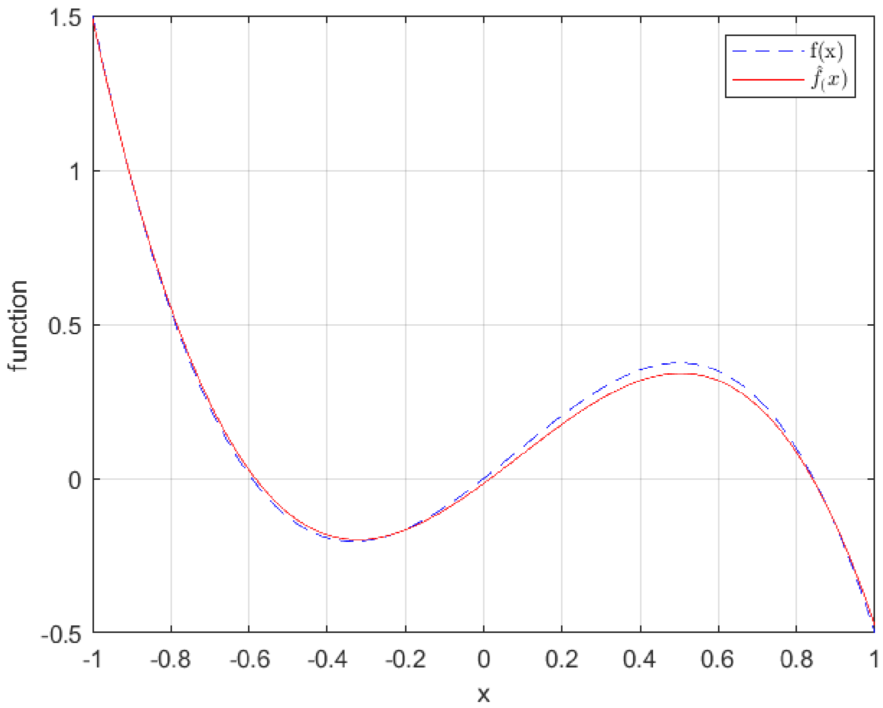
יוצר וקטור x של $m = 10000$ נקודות מדידה הנדגמות באקראי מתוך פילוג אחיד על פני הקטע $[-1,1]$.

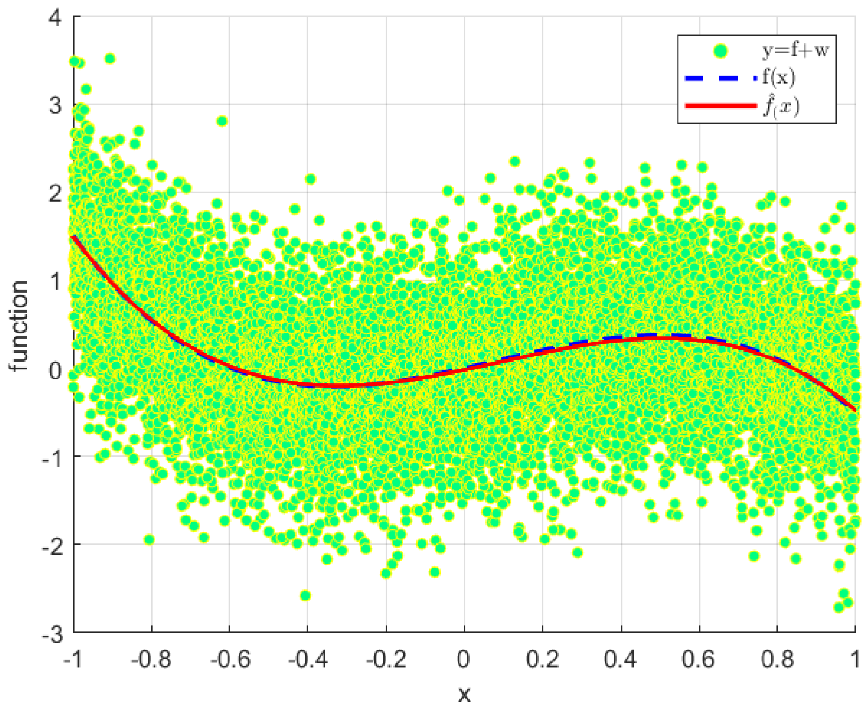
עבור וקטור זה חושב f על פני נקודות מדידה ונוצר וקטור $y = f + w$ ע"י הוספת רעש דגאוסי בעל שונות σ^2.

חלק 1 סעיף 3 +סעיף 4

מתוך *matlab* יוצר גרף שכולל את הוקטורים של :

- פולינום $f(x)$
- וקטר המדידות המורעשות, y
- הפולינום המשוערך $\hat{f}(x)$





Cv

כעת, הניחו כי הפתרון האופטימלי לבעיה (P) מתקבל בתחום $C = \left\{a \in \mathbb{R}^{n+1}: \left\|a\right\|_2 \le r,\ r = 4\right\}$.
הערה: שימו לב כי וקטור המקדמים ה"אמיתי" (זה שבעזרתו מיוצרות המדידות) אכן ב-C.

כעת בעייה מאולצת: $\min_a \left\{h(a) = \frac{1}{2m}\|\mathrm{y} - \mathrm{Xa}\|_2^2\right\}$ $s.t. \|\mathrm{a}\|_2 \le r$ (P') כאשר $D = diamC = 2r$

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

$$G = \frac{1}{m}\left[r \cdot \lambda\_\max(X^T X) + \left\|\mathrm{X^T y}\right\|_2\right]$$

כאשר $\lambda\_\max(X^T X)$ הוא הערך העצמי הכי גדול של $X^T X$

<u>חלק 2</u> סעיף 5+6

כתבו פונקציה המקבלת כקלט את הוקטור x,y וממשת את (P').

- הוגרלו את נקודת ההתחלה מפילוג נורמלי, $a_0 \sim \mathrm{N}(o, I)$
- נקבעו תנאי עצירה מהצורה $\|\nabla h(a)\|_2 \le \epsilon$
- מספר הצעדים הכולל של האלגוריתם מוגבל ע"י תנאי עצירה.

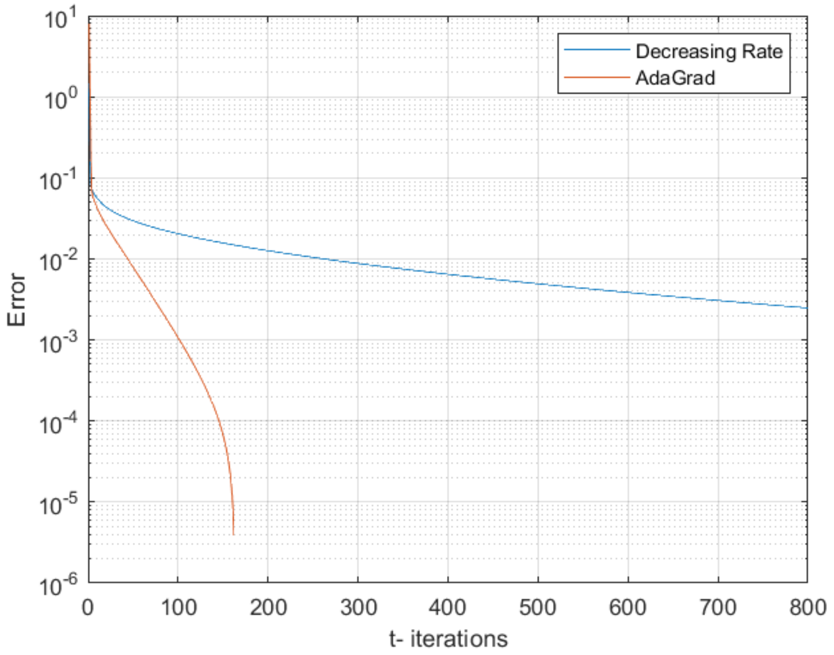
לצורך כתיבת הפונקציה נדרש להיעזר בגרדיאנט של פונקציית מטרה. כלומר:

$$h(a) = \frac{1}{2m}\|\mathrm{y} - \mathrm{Xa}\|_2^2 \qquad g_\tau = \nabla h(a)|_{\partial a} = -\frac{1}{2m}2\mathrm{X}^T(\mathrm{y} - \mathrm{Xa}) = -\frac{1}{m}\mathrm{X}^T(\mathrm{y} - \mathrm{Xa}) = \frac{1}{m}\mathrm{X}^T(\mathrm{Xa} - y)$$

בנוסף, נגדיר בתוך הפונקציה אופציה לבחירת גודל הצעד בין :

- גודל הצעד דועך ... $t = 1,2,3$ $\eta = \frac{D}{G\sqrt{t}}$
- גודל הצעד אדפטיבי... $t = 1,2,3$ $\eta = \frac{D}{\sqrt{2\sum_{\tau=1}^{t}\|g_\tau\|_2^2}}$

הוצגה בגרף פונקצית השגיאה כתלות במספר האיטרציות של האלגוריתם $e_t = h(a_t) - h^*$ עבור שתי בחירות גודל הצעד.



נשים לב כי עבור גודל צעד הדועך ההתכנסות היא יותר איטית ותנאי עצירה שפעל הוא המספר צעדים. לעומת זאת, עבור גודל צעד אדפטיבי, ההתכנסות הרבה יותר מהירה ותנאי העצירה שפעל הוא $\|\nabla h(a)\|_2 \le \epsilon$. בנוסף, עבור גודל הצעד הדועך נקבל שגיאה יותר קטנה. זה תואם את מה שראינו בשיעור לגבי קבצי התכנסות.

סעיף 7

נוכיח זאת מתוך הגדרת חלקות. יהיו $c, b \in R^n$, היות ומדובר בעיה קמורה, לכן, עבור פ' מטרה גזירה ברציפות יתקיים כי

$$\| \ \nabla h(c) - \nabla h(b)\| \le L\| \ \mathrm{c} - \mathrm{b}\|$$

מכיוון ששני הביטויים חיוביים, נבודד את L ונקבל:

$$\frac{\| \ \nabla h(c) - \nabla h(b)\|}{\| \ \mathrm{c} - \mathrm{b}\|} \le L$$

גרדיאנט חושב בסעיפים הקודמים: $\nabla h(a) = -\frac{1}{2m} 2\mathrm{X}^T(\mathrm{y} - \mathrm{Xa}) = -\frac{1}{m}\mathrm{X}^T(\mathrm{y} - \mathrm{Xa})$ נציבו עבור c, b :

$$\frac{\left\| -\frac{1}{m}\mathrm{X}^T(\mathrm{y} - \mathrm{X}c) + \frac{1}{m}\mathrm{X}^T(\mathrm{y} - \mathrm{Xb})\right\|}{\| \ \mathrm{c} - \mathrm{b}\|} \le L$$

$$\frac{\left\| -\frac{1}{m}[\mathrm{X}^T(\mathrm{y} - \mathrm{X}c) - \mathrm{X}^T(\mathrm{y} - \mathrm{Xb})]\right\|}{\| \ \mathrm{c} - \mathrm{b}\|} \le L$$

$$\frac{\left\| -\frac{1}{m}[\mathrm{X}^T\mathrm{y} - \mathrm{X}^T\mathrm{X}c - \mathrm{X}^T\mathrm{y} + \mathrm{X}^T\mathrm{Xb}]\right\|}{\| \ \mathrm{c} - \mathrm{b}\|} \le L$$

נשים לב כי m ערך חיובי

$$\frac{\left\| -\frac{1}{m}[-\mathrm{X}^T\mathrm{X}c + \mathrm{X}^T\mathrm{Xb}]\right\|}{\| \ \mathrm{c} - \mathrm{b}\|} = \frac{\left\| \frac{1}{m}[\mathrm{X}^T\mathrm{X}c - \mathrm{X}^T\mathrm{Xb}]\right\|}{\| \ \mathrm{c} - \mathrm{b}\|} = \frac{\left\| \frac{1}{m}[\mathrm{X}^T\mathrm{X}(c - \mathrm{b})]\right\|}{\| \ \mathrm{c} - \mathrm{b}\|} = \frac{1}{m}\frac{\| \ \mathrm{X}^T\mathrm{X}\|\| \ c - \mathrm{b}\|}{\| \ \mathrm{c} - \mathrm{b}\|} \le L$$

הניחו כי דרגת המטריצה $\mathrm{X}^T\mathrm{X}$ היא $(n+1)x(n+1)$, נסמן אותה ב-A: $\mathrm{A} = \ \mathrm{X}^T\mathrm{X} \in \mathrm{R}^{\mathrm{n+1\ x\ n+1}}$, נגדיר מטריצה $\Lambda = diag(\lambda_1, \lambda_2 \ldots.)$ כך ש - λ_n - הם הערכים העצמיים של A. כפי שראינו מתוך הגדרה של מכפלה פנימית בתרגול, מטריצה A היא מוגדרת חיובית. היא גם סימטרית (זה גורר מסקנה כי כל הע"ע הם ממשים).
לכן ניתן להגדיר מטריצה יוניטרית וממשית $U \in R^{n+1\ x\ n+1}$ כך ש:

$$\Lambda = U^T A U \rightarrow \quad A = U\Lambda U^T$$

מתוך תכונות של נורמה ומטרוצות יוניטריות, (בפרט, מתוך $\| \ U\Lambda\|_n^2 = tr[(U\Lambda) * U\Lambda] = tr[\Lambda * U * U\Lambda] = tr[\Lambda * \Lambda] = \| \ \Lambda\|_n^2$

נקבל $\| \ \mathrm{A}\| = \| \ U\Lambda U^T\| = \| \ \Lambda\| = \left\| \sum_{k=1}^{n+1} \lambda_k\right\|$ ולכן, אם נרצה למקסם את הביטוי, נוכל לקחת את הע"ע המקסימלי במקום כל אחד הע"ע האחרים ונקבל:

$$\| \ \mathrm{X}^T\mathrm{X}\| = \left\| \sum_{k=1}^{n+1} \lambda_k \right\| \le \left\| \sum_{k=1}^{n+1} \lambda_{max} \right\| = \lambda_{\max\ (\mathrm{X}^T\mathrm{X})}$$

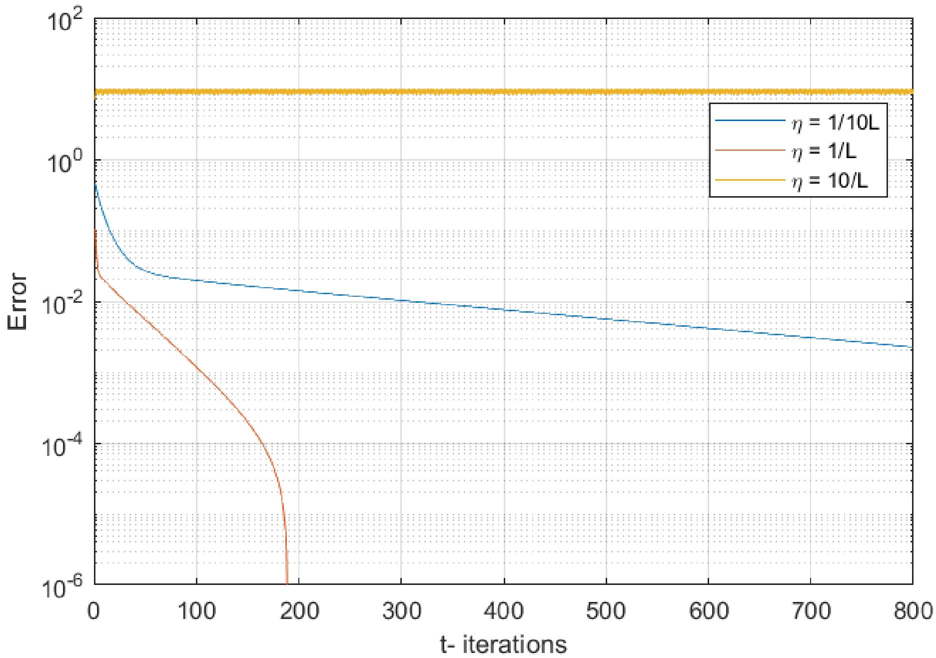
$$\frac{1}{m}\frac{\| \ \mathrm{X}^T\mathrm{X}\|\| \ c - \mathrm{b}\|}{\| \ \mathrm{c} - \mathrm{b}\|} = \frac{1}{m}\| \ \mathrm{X}^T\mathrm{X}\| \le \frac{1}{m}\lambda_{\max(\mathrm{X}^T\mathrm{X})} = L$$

סעיף 8

הריצו את הפונקציה מסעיף 5 עבור שלושת הבחירות הבאות של גודל צעד קבוע:

$$\eta_3 = \frac{1}{10L} = \frac{m}{10\lambda_{\max(X^TX)}}, \quad \eta_4 = \frac{1}{L} = \frac{m}{\lambda_{\max(X^TX)}}, \quad \eta_5 = \frac{10}{L} = \frac{10m}{\lambda_{\max(X^TX)}}$$

שרטטו על גרף אחד, בסקלה לוגריתמית, את פונקצית השגיאה כתלות במספר האיטרציות של האלגוריתם עבור הבחירות הנ"ל.

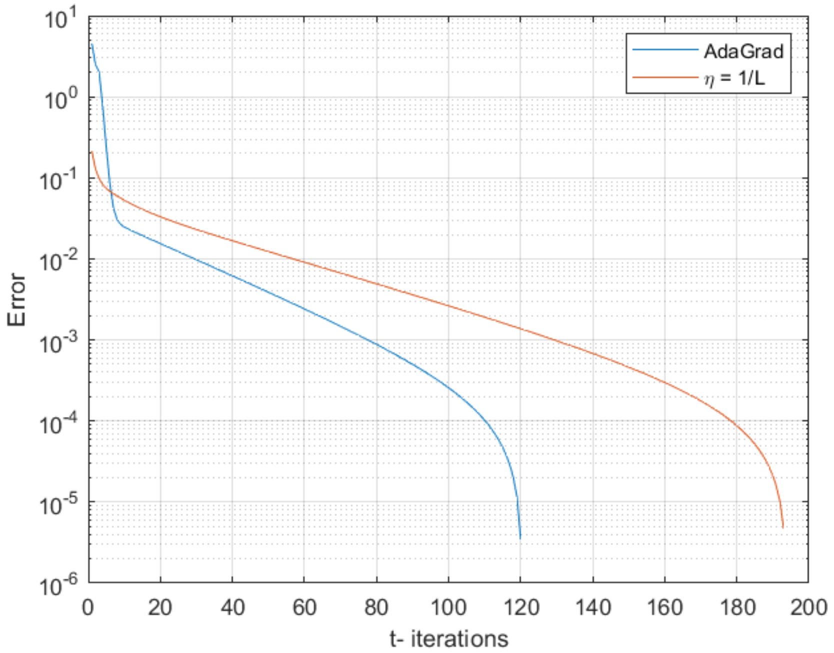


כפי שניתן לראות בגרף הזה, בברור גודל הצעד משפיע על גודל השגיאה וגם על הקצת בבתכנסות.

עבור צעד גדול מדי , $\eta_5 = \frac{10}{L}$ ניתן לראות כי השיטה לא מתכנסת אלה שומרת על השגיאה באיזור 10^1. לעומת זאת, עבור $\eta_3 = \frac{1}{10L}$ ההתכנסות הינה איטית (עבור כ1200 איטרציות נדרשות כדי להגיע לשגיאה של 10^{-3}, וכדי להגיע לשגיאה של 10^{-5} כ-1800 איטרציות). גודל הצעד האופטימלי מבין שלושת הנתונים הוא $\eta_4 = \frac{1}{L}$.
חשוב לציין כי זה תואם את התאוריה שראינו בתרגול 6 כי פ' מטרה שלנו קמורה ו-Lחלקה, גודל הצעד הוא $\frac{1}{L}$ עבור שיטת גרדיאנט המוטל.

סעיף 9

בוצעה השוואה בין גודל הצעד אדפטיבי... $t = 1,2,3$ $\eta = \frac{D}{\sqrt{2\sum_{\tau=1}^{t}\|g_\tau\|_2^2}}$ וגודל הצעד $\eta_4 = \frac{1}{L}$.



כפי שראינו בתרגול , עבור גודל הצעד אדפטיבי If $\|\mathbf{g}_t\|_2 \le G$, same convergence rate of $O(1/\sqrt{t})$

וגם, עבור גודל הצעד הקבוע:

דף עזר: קצבי התכנסות של אלגוריתמים

Algorithm	f	Step size (η_t)	Convergence rate ($O(\cdot)$)
PGD	smooth	$1/L$	LD^2/t

ניתן גם לראות כי ההתכנסות של גודל הצעד אדפטיבי היא טובה יותר (עבורה נקבל כי γ הוא גדול יותר).

תואם את התאוריה.

<u>חלק 3</u> סעיף 10-11

כתבו פונקציה המקבלת כקלט את הוקטורי x, y ופרמטר $b \in \{1, \ldots .. m\}$ (ייתכן 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 מדידות.

כלל העדכון של האלגוריתם הינו מהצורה: $a_{t+1} = P_C(a_t - \eta_{t+1} g_t)$

עבור גודל הצעד דועך ... $t = 1,2,3$ $\eta = \frac{D}{G\sqrt{t}}$

הריצו את הפונקציה הנ"ל עבור הבחירות הבאות של `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.

בניתי את הפונקציה, אבל יש איתה בעיה בקוד, כי לא מצליחה לקבל את ההתנהגות הרצויה ולא הספקתי למצוא איפה בעיה.

מה שכן, עבור `b=10000` מתקבל תוצאה תואמת את התוצאות שהתקבלו בריצות של אלגוריטם הגרדיאט.

בכל מקרה אכניס את הגרפים: