חלק 1:

1. מצא מטריצה $X \in R^{m\times n+1}$ כך שהבעיות שקולות:

פתרון:

נגדיר $X = \begin{pmatrix} x_1^0 & x_1^1 & \dots & x_1^n \\ x_2^0 & x_2^1 & \dots & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ x_m^0 & x_m^1 & \dots & x_m^n \end{pmatrix}$ כלומר האיבר ה- i,j הוא **הדגימה ה -** i **בחזקת** j .

נשים לב כי: $Xa = \begin{pmatrix} x_1^0 & x_1^1 & \dots & x_1^n \\ x_2^0 & x_2^1 & \dots & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ x_m^0 & x_m^1 & \dots & x_m^n \end{pmatrix}\begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} x_1^0\cdot a_0 + x_1^1\cdot a_1 + \dots + x_1^n \cdot a_n \\ \vdots \\ x_m^0\cdot a_0 + x_m^1\cdot a_1 + \dots + x_m^n \cdot a_n \end{pmatrix} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{pmatrix}$

$\Rightarrow (Xa)_i = f(x_i)$

לכן: $h(a) = \frac{1}{2m}\|y - Xa\|_2^2 \underset{\text{L2 Norm}}{=} \frac{1}{2m}\sum_{i=1}^{m}\left(y_i - (Xa)_i\right)^2 = \frac{1}{2m}\sum_{i=1}^{m}\left(y_i - f(x_i)\right)^2$ והבעיות שקולות עבור המטריצה הזאת.

2. נראה כעת כי $(P) = \min_{a\in R^{n+1}}\left\{h(a) = \frac{1}{2m}\sum_{i=1}^{m}\left(y_i - (Xa)_i\right)^2 = \frac{1}{2m}\sum_{i=1}^{m}\left(y_i - f(x_i)\right)^2\right\}$ קמורה,

ונמצא פתרון סגור.

פתרון:

$$h(a) = \frac{1}{2m}\|y - Xa\|_2^2 = (y - Xa)^T(y - Xa) =$$
$$\frac{1}{2m}\left(y^T - a^TX^T\right)(y - Xa) = \frac{1}{2m}\Big(y^Ty - \underbrace{a^TX^Ty}_{scalar} - \underbrace{y^TXa}_{scalar} + a^TX^TXa\Big) = \frac{1}{2m}\left(\|y\|_2^2 - \left(a^TX^Ty\right) - \left(a^TX^Ty\right)^T + a^TX^TXa\right) =$$
$$\frac{1}{2m}\left(\|y\|_2^2 - 2\left(a^TX^Ty\right) + a^TX^TXa\right)$$
$$\nabla h(a) = \frac{1}{2m}\left(-2X^Ty + 2X^TXa\right)$$
$$\nabla^2 h(a) = \frac{1}{2m}\left(2X^TX\right) = \frac{1}{m}\left(X^TX\right)$$

X^TX is PSD (by Homework 1) and $m > 0 \Rightarrow \nabla^2 h(a)$ is PSD

והפתרון הסגור:

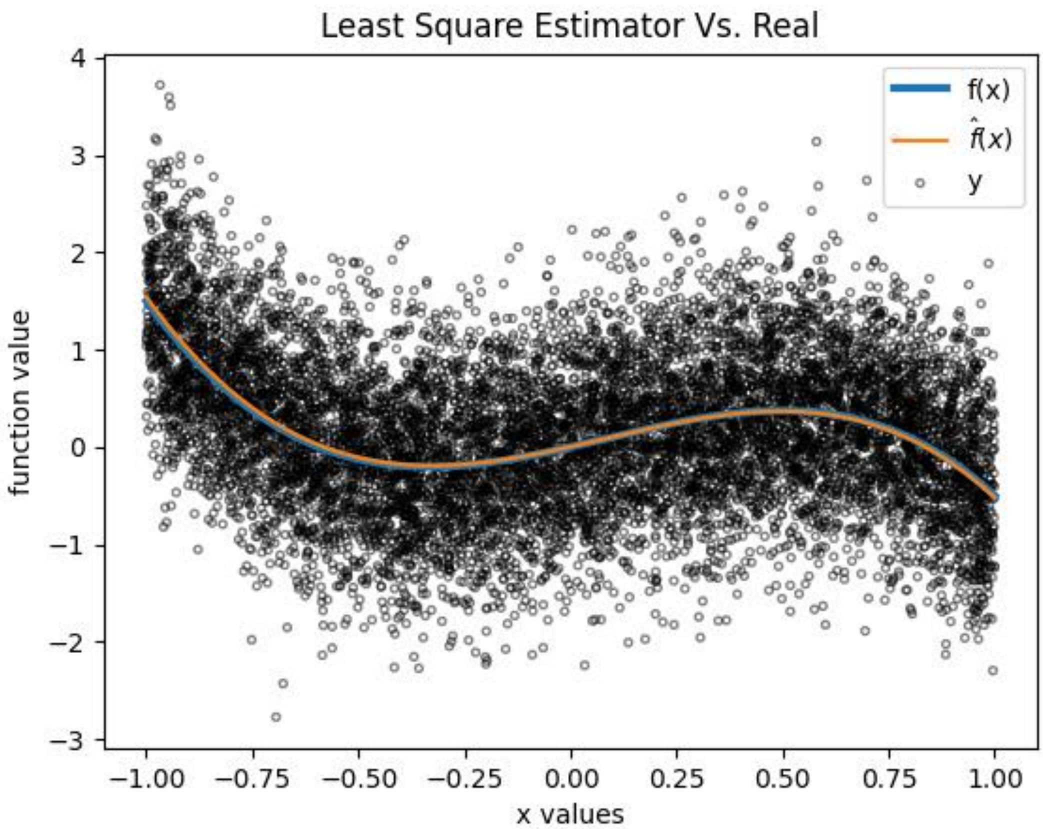
$$\nabla h(a) = \frac{1}{2m}\left(-2X^Ty + 2X^TXa\right) = 0$$
$$\Leftrightarrow X^TXa = X^Ty$$
$$\Leftrightarrow a^* = \left(X^TX\right)^{-1}X^Ty$$

נשים לב כי הפתרון קיים אמ"מ X^TX הפיכה.

3. המקדמים שהפונקציה שלנו החזירה הם:

0	1	2	3
0.00294	0.97805	0.50909	-2.00785

4. **<u>Least Square Estimator Vs. Real:</u>**



ניתן לראות מהגרף, כי הדגימות הרועשות y מופלגות בצורה גאוסית סביב הפונקציה המקורית, כלומר הסבירות לדגימה רועשת ממש (רחוקה מאוד מהפונקציה) קטנה (יש מעט דגימות שכאלה).

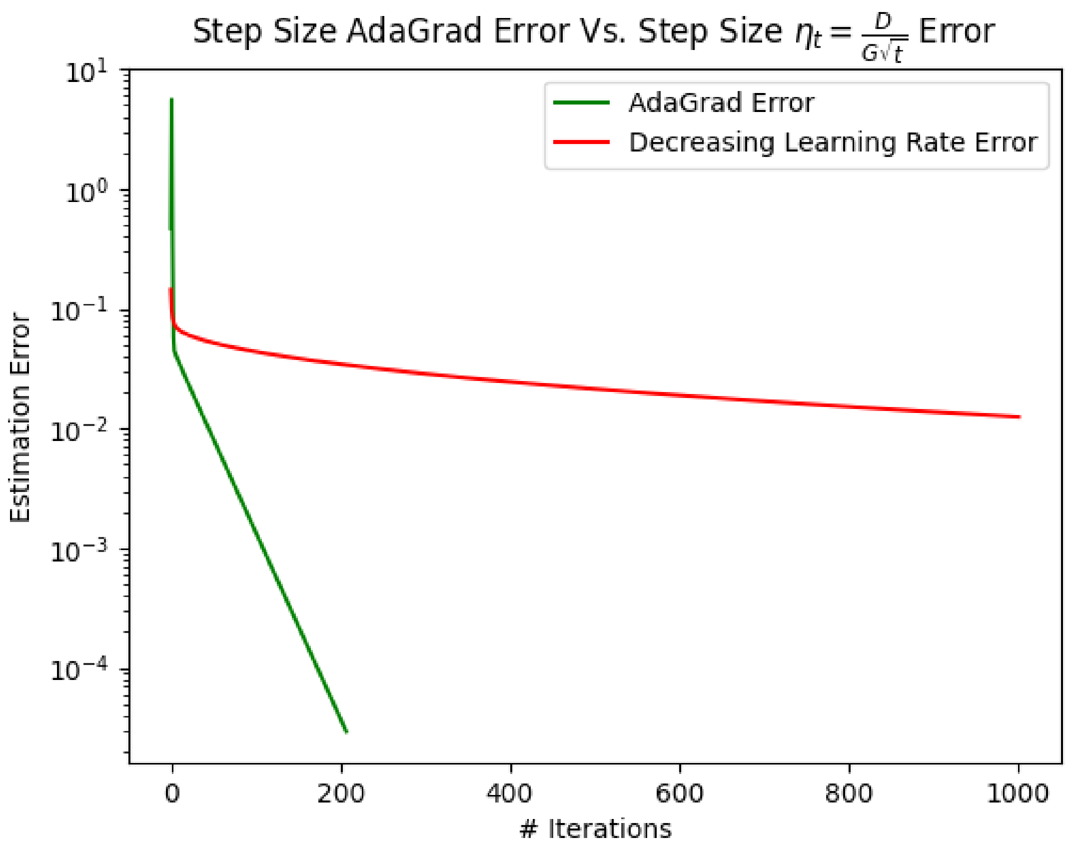
בנוסף ניתן לראות שלמרות שהדגימות שלנו רועשות אכן שערוך שמצאנו משערך היטב את הפונקציה.

חלק Projected Gradient Descent 2-

5. מהחלק הקודם של השאלה ראינו כי:

$$h(a)=\frac{1}{2m}\|y-Xa\|_2^2;\nabla h(a)=\frac{1}{2m}\left(-2X^Ty+2X^TXa\right)\cdot$$

6. השוואה בין ה גודל הצעד כאשר הוא AdaGrad או $\eta_t=\frac{D}{G\sqrt{t}}$:



הרצת אלגוריתם *Projected gradient descent* המבוקש עם פרמטר סף של $\varepsilon=0.001$.

כפי שניתן לראות מהגרף, עבור גודל צעד דועך האלגוריתם עדיין לא התכנס לטווח השגיאה הרצוי כי ראינו שלצורך כך דרושות $O\left(\frac{1}{\varepsilon^2}\right)$ איטרציות, כלומר סדר גודל של 1,000,000 איטרציות. לעומת זאת, כאשר גודל הצעד הוא AdaGrad הביצועים הם הרבה יותר טובים, אנו משיגים שגיאה קטנה מהסף בהרבה פחות איטרציות. לפי מה שראינו בתרגול 6, AdaGrad יכול להשיג ביצועים הרבה יותר טובים כאשר הנורמה של הגראדינטים קטנה משמעותית מ-G.

7. צריך להוכיח כי קבוע החלקות של פונקציית המטרה $h(a)$ הוא: $L=\lambda_{\max}\left(\frac{X^TX}{m}\right)$.

<u>הוכחה:</u> נראה כי הפונקציה $h(a)$ היא בעלת גראדינט ליפשצי עם קבוע $L=\lambda_{\max}\left(\frac{X^TX}{m}\right)$:

$a_1, a_2 \in C$

$\left\|\nabla h(a_1)-\nabla h(a_2)\right\|_2^2=\frac{1}{2m^2}\left\|-2X^Ty+2X^TXa_1+2X^Ty-2X^TXa_2\right\|_2^2=\frac{1}{2m^2}\left\|2X^TXa_1-2X^TXa_2\right\|_2^2=\left\|\frac{X^TX}{m^2}a_1-\frac{X^TX}{m^2}a_2\right\|_2^2$

$let\ Z=\frac{X^TX}{m^2}; Z \in R^{n+1\times n+1}$

$\left\|Za_1-Za_2\right\|_2^2=\left\|Z(a_1-a_2)\right\|_2^2=\left(Z(a_1-a_2)\right)^T Z(a_1-a_2)=(a_1-a_2)^T Z^TZ(a_1-a_2)\underset{Tutorial\_5}{\leq}\lambda_{\max}\left(Z^TZ\right)\left\|a_1-a_2\right\|_2^2$

נסמן ב- $L=\lambda_{\max}\left(\frac{X^TX}{m}\right)$, כמו כן $L \geq 0$ כי זאת מטריצה PSD ובפרט,

$$\lambda_{\max}\left(Z^TZ\right)=\lambda_{\max}\left(\frac{X^TX}{m}\times\frac{X^TX}{m}\right)=L\cdot L=L^2$$

זאת מכיוון ש $L-$ הוא ערך עצמי של $\frac{X^TX}{m}$ ולכן עם אותו וקטור עצמי, אזי L^2 הוא גם ערך עצמי של $\frac{X^TX}{m}\times\frac{X^TX}{m}$ והוא גם מקסימלי עבור מטריצה זו אחרת זו סתירה ש- L מקסימלי עבור המטריצה $\frac{X^TX}{m}$.

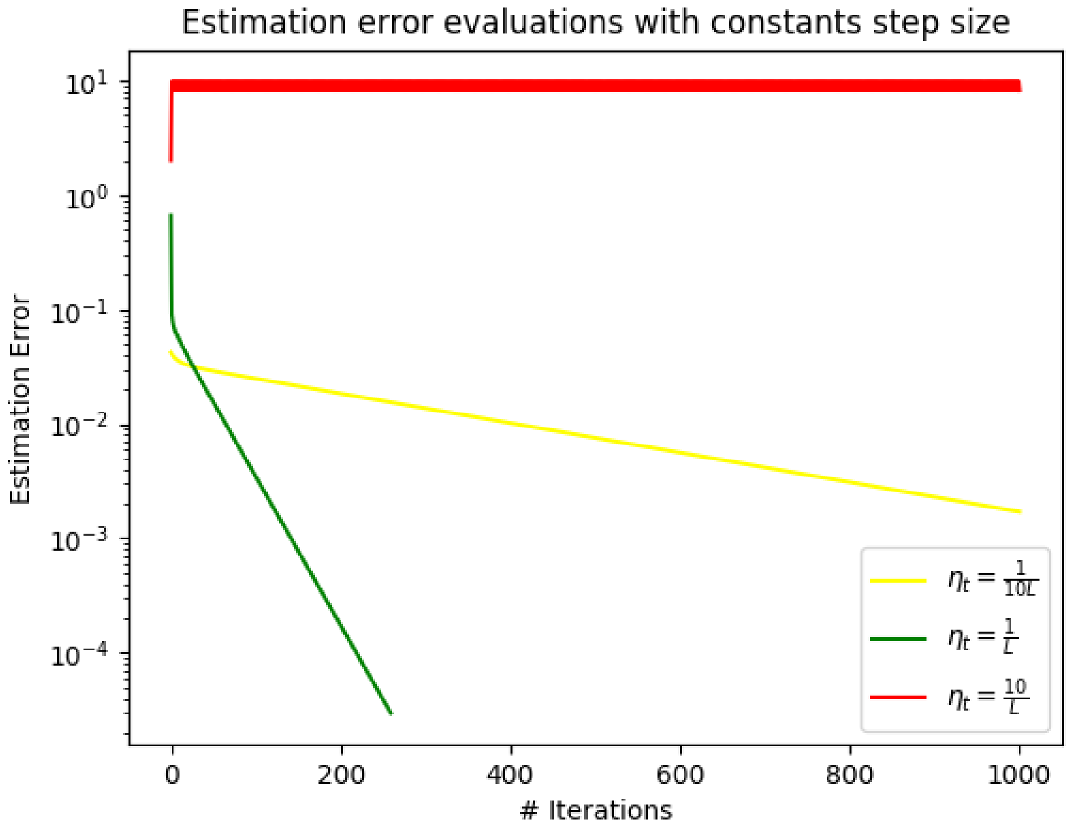
(מכיוון שהדרגה של X מלאה הרי שהדטרמיננטה שלה שונה מאפס ולכן המטריצה $\frac{X^TX}{m}$ היא מוגדרת חיובית ממש ולכן יש לה n+1 ערכים עצמים שונים, כי ניתנת ללכסון, וכל ערך עצמי שלה חיובי ואם נעלה אותו בריבוע הוא גם ערך עצמי של $\frac{X^TX}{m}\times\frac{X^TX}{m}$ ולכן הערך העצמי המקסימלי של $\frac{X^TX}{m}$ אם נעלה אותו בריבוע הוא המקסימלי של $\frac{X^TX}{m}\times\frac{X^TX}{m}$.

לכן:

$$\begin{aligned}&a_1, a_2 \in C\\ &\left\|\nabla h(a_1)-\nabla h(a_2)\right\|_2^2 \leq L^2\left\|a_1-a_2\right\|_2^2 \quad .\\ &\Rightarrow \left\|\nabla h(a_1)-\nabla h(a_2)\right\|_2 \leq L\left\|a_1-a_2\right\|_2\end{aligned}$$

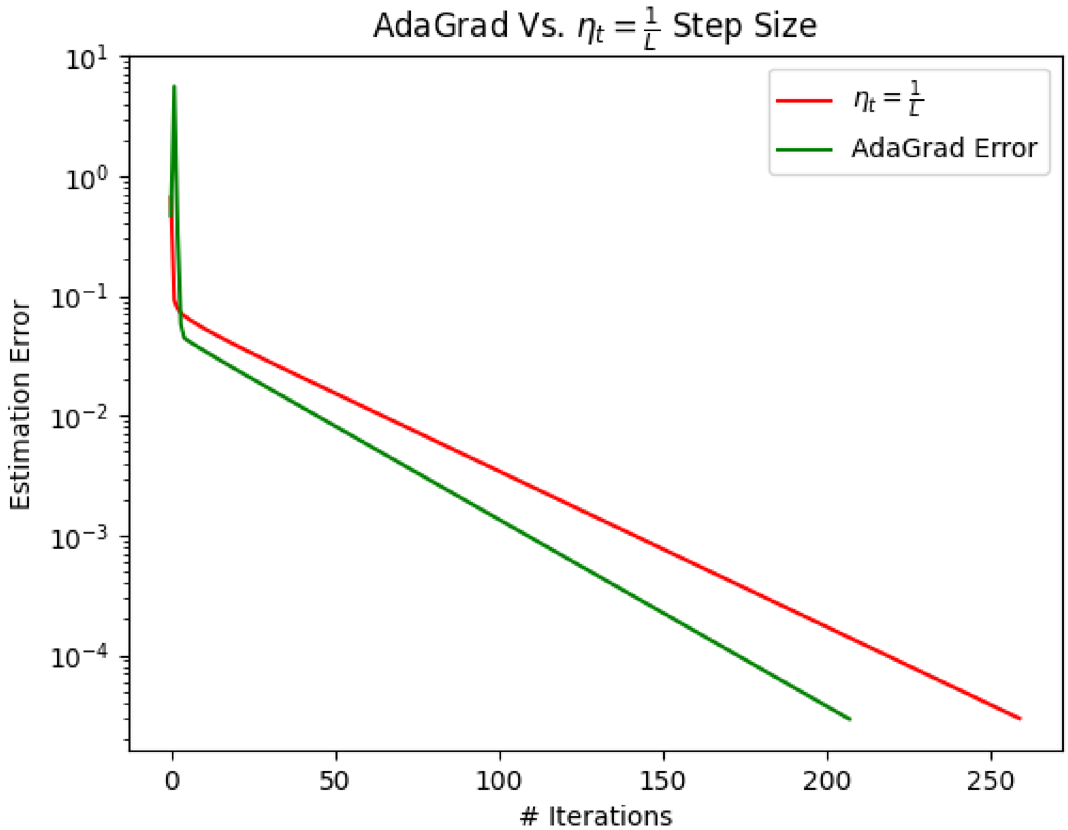
ומכיוון שהפונקציה קמורה וגזירה ברציפות ובעל גראדינט ליפשצי עם קבוע L אזי היא L חלקה.

8. **:Constant Step Size**



מכיוון שאנו מניחים כי דרגת המטריצה X היא $n+1$ הרי הדטרמיננטה שלה שונה מאפס בוודאות ומכיוון ש $\nabla^2 h(a) = \frac{1}{m}\left(X^T X\right)$, אזי ההיסאן של פונקציית המטרה בעל דטרמיננטה שונה מאפס בוודאות ולכן בהכרח $0 \succ$ $\nabla^2 h(a) = \frac{1}{m}\left(X^T X\right)$ כלומר פונקציית המטרה היא strongly convex ו L חלקה אזי כפי ראינו בכיתה נוכל לבחור בגודל צעד קבוע השווה ל- $\frac{1}{L}$, **וקצב ההתכנסות הוא מעריכי במספר איטרציות.** בנוסף, ניתן לראות בגרף שגודל צעד קטן מידי מוביל להתכנסות איטית יותר וגודל צעד גדול לא מוביל להתכנסות.

.9

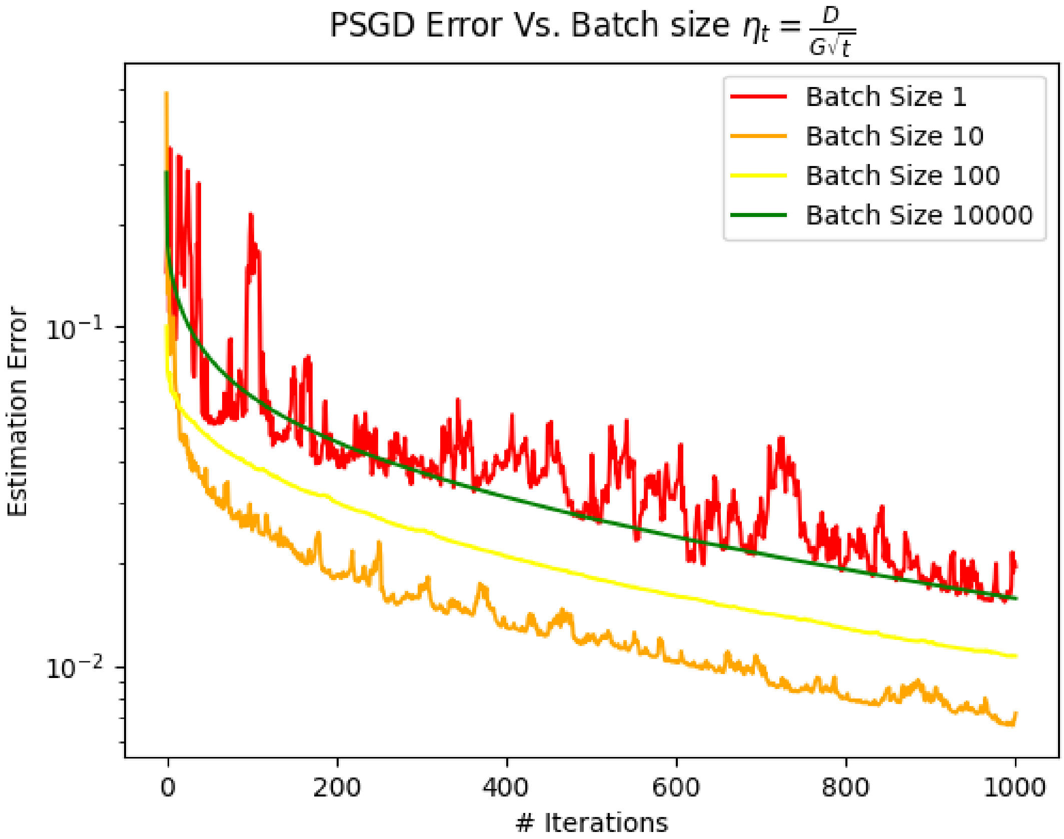


כפי שניתן לראות מהגרף שני גדלי הצעד משיגים קצב התכנסות אקספוננציאלי זאת מכיוון שהגרף הוא לינארי בסקלה לוגריתמית.

עבור גודל הצעד הקבוע ראינו והוכחנו זאת בתרגול. לעומת זאת, בתרגול הוכחנו כי עבור המקרה הגרוע ביותר קצב ההתכנסות של גודל צעד *AdaGrad* הוא $O\left(\frac{1}{\sqrt{t}}\right)$. אמנם, מה שמייחד את *AdaGrad* שהוא אדפטיבי והוא גורם לגודל הצעד להתאים את עצמו לתהליך עצמו על ידי שימוש בגראדינטים. לכן, הוא עבור מקרה זה גם מושג קצב מעריכי.

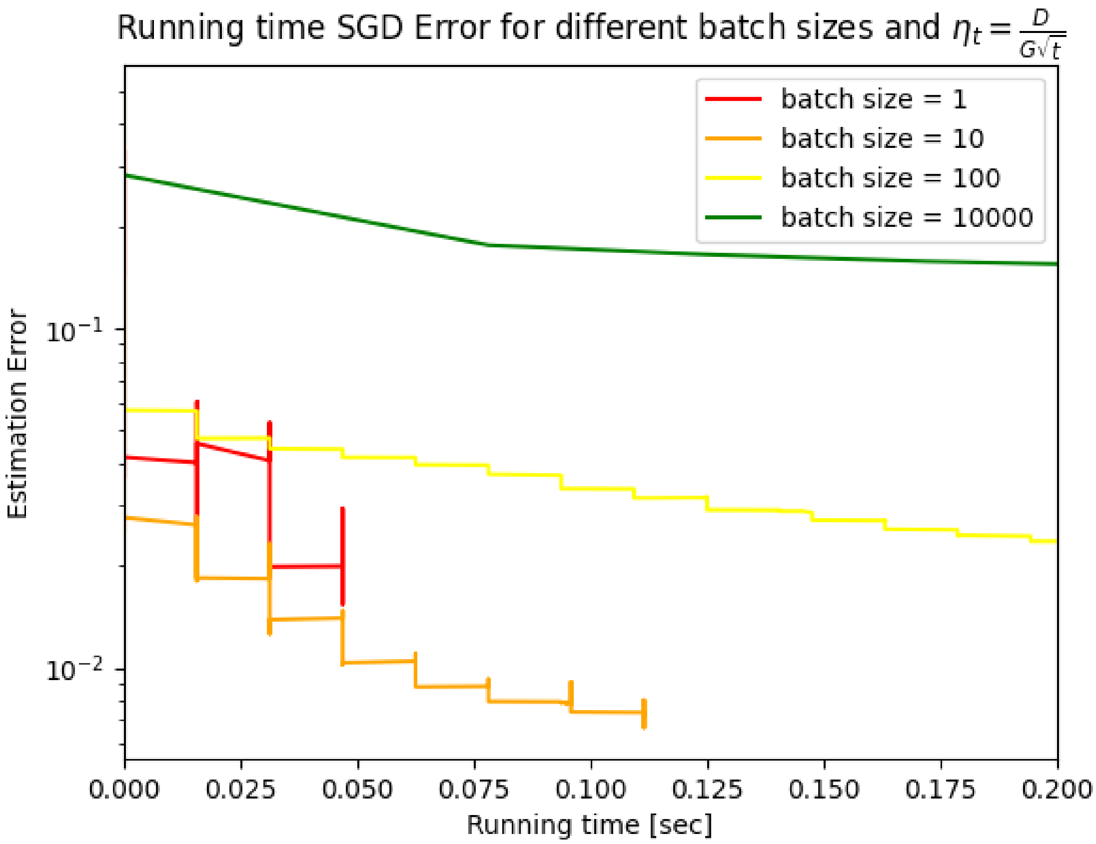
חלק 3 Projected Stochastic Gradient Descent-

.10



ניתן להבין מהגרף כי אלגוריתם ה-SGD הוא יותר רועש מהאלגוריתם ה-GD, זאת מכיוון שהגראדינט נמדד רק על פה כמה מדידות ולא על פי כולן לכן לא בהכרח שבכל איטרציה אנחנו נתקדם לעבר המינימום של הפונקציה הכוללת. כמו שעוד ניתן לשים בכל שאנחנו דוגמים יותר דגימות יחד, כלומר batch גדול יותר התהליך פחות רועש כי כיוון הגרדיאנט יותר נכון.
נשים לב, למרות שישנה צפייה כי כאשר אנו דוגמים יותר אז ההתכנסות תהיה מהירה יותר, זה אמנם לא תמיד כך (הדבר משתנה בין הריצות) וזאת מכיוון שההתכנסות תלויה גם בווריאנס של הגראדינט הסטוכטי.

.11



בגרף הזה אנחנו רואים בבירור את היתרון של SGD ביחס לGD. נכון שביחס לכמות איטרציות batch גדול יותר עשוי להתכנס מהר יותר. אמנם, אם נסתכל ביחס **לזמן** אנחנו רואים כי ככל ש- batch גדול כמו כל הדגימות, אזי השגיאה יורדת בקצב יותר איטי וזאת מכיוון **שבכל איטרציה** עבור batch גדול יותר לוקחת יותר זמן (חישוב הגראדינט בכל הדגימות).

וזה החיסרון של batch גדול יותר, **הזמן בפועל** של ההתכנסות יותר איטי כי כל איטרציה לוקחת יותר זמן, יותר כבדה חישובית, והיתרון הוא שההתכנסות תהיה יותר חלקה ופחות רועשת.