29.12.2022 **פרטיות מידע**

Lecture 10: Projected Gradient Decent

Textbook: Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy

מרצה: אורי שטמר

בפעמים הקודמות הכרנו את בעיית ה ERM (עבור פונק' הפסד פריקה) וראינו אלג' גנרי המבוסס על המ.אקספ. האלג' הזה לא תמיד יעיל חישובית (וגם במקרה הכי טוב זמן הריצה שלו אולי פולינומי אבל לא באמת פרקטי להריץ אותו...). היום נראה רעיון אחר שמבוסס על gradient decent שהוא הרבה יותר פרקטי.

הגדרת הבעיה:

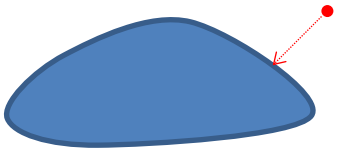
קלט:

- מרחב פתרונות $C \subseteq \mathbb{R}^d$
- דטהבייס $S = (x_1, \dots, x_n) \in D^n$
- פונקציית הפסד $\ell: C \times D \to \mathbb{R}$

הנחה על C: ניתן "להטיל" (project) נקודות על C ביעילות, כאשר ההטלה של נקודה $x \in \mathbb{R}^d$ על C היא:

$$\Pi_C(x) = \underset{w \in C}{\operatorname{argmin}} \|x - w\|$$

לדוגמה:

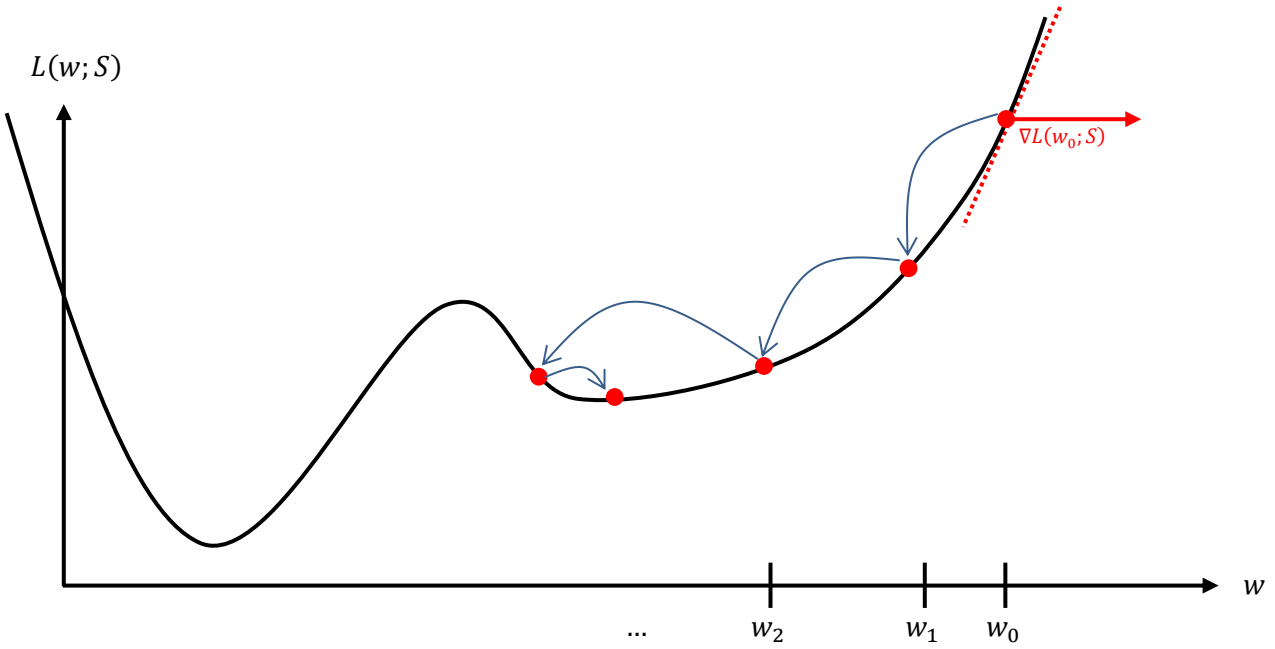


הערה: הטלות כאלה אפשר לחשב ביעולות בהרבה מקרים, למשל כאשר C הוא כדור היחידה, או כאשר $C = [-1,1]^d$ או כאשר $C = \mathbb{R}^d$.

יש למצוא: $\hat{w} \in C$ המקיים $L(\hat{w}; S) \approx \min_{w \in C} L(w; S)$

(כאשר $L(w; S) = \frac{1}{n} \sum_{x \in S} \ell(w; x)$)

הרעיון של *gradient decent* הוא להתחיל מאיזושהי נקודה שרירותית $w_0 \in C$ ואז בכל שלב נחשב את הגרדיאנט בנקודה הנוכחית (מזה אנחנו לומדים את השיפוע של הפונקציה בנקודה הנוכחית בכל הכיוונים) ואז נתקדם מהנקודה הנוכחית בכיוון שבו הפונקציה קטנה הכי מהר, שזה הכיוון המנוגד לכיוון הגרדיאנט.



בנוסף, מכיוון שאנחנו מתעניינים רק בפתרונות $w \in C$, אז בכל פעם אחרי שנעשה צעד בכיוון המנוגד לגרדיאנט אנחנו נטיל את הנקודה שהגענו אליה בחזרה לקבוצה C כדי לדאוג שבכל שלב הנקודה w שאנחנו בוחנים כרגע היא פתרון חוקי.

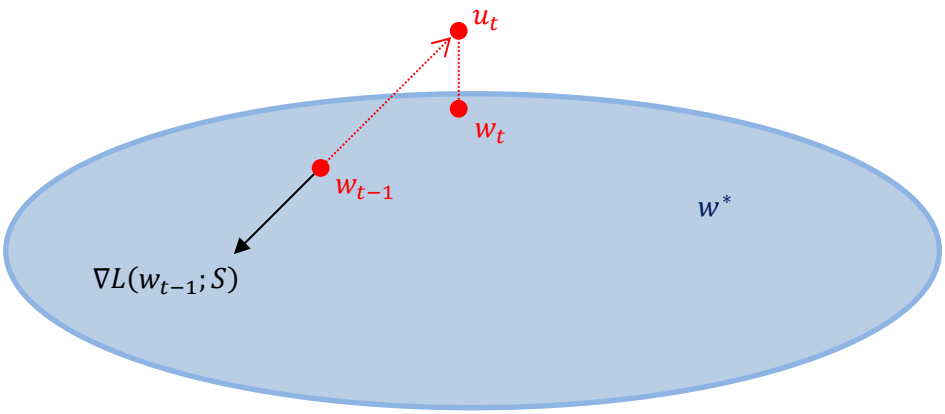
אז אנחנו מקבלים את האלגוריתם הבא:

Projected Gradient Decent (PGD)

Input: *Set $C \subseteq \mathbb{R}^d$ for which the projection Π_C is easy to compute.*
Database $S = (x_1, \dots, x_n) \in D^n$
Decomposable loss function $L: C \times D^n \to \mathbb{R}$ where $L(w;S) = \frac{1}{n}\sum_i \ell(w;x_i)$
Learning rate (or step size) $\eta > 0$

(1) $w_0 \leftarrow$*arbitrary point in C*
(2) For $t = 1,2,\dots,T$ *do*
  a. $g_t \leftarrow \nabla L(w_{t-1};S) = \frac{1}{n}\sum_i \nabla \ell(w_{t-1};x_i)$
  b. $u_t \leftarrow w_{t-1} - \eta \cdot g_t$
  c. $w_t \leftarrow \Pi_C(u_t)$
(3) Return $\widehat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$

הערה: יש הרבה גרסאות לאלגוריתם הזה ומה שהצגנו כאן זאת לא הדרך היחידה. בפרט, בסיום הריצה אנחנו כאן מחזירים את הממוצע של כל הנקודות שעברנו דרכם בתהליך. זאת אופציה פשוטה אבל לא בהכרח הכי טובה. לפעמים כדאי לקחת את הנקודה האחרונה שהגענו אליה.



<u>זמן ריצה:</u> כמו שהאלגוריתם כתוב, הוא דורש זמן $O\left(T\left(n(\text{Time}(\nabla\ell)+d)+\text{Time}(\Pi_C)\right)\right)$ כי בשלב *2a* אנחנו מחשבים בכל איטרציה את כל n הגרדיאנטים ואז ממצעים את ווקטורי הגרדיאנט שהם ווקטורים ב $\mathbb{R}^d$ ובנוסף בכל איטרציה אנחנו מחשבים הטלה.

אז אם מספר האיטרציות T וגודל הדטהבייס n הם לא יותר מדי גדולים, זה כבר נותן אלגוריתם בזמן סביר.
בהמשך נדבר על איך אפשר ליעל את האלגוריתם הזה עוד יותר.

תכונה חשובה של *GD*: מסתבר שהאלגוריתם הזה מאוד חסין לשינויים. בפרט, האלגוריתם בד"כ ממשיך לעבוד טוב גם אם במקום לחשב את הגרדיאנטים במדוייק בכל שלב, נחשב רק <u>הערכה</u> שלהם.

למה זה טוב לנו? זה יאפשר לנו לתכנן גרסה פרטית של האלגוריתם הזה. הנה ניסיון ראשון (בהמשך נראה איך אפשר לשפר).

Noisy Projected Gradient Decent (Noisy PGD)

Input: *Set* $C \subseteq \mathbb{R}^d$ *for which the projection* Π_C *is easy to compute.*
Database $S = (x_1, \dots, x_n) \in D^n$
Decomposable loss function $L: C \times D^n \to \mathbb{R}$ *where* $L(w;S) = \frac{1}{n}\sum_i \ell(w;x_i)$ *and where* ℓ *is G-Lipchitz*
Learning rate (or step size) $\eta > 0$

(1) $w_0 \leftarrow$*arbitrary point in* C
(2) For $t = 1,2,\dots,T$ *do*
  a. $g_t \leftarrow \nabla L(w_{t-1};S) = \frac{1}{n}\sum_i \nabla\ell(w_{t-1};x_i)$
  b. $\tilde{g}_t \leftarrow g_t + \left(\text{Lap}(b)\right)^d$ *for* $b = \Theta\left(\frac{G\sqrt{T\cdot d\cdot\log(1/\delta)}}{n\varepsilon}\right)$
  c. $u_t \leftarrow w_{t-1} - \eta\cdot\tilde{g}_t$
  d. $w_t \leftarrow \Pi_C(u_t)$
(3) Return $\widehat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$

<u>טענת עזר 1:</u>
אם פונקציה f היא קמורה וגם G-ליפשיץ אזי לכל x מתקיים $\|\nabla f(x)\| \leq G$.

<u>הוכחה:</u>
נשתמש בהגדרה האלטרנטיבית לקמירות של פונקציה (ראיתם בתרגול), כאשר נסתכל על שתי הנקודות x , $x + \nabla f(x)$. נקבל:

$$G \cdot \|\nabla f(x)\| \geq f\big(x + \nabla f(x)\big) - f(x) \geq \langle \nabla f(x), \nabla f(x) \rangle = \|\nabla f(x)\|^2$$

ולכן

$$\|\nabla f(x)\| \leq G$$

מ.ש.ל. (טענת עזר 1)

<u>ניתוח פרטיות (סקיצה):</u>
הרעיון בניתוח הפרטיות מבוסס על האבחנה שמספיק להראות שכל הגרדיאנטים הרועשים שמחושבים לאורך הריצה הם תוצאה של חישוב פרטי. אכן מספיק להראות זאת משום שאת שאר הדברים באלגוריתם ניתן לראות כ *post-processing* של הגרדיאנטים הרועשים האלה.
כעת, ישנן T שלבים בריצה ובכל שלב אנחנו מחשבים גראדירנט שהמורכב מ d חישובי נגזרות. לפי טענת העזר הנ"ל, תחת ההנחה שהפונקציה ℓ היא G-ליפשיץ, אז לכל חישוב נגזרת כזה יש רגישות $\frac{2G}{n}$.
אז בסה"כ ישנן Td הפעלות של המ.לפלאס לאורך הריצה. פרטיות נובעת לכן לפי משפט הקומפוזיציה (ולפי בחירת הפרמטר b בשלב $2b$ באלגוריתם).

אוקיי, אז האלגוריתם הזה משמר פרטיות. מתי הוא טוב?
נתחיל מניתוח פשוט של האלגוריתם הלא פרטי ואח"כ נראה איך להתאים את הניתוח גם לאלגוריתם הפרטי.

<u>משפט 1:</u> נניח שפונקציית ההפסד $L: C \to \mathbb{R}$ היא קמורה וגם G-ליפשיץ. בנוסף נניח ש C היא קבוצה סגורה וקמורה עם קוטר R. נסמן $w^* = \underset{w \in C}{\operatorname{argmin}} L(w)$. עבור $\eta = \frac{R}{G\sqrt{T}}$ מתקיים שאלג' PGD מחזיר נקודה $\widehat{w}$ המקיימת

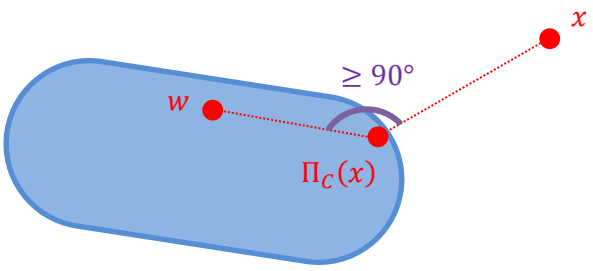
$$L(\widehat{w}) - L(w^*) \leq \frac{RG}{\sqrt{T}}$$

לצורך ההוכחה נשתמש בטענת העזר הבאה:

<u>טענת עזר 2:</u> אם $C \subseteq \mathbb{R}^d$ היא קבוצה סגורה וקמורה אז לכל $x \in \mathbb{R}^d$ ולכל $w \in C$ מתקיים

$$\|\Pi_C(x) - w\| \leq \|x - w\|$$

<u>הוכחה בציור לטענת עזר 2:</u>



<u>רעיון ההוכחה של משפט 1:</u> אנחנו נעקוב במקביל אחרי שני הביטויים הבאים:

- $L(w_t) - L(w^*)$ (זהו ה *excess risk*)
- $\|w_t - w^*\|^2$ (זהו המרחק בריבוע לנקודת האופטימום)

מה שאנחנו נטען זה ששני הביטויים האלה מתכווצים לאורך הריצה של האלגוריתם ונשתמש באחד מהם כדי לשלוט על השני.

<u>טענת עזר 3 ("טענת התקדמות"):</u>

לכל $t = 0,1,\dots,T-1$ מתקיים:

$$L(w_t) - L(w^*) \le \frac{\eta\|g_{t+1}\|^2}{2} + \frac{1}{2\eta}(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2)$$

מה הטענה הזאת אומרת לי?
אם בשלב מסויים $L(w_t) - L(w^*)$ הוא גדול, אז גם $(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2)$ צריך להיות גדול, מה שאומר שבשלב הבא באלגוריתם אנחנו מתקרבים משמעותית לנקודת האופטימום.
אינטואיטיבית זה אומר ש $L(w_t) - L(w^*)$ לא יכול להיות גדול יותר מדי פעמים לאורך הריצה, כי זה יגיד שאנחנו שוב ושוב מתקרבים משמעותית לאופטמייזר w^* אבל התהליך הזה חייב לעצור כי בסוף נגיע אליו.

<u>הוכחת טענת עזר 3:</u>

אנחנו מניחים שפונקציית ההפסד L היא קמורה. לפי ההגדרה השקולה לקמירות של פונקציה (מהתרגול), אנחנו מקבלים

$$L(w^*) \ge L(w_t) + \langle \nabla L(w_t), w^* - w_t\rangle = L(w_t) + \langle g_{t+1}, w^* - w_t\rangle$$

לכן

$$L(w_t) - L(w^*) \le \frac{1}{\eta}\langle \eta\, g_{t+1}, w_t - w^*\rangle$$

$$\underbrace{\le}_{(**)} \frac{1}{2\eta}\left(\|\eta\, g_{t+1}\|^2 + \|w_t - w^*\|^2 - \underbrace{\|w_t - w^* - \eta g_{t+1}\|^2}_{u_{t+1}-w^*}\right)$$

$$= \frac{\eta}{2}\|g_{t+1}\|^2 + \frac{1}{2\eta}(\|w_t - w^*\|^2 - \|u_{t+1} - w^*\|^2)$$

$$\underbrace{\le}_{\text{ט.ע. 2}} \frac{\eta}{2}\|g_{t+1}\|^2 + \frac{1}{2\eta}(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2)$$

כאשר המעבר *(**)* נובע מכך שלכל $a, b \in \mathbb{R}^d$ מתקיים

$$\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|b - a\|^2)$$

מ.ש.ל. (טענת עזר 3)

<u>הוכחת משפט 1:</u>

נזכור כי התשובה שאנחנו מחזירים בסיום הריצה היא

$$\widehat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

מכיוון ש- L היא קמורה, אנחנו יודעים שמתקיים

$$L(\widehat{w}) \le \frac{1}{T}\sum_{t=1}^{T} L(w_t)$$

(ניתן לראות את העובדה הזאת לגבי פונק' קמורות דרך אי-שוויון ינסן; ראו תרגול)

ולכן

$$\underbrace{L(\widehat{w}) - L(w^*)}_{excess\ risk} \le \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(L(w_t) - L(w^*)\right)}_{average\ excess\ risk}$$

$$\underbrace{\le}_{\text{ט.ע. 3}} \frac{\eta}{2}\max_t \|g_t\|^2 + \frac{1}{2\eta T}\sum_{t=1}^{T}\left(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2\right)$$

$$\underbrace{\le}_{\text{ט.ע. 1}} \frac{\eta}{2}G^2 + \frac{1}{2\eta T}\sum_{t=1}^{T}\left(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2\right)$$

$$\underbrace{\le}_{\substack{\text{טור}\\ \text{טלסקופי}}} \frac{\eta}{2}G^2 + \frac{1}{2\eta T}\left(\underbrace{\|w_1 - w^*\|^2}_{\le R} - \underbrace{\|w_{T+1} - w^*\|^2}_{\ge 0}\right)$$

$$\le \frac{\eta}{2}G^2 + \frac{R^2}{2\eta T}$$

$$= \frac{GR}{\sqrt{T}} \qquad \text{for } \eta = \frac{R}{G\sqrt{T}}$$

מ.ש.ל. (משפט 1)

מה קרה פה אינטואיטיבית? היו לנו בהוכחה 2 נקודות חשובות:

1. טענת עזר 3 אמרה לנו שלאורך הריצה, ה *excess risk* הוא בדרך כלל נמוך (כי הוא לא יכול להיות גדול יותר מדי פעמים כי אז המרחק שלנו לאופטימייזר הולך וקטן עד שמגיעים אליו והתהליך חייב לעצור)
2. מכיוון שבסיום הריצה של האלגוריתם אנחנו מחזירים את ממוצע הנקודות שביקרנו בהם בתור התשובה שלנו $\widehat{w}$ (ומכיוון שפונקציית ההפסד היא קמורה) אז אנחנו יודעים שה *excess risk* שלנו בסיום הריצה יהיה שווה לממוצע ה *excess risks* שהיו לנו לאורך הריצה, אבל הממוצע הזה חייב להיות נמוך לפי התובנה הקודמת.

סיימנו את ניתוח utility עבור הגרסה הלא פרטית של האלג' הזה (ללא הוספת הרעש לגרדיאנטים). עכשיו אנחנו רוצים לנתח את האלג' עם הרעשים.

<u>משפט 2:</u> נניח שפונקציית ההפסד $L: C \to \mathbb{R}$ היא קמורה וגם G-ליפשיץ. בנוסף נניח ש C היא קבוצה סגורה וקמורה עם קוטר R. נסמן $w^* = \underset{w \in C}{\operatorname{argmin}} L(w)$. אלג' Noisy-PGD מחזיר נקודה $\widehat{w}$ המקיימת

$$\mathbb{E}\left[L(\widehat{w}) - L(w^*) \right] \lesssim \eta\, G^2 \left(1 + \frac{T \cdot d^2 \cdot \log(1/\delta)}{n^2 \varepsilon^2} \right) + \frac{R^2}{\eta T}$$

כאשר התוחלת היא מעל הגרלת הרעשים בחישוב הגרדיאנטים.

בפרט, אם נבחר $T = \frac{\varepsilon^2 \cdot n^2}{d^2 \cdot \log\left(\frac{1}{\delta}\right)}$ ונבחר $\eta = \frac{R}{G\sqrt{T}}$ אז נקבל שתוחלת ה *excess risk* היא לכל היותר (תוך הזנחת קבועים)

$$\frac{GR}{\sqrt{T}}$$

כלומר קיבלנו את אותו ביטויי כמו שהיה לנו במשפט 1 (ללא הפרטיות)!
שאלה: אז איפה העלות שאנחנו משלמים עבור פרטיות?
תשובה: (1) חסמנו את ה *excess risk* רק בתוחלת, אבל זה כ"כ לא נורא...
(2) בשביל לקבל את החסם הזה דרשנו $T = \frac{\varepsilon^2 \cdot n^2}{d^2 \cdot \log\left(\frac{1}{\delta}\right)}$. כלומר בניגוד לגרסה הלא פרטית שבה החסם על ה *excess risk* היה נכון ללא קשר לגודל של הדטה, עכשיו כדי לקבל את אותו חסם אנחנו צריכים להניח שיש לנו מספיק דטה. ספציפית אנחנו צריכים $n \geq \frac{\sqrt{T \cdot \log\left(\frac{1}{\delta}\right)} \cdot d}{\varepsilon}$. בהמשך נתווכח קצת עם הדרישה הזאת...

ההוכחה של משפט 2 תהייה דומה להוכחה שראינו עבור *PGD* ללא הרעשים.

<u>טענת עזר 4 ("טענת התקדמות"):</u>
לכל $t = 0,1,\dots,T-1$ מתקיים:

$$\mathbb{E}\left[L(w_t) - L(w^*) \right] \leq \mathbb{E}\left[\frac{\eta}{2} \|\tilde{g}_{t+1}\|^2 + \frac{1}{2\eta}\left(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2 \right) \right]$$

<u>הוכחת טענת עזר 4:</u>
אנחנו מניחים שפונקציית ההפסד L היא קמורה. לפי ההגדרה השקולה לקמירות של פונקציה אנחנו מקבלים

$$L(w_t) - L(w^*) \leq \langle \nabla L(w_t)\, , w_t - w^* \rangle = \langle g_{t+1}\, , w_t - w^* \rangle$$

ניקח תוחלת משני האגפים ונקבל

$$\mathbb{E}\left[L(w_t) - L(w^*) \right] \leq \mathbb{E}\left[\langle g_{t+1}\, , w_t - w^* \rangle \right]$$

$$= \mathbb{E}\left[\langle \underbrace{g_{t+1} + h_{t+1}}_{=\tilde{g}_{t+1}}\, , w_t - w^* \rangle \right] - \underbrace{\mathbb{E}\left[\langle h_{t+1}\, , w_t - w^* \rangle \right]}_{=0}$$

כי h_{t+1} בלתי תלוי w_t
וגם $\mathbb{E}[h_{t+1}]=0$

$$= \mathbb{E}\left[\langle \tilde{g}_{t+1}, w_t - w^* \rangle \right] \underbrace{\leq}_{\text{כמו בטענת עזר 3}} \mathbb{E}\left[\frac{\eta}{2} \|\tilde{g}_{t+1}\|^2 + \frac{1}{2\eta}(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2) \right]$$

מ.ש.ל. (טענת עזר 4)

<u>הוכחת משפט 2:</u>

נזכור כי התשובה שאנחנו מחזירים בסיום הריצה היא

$$\hat{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$$

מכיוון ש- L היא קמורה, אנחנו יודעים שמתקיים

$$L(\hat{w}) \leq \frac{1}{T}\sum_{t=1}^{T} L(w_t)$$

ולכן

$$\mathbb{E}\left[L(\hat{w}) - L(w^*) \right] \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[L(w_t) - L(w^*) \right]$$

$$\leq \frac{\eta}{2T}\left(\sum_{t=1}^{T} \mathbb{E}\left[\|\tilde{g}_{t+1}\|^2 \right] \right) + \frac{1}{2T\eta}\sum_{t=1}^{T} \mathbb{E}\left[\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2 \right]$$

$$\leq \frac{\eta}{2}\max_t \mathbb{E}\left[\|\tilde{g}_{t+1}\|^2 \right] + \frac{1}{2T\eta}\sum_{t=1}^{T} \mathbb{E}\left[\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2 \right]$$

$$\underbrace{\leq}_{\text{טור טלסקופי}} \frac{\eta}{2}\max_t \mathbb{E}\left[\|\tilde{g}_{t+1}\|^2 \right] + \frac{1}{2\eta T}\mathbb{E}\left[\underbrace{\|w_1 - w^*\|^2}_{\leq R} - \underbrace{\|w_{T+1} - w^*\|^2}_{\geq 0} \right]$$

$$\leq \frac{\eta}{2}\max_t \mathbb{E}\left[\|\tilde{g}_{t+1}\|^2 \right] + \frac{R^2}{2\eta T} = ((1))$$

כמעט סיימנו להוכיח את משפט 2. הדבר האחרון שאנחנו צריכים להבין כאן זה איך נראה $\mathbb{E}\left[\|\tilde{g}_{t+1}\|^2 \right]$.

לצורך כך, נזכור שאם אנחנו מוסיפים רעש עם תוחלת אפס H לווקטור קבוע g אזי מתקיים

$$\mathbb{E}\left[\|g + H\|^2 \right] = \mathbb{E}\left[\|g\|^2 \right] + \mathbb{E}\left[\|H\|^2 \right] + 2\underbrace{\mathbb{E}\left[\langle g, H \rangle \right]}_{=0}$$

במקרה שלנו, אנחנו יודעים ש

$$\mathbb{E}\left[\ \|\tilde{g}_{t+1}\|^2\ \right] = \mathbb{E}\left[\ \|g_{t+1} + h_{t+1}\|^2\ \right]$$

$$= \mathbb{E}\left[\ \|g_{t+1}\|^2\ \right] + \mathbb{E}\left[\ \|h_{t+1}\|^2\ \right] \underbrace{\leq}_{\text{ליפשיץ}} G^2 + \mathbb{E}\left[\ \|h_{t+1}\|^2\ \right]$$

$$\underbrace{\lesssim}_{\substack{\text{תכונות הרעש}\\(**)}} G^2 + \frac{G^2 \cdot T \cdot d^2 \cdot \log(1/\delta)}{n^2\varepsilon^2}$$

(** לפי לנאריות התוחלת: התוחלת של כל קואו בריבוע זה השונות)

לכן

$$((1)) \lesssim \eta\, G^2 \left(1 + \frac{T \cdot d^2 \cdot \log(1/\delta)}{n^2\varepsilon^2}\right) + \frac{R^2}{\eta T}$$

מ.ש.ל. (משפט 2)

אז האלגוריתם הזה כמו שהצגנו אותו משמר פרטיות ומשיג תוצאות *utility* לא רעות. אבל הניתוח שלנו היה מאוד בזבזני ואפשר בקלות לקבל תוצאה הרבה יותר טובה מבחינת ה *excess risk* שאנחנו משיגים. הנקודה היא שכשניתחנו את הבטחת הפרטיות של האלגוריתם, אמרנו:

ישנן T שלבים בריצה ובכל שלב אנחנו מחשבים גראדירנט (שהוא ווקטור d מיימדי). תחת ההנחה שהפונקציה ℓ היא G-ליפשיץ, אז הראינו שהנורמה של הגרדיאנט הזה היא לכל היותר $\frac{2G}{n}$ וזה בפרט אמר שכל קואו' היא בערך מוחלט לכל היותר $\frac{2G}{n}$.

מכאן הסקנו שלאורך כל הריצה יש לנו Td חישובים של קואו' (T איטרציות ו- d קואו בכל איטרציה) שאנחנו מרעישים עם לפלאס ולכל חישוב קואו כזה יש רגישות $\frac{2G}{n}$. לכן לפי שיקולי קומפוזיציה, כדי שכל האלגוריתם ישמר פרטיות הספיק לנו להוסיף רעש מסדר גודל $\sqrt{Td} \cdot \frac{2G}{n} \approx$ לחישוב של כל קואו'.

זה נכון, אבל זה מאוד בזבזני. הסיבה היא שברגע שהראינו שבכל איטרציה הנורמה של הגרדיאנט היא לכל היותר $\frac{2G}{n}$, אז לא צריך (ולא כדאי) להתחשבן על כל קואו' בנפרד. זה נכון שכל קואו' תמיד תהיה חסומה על ידי $\frac{2G}{n}$, אבל לא יכולות להיות יותר מדי קואו גדולות כאלה.

<u>הוספת רעש לפי נורמת *L2*</u>

<u>נתון:</u> דטהבייס $S = (x_1, \ldots, x_n) \in \mathbb{R}^d$ ופרמטר Δ כך שכל x_i הוא ווקטור *d*-מיימדי כך ש- $\|x_i\| \leq \Delta$.
<u>יש לחשב:</u> קירוב לסכום של S.

מה אנחנו יודעים להגיד על הבעייה הזאת לפי מה שלמדנו בתחילת הקורס?

<u>אפשרות א:</u> מ.לפלאס + קומפוזיציה רגילה:

$$\left(\sum_{i\in[n]} x_i\right) + \left(\mathrm{Lap}\left(\frac{\Delta d}{\varepsilon}\right)\right)^d$$

<u>אפשרות ב:</u> מ.לפלאס + קומפוזיציה חזקה:

$$\left(\sum_{i\in[n]} x_i\right) + \left(\mathrm{Lap}\left(\frac{\Delta\sqrt{d\cdot\log\left(\frac{1}{\delta}\right)}}{\varepsilon}\right)\right)^d$$

<u>משפט 3:</u> עבור הבעייה הנ"ל, האלגוריתם הבא משמר (ε,δ)-פ"ד:

$$\left(\sum_{i\in[n]} x_i\right) + \left(\mathrm{Lap}\left(\frac{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}{\varepsilon}\right)\right)^d$$

לצורך ההוכחה, הזכרו במשפט הקומפוזיציה החזקה שלמדנו:

<u>משפט קומפוזיציה חזקה:</u>
יהיו $0 < \varepsilon,\delta \le 1$. הפעלה אדפטיבית של k מכניזמים המשמרים (ε,δ)-פ"ד כל אחד (ללא גישה נוספת לדטהבייס) משמרת

$$-\text{פ"ד}\left(\sqrt{2k\ln\left(\frac{1}{k\delta}\right)}\cdot\varepsilon + 2k\cdot\varepsilon^2 \;\;,\;\; 2k\delta\right)$$

<u>מסקנה:</u> כדי לקבל אלגוריתם המשמר $(\tilde{\varepsilon},\tilde{\delta})$-פ"ד, מספיק לדאוג שכ"א מ- k המכניזמים שנריץ ישמר

$$-\text{פ"ד}\left(O\left(\frac{\tilde{\varepsilon}}{\sqrt{k\ln\left(\frac{1}{\tilde{\delta}}\right)}}\right)\;,\;\frac{\tilde{\delta}}{2k}\right)$$

כשלמדנו את משפט הקומפוזיציה הזה, התרכזנו במקרה בו לכל אחד מ k המכניזמים שעשינו עליהם קומפוזיציה היה בדיוק את אותם פרמטרים (ε,δ). אבל מסתבר שאפשר להכליל את זה באופן הבא (לא נראה את ההוכחה של ההכללה):

<u>משפט קומפוזיציה חזקה עם פרמטרים שונים (לא פורמלי):</u>
יהי $\delta \in [0,1]$. הפעלה אדפטיבית של k מכניזמים פרטיים (ללא גישה נוספת לדטהבייס) עם פרמטרים $(\varepsilon_1,0),\ (\varepsilon_2,0),\ \ldots,\ (\varepsilon_k,0)$, בהתאמה, משמרת

$$\text{פ"ד}-\left(\sum_{i=1}^{k}2\varepsilon_i^2+\sqrt{\sum_{i=1}^{k}2\varepsilon_i^2\log\left(\frac{1}{\delta}\right)}\ \ ,\ \ \delta\right)$$

<u>סקיצת הוכחת משפט 3:</u>

נקבע שני דטהבייסים שכנים $S,S'=S\cup\{x\}$ כך שמתקיים $\|x\|\le\Delta$.

נסמן $x=(\alpha_1,\dots,\alpha_d)$ כך ש-

$$\sum_{i=1}^{d}\alpha_i^2\le\Delta^2$$

זכרו כי האלגוריתם שלנו מחשב את סכום הווקטורים ומוסיף לכל קואו' רעש מ $\text{Lap}\left(\frac{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}{\varepsilon}\right)$.

מכיוון שהדטהבייסים השכנים שלנו שונים רק בווקטור $x=(\alpha_1,\dots,\alpha_d)$, ההבדל בין סכומי הדטהבייסים בכ"א מהקואו' i הוא α_i ולזה אנחנו מוסיפים רעש מסדר גודל $\frac{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}{\varepsilon}$.

אנחנו יכולים לחשוב על כל קואו' כזאת כאילו היא מקבילה להפעלה בודדת של המ.לפלאס. "הרגישות" בקואו' הזאת היא α_i ולכן פרמטר הפרטיות "האפקטיבי" עבור הקואו' הזאת הוא $\frac{\alpha_i\cdot\varepsilon}{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}$.

לפי משפט הקומפוזיציה הנ"ל, זה אומר שהאלגוריתם כולו (על פני כל d הקואורדינטות) משמר פרטיות עם פרמטר (בערך)

$$\sum_{i=1}^{d}\left(\frac{\alpha_i\cdot\varepsilon}{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}\right)^2+\sqrt{\sum_{i=1}^{d}\left(\frac{\alpha_i\cdot\varepsilon}{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}\right)^2\log\left(\frac{1}{\delta}\right)}$$

$$=\sum_{i=1}^{d}\left(\frac{\alpha_i\cdot\varepsilon}{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}\right)^2+\sqrt{\sum_{i=1}^{d}\left(\frac{\alpha_i\cdot\varepsilon}{\Delta}\right)^2}$$

$$=\left(\frac{\varepsilon}{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}\right)^2\sum_{i=1}^{d}{\alpha_i}^2+\frac{\varepsilon}{\Delta}\cdot\sqrt{\sum_{i=1}^{d}{\alpha_i}^2}$$

$$\le\left(\frac{\varepsilon}{\Delta\sqrt{\log\left(\frac{1}{\delta}\right)}}\right)^2\Delta^2+\frac{\varepsilon}{\Delta}\cdot\Delta\lesssim\varepsilon$$

מ.ש.ל.