מבוא למערכות לומדות - הרצאה 5 - בעיות למידה קמורות

24 ביוני 2015

בשיעור היום נלמד על כלל למידה הנקרא Regularized Loss Minimization (RLM). אנו נראה שאלגוריתמים העוקבים אחרי הכלל הנ"ל מסוגלים ללמוד בעיות למידה שהן קמורות, ליפשיציות וחסומות (נגדיר בהמשך למה הכוונה). כמו כן, נראה שכל עוד ניתן לחשב ביעילות את $l_{(x,y)}(w)$ ואת $\nabla l_{(x,y)}(w)$, קיימים אלגוריתמים יעילים העוקבים אחרי הכלל הנ"ל. האלגוריתמים הללו יאפשרו לנו לפתור ביעילות שורה של בעיות למידה קמורות. בהמשך נראה כיצד ניתן להשתמש בטכניקות הללו על מנת לטפל בבעיות שאינן קמורות.

1 בעיות למידה קמורות ולמידה באמצעות כללים יציבים - תזכורת

תהא $(X, Y, \mathcal{H}, l)$ בעיית למידה. פעמים רבות ניתן לתאר כל היפותזה במחלקה ע"י מספר, נאמר n, של פרמטרים. כלומר, קיימת העתקה $w \mapsto h_w$ מתת קבוצה W של $\mathbb{R}^n$ **על** $\mathcal{H}$. לדוגמא, אם $X \subset \mathbb{R}^n$, $Y = \mathbb{R}^n$, ו־$\mathcal{H}$ היא מחלקה של פונקציות לינאריות מ־X ל־Y, אז כל $h \in \mathcal{H}$ ע"י מתוארת ע"י n מספרים, שכן, h היא מהצורה

$$h(x) = \sum_{i=1}^{n} w_i x_i$$

כאשר עובדים עם מחלקות כאלו, עושים שימוש נרחב בפרמטריזציה הנ"ל של ההיפותזות - כאשר מחפשים היפותזה טובה, בפועל עושים אופטימיזציה על הפרמטרים. גם לאחר שלב הלימוד, הדרך בה ההיפותזה שלמדנו נשמרת בזיכרון היא ע"י כך ששומרים את ערכי הפרמטרים המתארים אותה. לאור הנ"ל, יהיה נוח להשתמש במינוח הבא:

הגדרה 1.1 **פרמטריזציה קמורה** של בעיית למידה $(X, Y, \mathcal{H}, l)$ היא העתקה $w \mapsto h_w$ מקבוצה קמורה W **על** $\mathcal{H}$.

נעיר שהרבה פעמים הפרמטריזציה תהיה ברורה מן ההקשר ולא נציין אותה בפירוש.

סימונים. יהיה נוח לסמן ב־$L_{\mathcal{D}}(w)$ ו־$L_S(w)$ את הפונקציות $L_{\mathcal{D}}(w) = L_{\mathcal{D}}(h_w)$ ו־$L_S(w) = L_S(h_w)$. כמו, יהיה נוח להשתמש בסימון

$$l_{(x,y)}(h_w) := l_{(x,y)}(w) := l(h_w(x), y)$$

הגדרה 1.2 בעיית למידה $(X, Y, \mathcal{H}, l)$ עם פרמטריזציה קמורה $w \mapsto h_w$ תיקרא **קמורה** אם לכל $(x, y) \in X \times Y$ הפונקציה $l_{(x,y)}(w)$ הינה קמורה.

הגדרה 1.3 נאמר שהבעיה היא ρ**־ליפשיצית** אם לכל $(x, y) \in X \times Y$ $l_{(x,y)}(w)$ היא ρ־ליפשיצית[1]. נאמר שהבעיה R**־חסומה** אם W מוכלת בכדור ברדיוס R.

הערה: אם הבעיה היא ρ־ליפשיצית, אז מפני שממוצע של פונקציות ρ־ליפשיציות הוא ρ־ליפשיצי, מתקיים שלכל מדגם S הפונקציה $L_S(w)$ הינה ρ־ליפשיצית. לכן, עבור בעיה כנ"ל, בעיית מיזעור השגיאה האמפירית הינה בעיה של מיזעור פונקציה קמורה ו־ρ־ליפשיצית.

הגדרה 1.4 עבור פונקציה $\epsilon : \mathbb{N} \to \mathbb{R}_+$ אנו נאמר ש־$\mathcal{A}$ הוא $\epsilon(m)$**־יציב** אם לכל מדגם

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset X \times Y$$

ולכל $1 \leq i \leq m$ ו־$(x, y) \in X \times Y$ מתקיים

$$l_{(x_i,y_i)}(\mathcal{A}(S^i)) \leq l_{(x_i,y_i)}(\mathcal{A}(S)) + \epsilon(m)$$

כאשר S^i הוא המדגם המתקבל מ־S ע"י החלפת (x_i, y_i) ב־(x, y).

למה 1.5 (אלגוריתם יציבים לא עושים overfit) אם $\mathcal{A}$ הינו $\epsilon(m)$**־יציב** אז

$$E_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))\right] \leq \epsilon(m)$$

2 למידות של בעיות קמורות

כאמור, אנו נראה אלגוריתם, שהינו יעיל תחת דרישות לא מחמירות, המסוגל ללמוד בעיות למידה קמורות וליפשיציות. לפני שנציג וננתח אותו, נראה שלא ניתן ללמוד בעיות קמורות כלליות.

2.1 מדוע יש לדרוש חסימות וליפשיציות?

כאמור, אנו נראה שבעיות חסומות וליפשיציות הן למידות. לפני כן, נעיר שבעיות קמורות כלליות אינן למידות. נביט בבעיית הלמידה הקמורה הבאה:

$$X = [0, 1], \;\; Y = \mathbb{R}, \;\; l(\hat{y}, y) = |\hat{y} - y|, \;\; \mathcal{H} = \{h_w(x) = w \cdot x \mid w \in \mathbb{R}\}$$

אנו נראה שלכל אלגוריתם $\mathcal{A}$ מתקיים $m_{\mathcal{A}}\left(1, \frac{1}{10}\right) = \infty$. כלומר, נראה שלא קיים $m > 0$ כך שלכל התפלגות $\mathcal{D}$ על $X \times Y$ מתקיים

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\mathcal{A}(S)) \leq L_{\mathcal{D}}(\mathcal{H}) + 1\right) \geq \frac{9}{10}$$

[1] כזכור, עבור $C \subset \mathbb{R}^n$ פונקציה $f : C \to \mathbb{R}$ תיקרא ρ**־ליפשיצית** אם $|f(x) - f(y)| \leq \rho\|x - y\|$ לכל $x, y \in C$.

לשם פשטות, נראה זאת רק עבור אלגוריתמים דטרמיניסטיים. נניח שזה לא המצב. נסמן ב־$\bar{h}$ את ההיפותזה אותה האלגוריתם מחזיר על מדגם בן m איברים שכל הדוגמאות בו הן $(0,0) \in X \times Y$. נניח בה"כ ש־$\bar{h}(1) \leq 0$. נבחר μ מספיק קטן כך ש־$(1-\mu)^m \geq \frac{1}{2}$ (למה יש כזה?). נביט כעת בהתפלגות $\mathcal{D}$ המוגדרת באופן הבא:

$$\Pr_{(x,y)\sim\mathcal{D}}((x,y)=(0,0)) = 1-\mu,\ \Pr_{(x,y)\sim\mathcal{D}}\left((x,y)=\left(1,\frac{2}{\mu}\right)\right) = \mu$$

לא קשה להראות ש־$L_{\mathcal{D}}(\mathcal{H}) = 0$ (שכן להיפותזה $h(x) = \frac{2}{\mu}\cdot x$ יש שגיאה 0). כעת, כאשר $S \sim \mathcal{D}^m$, בהסתברות

$$(1-\mu)^m \geq \frac{1}{2}$$

כל הדוגמאות במדגם הן $(0,0)$, ולכן האלגוריתם יחזיר את $\bar{h}$. במקרה הזה יתקיים

$$\begin{aligned} L_{\mathcal{D}}(\mathcal{A}(S)) &= L_{\mathcal{D}}\left(\bar{h}\right) \\ &= (1-\mu)|\bar{h}(0)-0| + \mu\left|\bar{h}(1)-\frac{2}{\mu}\right| \\ &\geq \mu\left|\bar{h}(1)-\frac{2}{\mu}\right| \geq 2 > 1 + L_{\mathcal{D}}(\mathcal{H}) \end{aligned}$$

כלומר, קיבלנו שבהסתברות $\geq \frac{1}{2}$ מתקיים $L_{\mathcal{D}}(\mathcal{A}(S)) > L_{\mathcal{D}}(\mathcal{H}) + 1$. מצד שני, הנחנו שבהסתברות $\geq \frac{9}{10}$ מתקיים $L_{\mathcal{D}}(\mathcal{A}(S)) \leq L_{\mathcal{D}}(\mathcal{H}) + 1$. כלומר, הצבענו על שני מאורעות זרים ($L_{\mathcal{D}}(\mathcal{A}(S)) > L_{\mathcal{D}}(\mathcal{H})+1$ ו־$L_{\mathcal{D}}(\mathcal{A}(S)) \leq L_{\mathcal{D}}(\mathcal{H})+1$) שסכום ההסתברויות שלהם גדול מ־1. סתירה.

2.2 למידה באמצעות Regularized Loss Minimization

תהא $(X, Y, \mathcal{H}, l)$ בעיית למידה קמורה ביחס לפרמטריזציה $w \mapsto h_w$. נניח שמרחב הפרמטרים נתון ע"י $W \subset \mathbb{R}^n$. נניח, כמו כן, שהבעיה הינה ρ־ליפשיצית ו־R־חסומה. נאמר שאלגוריתם **מממש את כלל ה־RLM עם פרמטר רגולריזציה** $\lambda > 0$ אם הוא ממזער את הפונקציה

$$L_S^\lambda(w) := L_S(w) + \lambda\|w\|^2 = \frac{1}{m}\sum_{i=1}^m l_{(x_i,y_i)}(w) + \lambda\sum_{j=1}^n w_j^2$$

על פני $w \in W$.

האפקט של הוספת גורם הרגולריזציה. כלל ה־RLM דומה מאד לכלל ה־ERM. ההבדל היחיד הוא ההוספה של **גורם הרגולריזציה** $\lambda\|w\|^2$. לתוספת הנ"ל שני אפקטים:

- כפי שנראה, התוספת הנ"ל "תייצב" כלל ה־ERM, וככל ש־λ יהיה גדול יותר, האלגוריתם יהיה יציב יותר. קונקרטית נראה שכלל ה־RLM הינו $\frac{2\rho^2}{\lambda m}$־יציב.

- מצד שני, כאשר λ גדול, כלל ה־RLM יתרחק מכלל ה־ERM.

הנקודה הראשונה מעודדת אותנו לקבוע λ גדול, בעוד השנייה מעודדת לקבוע λ קטן. אנו נראה כיצד לקבוע את λ לפי m, כך שיתקבל אלגוריתם עם סיבוכיות מדגם קרובה לאופטימלית.

יעילות. כפי שתראו בתרגול, כאשר ניתן לחשב ביעילות את $l_{(x,y)}(w)$ ואת $\nabla l_{(x,y)}(w)$ בהינתן (x,y) ו־w, וכאשר W היא קבוצה "יפה" (למשל כאשר W הוא המרחב $\mathbb{R}^n$ כולו או כדור במרחב), קיימים אלגוריתמים יעילים המממשים את כלל ה־RLM.

ניגש כעת להוכיח שכלל ה־RLM מאפשר ללמוד בעיות קמורות, ליפשיציות וחסומות. הלמה הראשונה, והעיקרית, מראה שכלל ה־RLM הינו יציב.

למה 2.1 כלל ה־RLM הינו $\frac{2\rho^2}{\lambda m}$־יציב.

לפני שנוכיח את הלמה, נסיק ממנה מספר מסקנות

מסקנה 2.2 עבור כלל ה־RLM עם $\lambda=\sqrt{\frac{2\rho^2}{R^2 m}}$ מתקיים

$$E_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}(S))\right]\le L_{\mathcal{D}}\left(\mathcal{H}\right)+\rho R\sqrt{\frac{8}{m}}$$

נעיר כמה הערות

- **סיבוכיות המדגם.** עבור $\epsilon>0$, אם ניקח $m\ge\frac{32\rho^2R^2}{\epsilon^2}$ המסקנה תראה לנו שמתקיים

$$E_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}(S))\right]\le L_{\mathcal{D}}\left(\mathcal{H}\right)+\rho R\sqrt{\frac{8}{m}}\le L_{\mathcal{D}}\left(\mathcal{H}\right)+\frac{\epsilon}{2}$$

  מאי־שיוויון מרקוב עבור עבור המשתנה המקרי $L_{\mathcal{D}}(\mathcal{A}(S))-L_{\mathcal{D}}\left(\mathcal{H}\right)$, נקבע שבהסתברות לפחות $\frac{1}{2}$ על פני בחירת המדגם יתקיים

$$L_{\mathcal{D}}(\mathcal{A}(S))\le L_{\mathcal{D}}\left(\mathcal{H}\right)+\epsilon$$

  במילים אחרות, $m_{\mathcal{A}}\left(\epsilon,\frac{1}{2}\right)\le\frac{32\rho^2R^2}{\epsilon^2}$. נעיר ש־

  - החסם הנ"ל הדוק עד כדי קבוע במובן הבא: קיימות בעיות למידה קמורות שהן ρ־ליפשיציות ו־R־חסומות וקיים קבוע $c>0$ כך שלכל אלגוריתם $\mathcal{A}$ מתקיים $m_{\mathcal{A}}\left(\epsilon,\frac{1}{2}\right)>c\frac{\rho^2R^2}{\epsilon^2}$.
  - בהמשך, נראה כיצד ניתן לקבל מהחסם הנ"ל אלגוריתם (טיפה שונה) המקיים $m_{\mathcal{A}}\left(\epsilon,\delta\right)\le C\frac{\rho^2R^2\log\left(\frac{1}{\delta}\right)}{\epsilon^2}$ עבור קבוע אוניברסלי $C>0$.

- **נורמה מול מימד.** בד"כ ρ יהיה קבוע קטן. לכן, הגורם הדומיננטי בחסם על סיבוכיות המדגם שקיבלנו הינו $\left(\frac{R}{\epsilon}\right)^2$. זה בשונה ממה שהיה לנו עבור בעיות קלסיפיקציה, שם הגורם הדומיננטי היה $\frac{VC^2}{\epsilon}$.

- **המקרה** $W = \mathbb{R}^n$**.** הרבה פעמים, הבעיה תהיה ρ־ליפשיצית, אבל מרחב הפרמטרים הטבעי יהיה $\mathbb{R}^n$ שאיננו חסום. במקרה הזה, טיעון דומה לטיעון המוכיח את המסקנה יראה שאלגוריתם המממש את כלל ה־RLM עם פרמטר $\lambda = \sqrt{\frac{2\rho^2}{R^2 m}}$ (כלומר, ממזער את $L_S^\lambda(w)$ על פני $\lambda \in \mathbb{R}^n$) יקיים

$$E_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}(S))\right] \leq L_{\mathcal{D}}\left(\mathcal{H}_R\right) + \rho R\sqrt{\frac{8}{m}}$$

כאשר $\mathcal{H}_R = \{h_w : \|w\| \leq R\}$. נשים לב שככל שאנו מגדילים את R (או, באופן שקול, מקטינים את λ) הגורם $L_{\mathcal{D}}^{\text{hinge}}(\mathcal{H}_R)$ קטן, שכן אנו נשתמש במחלקה יותר גדולה. לעומת זאת, הגורם השני, $\rho R\sqrt{\frac{8}{m}}$, יגדל. נעיר שבפועל רצים על כמה ערכי λ ובוחרים את זה שהניב את היפותזה עם הביצועים הכי טובים. התהליך הנ"ל נקרא model-selection ונדבר עליו יותר בפירוט בשבוע הבא.

הוכחה: (של המסקנה) יהא $w^* \in W$ וקטור המקיים $L_{\mathcal{D}}(w^*) = L_{\mathcal{D}}(\mathcal{H})$ (נניח לשם פשטות שקיים כזה). מהלמה ומהמשפט שהוכחנו עבור אלגוריתמים יציבים נקבל שמתקיים

$$\begin{aligned}
E_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}(S))\right] &\leq E_{S\sim\mathcal{D}^m}\left[L_S(\mathcal{A}(S))\right] + \frac{2\rho^2}{\lambda m} \\
&\leq E_{S\sim\mathcal{D}^m}\left[L_S(\mathcal{A}(S)) + \lambda\|\mathcal{A}(S)\|^2\right] + \frac{2\rho^2}{\lambda m} \\
&\leq E_{S\sim\mathcal{D}^m}\left[L_S(w^*) + \lambda\|w^*\|^2\right] + \frac{2\rho^2}{\lambda m} \\
&= L_{\mathcal{D}}(w^*) + \lambda\|w^*\|^2 + \frac{2\rho^2}{\lambda m} \\
&\leq L_{\mathcal{D}}(w^*) + \lambda R^2 + \frac{2\rho^2}{\lambda m} \\
&= L_{\mathcal{D}}\left(\mathcal{H}\right) + \rho R\sqrt{\frac{8}{m}}
\end{aligned}$$

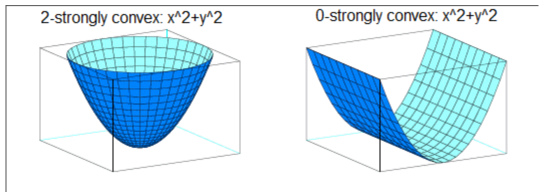
אי השיוויון השלישי נובע מכך ש־$\mathcal{A}$ מממש את כלל ה־RLM. אי השיוויון הרביעי נובע מכך שהבעיה הינה R־חסומה. ■

2.2.1 הוכחת למה 2.1

הלמה נכונה באופן כללי, אך לשם פשטות, נצמצם למקרה בו הפונקציות $w \mapsto l_{(x,y)}(w)$ גזירות פעמיים.

מה גורם לכלל ה־RLM להיות יציב? ובכן, לפונקציה מהצורה $g(w) = f(w) + \lambda\|w\|^2$ עבור $f : W \to \mathbb{R}$ קמורה, יש את התכונה הבאה: g היא 2λ**־קמורה חזק**: אחד הדרכים לאפיין פונקציה קמורה היא שלכל נקודה $w \in W$, אם מתרחקים מ־w באיזשהו כיוון $e \in \mathbb{R}^n$ ומביטים על ערך הפונקציה, אז השיפוע לא קטן. כלומר, הנגזרת השנייה של הפונקציה $\alpha(t) = f(w + te)$ הינה אי שלילית. ל־g יש את התכונה שלא רק שהשיפוע לא קטן, למעשה הוא עולה. קונקרטית, אם $\|e\| = 1$ אז $\forall t,\ \alpha''(t) \geq 2\lambda$. בציור הבא מופיעות

שתי פונקציות - השמאלית מבינהן 2־קמורה חזק, בעוד הימנית איננה קמורה חזק כלל (אם מתחילים מכל נקודה והולכים בכיוון ציר ה־y, השיפוע לא גדל).



מה לקמירות חזקה וליציבות? ובכן, אפשר לקבל איזשהי אינטואיציה מהתמונה השמאלית - אנו רואים שהנקודה הממזערת את הפונקציה (כלומר $(0,0)$) היא יציבה במובן הבא: ככל שמתרחקים ממנה ערך הפונקציה גדל. זה לא נכון בתמונה הימנית! שם, כאשר מתרחקים מהממזער (שוב - $(0,0)$) בכיוון ציר ה־y ערך הפונקציה לא גדל כלל! הטענה הבאה מראה שהתכונה הזו נכונה לכל g מהצורה הנ"ל.

טענה 2.3 תהא $f : W \to \mathbb{R}$ פונקציה קמורה. נניח ש־$w^* \in W$ ממזער את הפונקציה $g(w) = f(w) + \lambda\|w\|^2$ אזי, לכל $w \in W$,

$$g(w) \geq g(w^*) + \lambda\|w - w^*\|^2$$

הוכחה: נביט על הפונקציה $\alpha(t) = g(w^* + te)$ כאשר $e = \frac{w-w^*}{\|w-w^*\|}$ ו־$0 \leq t \leq \|w - w^*\|$. מכיוון ש־$w^*$ ממזער את g, $t = 0$ ממזער את α ולכן יתקיים

$$\alpha'(0) \geq 0 \tag{1}$$

כמו כן, אם נסמן $\beta(t) = f(w^* + te)$ נקבל שמתקיים

$$\begin{aligned}
\alpha(t) &= \beta(t) + \lambda\|w^* + te\|^2 \\
&= \beta(t) + \lambda\|w^*\|^2 + 2\lambda t\langle w^*, e\rangle + \lambda t^2\|e\|^2 \\
&= \beta(t) + \lambda\|w^*\|^2 + 2\lambda t\langle w^*, e\rangle + \lambda t^2
\end{aligned}$$

מכיוון ש־β קמורה (בדקו!), $\beta''(t) \geq 0$ ולכן

$$\alpha''(t) = \beta''(t) + 2\lambda \geq 2\lambda$$

כעת, ממשפט טיילור[2] קיים $t \in [0, \|w - w^*\|]$ עבורו מתקיים

$$\begin{aligned}
g(w) = \alpha(\|w - w^*\|) &= \alpha(0) + \alpha'(0) \cdot \|w - w^*\| + \frac{\alpha''(\xi)}{2}\|w - w^*\|^2 \\
&\geq g(w^*) + \lambda\|w - w^*\|^2
\end{aligned}$$

[2]נזכיר שמשפט טיילור אומר שאם $f : [0, a] \to \mathbb{R}$ גזירה פעמיים ברציפות אז יש $\xi \in (0, a)$ עבורו

$$f(a) = f(0) + f'(0)a + \frac{f''(\xi)}{2}a^2$$

■

כעת אנו מוכנים להוכיח את למה 2.1.

הוכחה: (של למה 2.1) נקבע מדגם $S \in (X \times Y)^m$, $(x, y) \in X \times Y$ ו־$1 \leq i \leq m$. יהא w_S הממזער של $L_S^\lambda(w)$ ויהא w_{S^i} הממזער של $L_{S^i}^\lambda(w)$. צריך להראות ש־

$$l_{(x_i,y_i)}(w_{S^i}) \leq l_{(x_i,y_i)}(w_S) + \frac{2\rho^2}{\lambda m}$$

מכיוון ש־$l_{(x_i,y_i)}$ היא ρ־ליפשיצית מתקיים

$$l_{(x_i,y_i)}(w_{S^i}) \leq l_{(x_i,y_i)}(w_S) + \rho\|w_{S^i} - w_S\| \tag{2}$$

ולכן די להראות ש־$\|w_{S^i} - w_S\| \leq \frac{2\rho}{\lambda m}$. אכן, מתקיים

$$L_{S^i}^\lambda(w) = L_S^\lambda(w) + \frac{l_{(x,y)}(w) - l_{(x_i,y_i)}(w)}{m} \tag{3}$$

עכשיו, מטענה 2.3,

$$L_S^\lambda(w_{S^i}) \geq L_S^\lambda(w_S) + \lambda\|w_{S^i} - w_S\|^2 \tag{4}$$

מכיוון ש־$l_{(x_i,y_i)}, l_{(x,y)}$ הינן ρ־ליפשיציות מתקיים

$$\begin{aligned}
\frac{l_{(x,y)}(w_{S^i}) - l_{(x_i,y_i)}(w_{S^i})}{m} &\geq \frac{l_{(x,y)}(w_S) - \rho\|w_{S^i} - w_S\| - l_{(x_i,y_i)}(w_S) - \rho\|w_{S^i} - w_S\|}{m} \\
&= \frac{l_{(x,y)}(w_S) - l_{(x_i,y_i)}(w_S)}{m} - \frac{2\rho\|w_{S^i} - w_S\|}{m} \qquad (5)
\end{aligned}$$

לסיום, מכיוון ש־w_{S^i} ממזער את $L_{S^i}^\lambda$, חייב להתקיים

$$\begin{aligned}
0 &\geq L_{S^i}^\lambda(w_{S^i}) - L_{S^i}^\lambda(w_S) \\
&\overset{(3)}{=} L_S^\lambda(w_{S^i}) - L_S^\lambda(w_S) \\
&\quad + \frac{l_{(x,y)}(w_{S^i}) - l_{(x_i,y_i)}(w_{S^i})}{m} - \frac{l_{(x,y)}(w_S) - l_{(x_i,y_i)}(w_S)}{m} \\
&\overset{(4),(5)}{\geq} \lambda\|w_{S^i} - w_S\|^2 - \frac{2\rho\|w_{S^i} - w_S\|}{m}
\end{aligned}$$

מכאן

$$\lambda\|w_{S^i} - w_S\|^2 \leq \frac{2\rho\|w_{S^i} - w_S\|}{m} \Rightarrow \|w_{S^i} - w_S\| \leq \frac{2\rho}{\lambda m}$$

■

3 מעבר לבעיות קמורות - ההופעה של הקושי החישובי

הדוגמאות העיקריות לבעיות למידה "טבעיות" שהן קמורות הן בעיות רגרסייה למינהן. כלומר, בעיות בהן $Y = \mathbb{R}$ והמרחק בין איברים ב־Y נמדד ע"י מדד מרחק (למשל $l(\hat{y}, y) =$ $(\hat{y} - y)^2$ או $l(\hat{y}, y) = |\hat{y} - y|$). כלל ה־RLM נותן לנו אלגוריתם יעיל ואפקטיבי לבעיות כאלו. עם זאת, אוסף הבעיות שנרצה לפתור מכיל הרבה מאד בעיות שאינן בעיות רגרסייה. למשל, בעיות קלסיפיקציה אינן קמורות (למשל, בגלל שבבעיות קלסיפיקציה $l_{(x,y)}$ מקבל רק את הערכים 0 ו־1, ואין פונקציה קמורה עם התכונה הנ"ל). באופן כללי יותר, בעיות בהן הפלט הוא דיסקרטי אינן קמורות.

בהמשך השיעור ובשלושת השיעורים הבאים, נתרכז בשיטות לתקוף בעיות קלסיפיקציה (ונדבר קצת גם על בעיות נוספות). נזכיר שבעיית למידה $(X, Y, \mathcal{H}, l)$ נקראית **בעיית קלסיפיקציה** אם Y היא קבוצה סופית ו־$l(\hat{y}, y) = l_{0-1}(\hat{y}, y) = \begin{cases} 0 & \hat{y} = y \\ 1 & \hat{y} \neq y \end{cases}$. למרבה הצער, רוב רובן של הבעיות הללו הן בעיות קשות חישובית. כלומר, ככל הנראה, לא קיים אלגוריתם יעיל המסוגל לפתור אותן.

נעיר שהקושי הוא חריף מאד: אפילו אם מובטח לנו ש־$L_{\mathcal{D}}(\mathcal{H}) = 0$ (כלומר, אנו במקרה הפריד), לא קיים אלגוריתם המסוגל להחזיר היפותזה עם שגיאה הקטנה מ־0.49999 (שגיאה של 0.5 ניתן להשיג באופן טריוויאלי, למשל, ע"י הטלת מטבע)! מקרה אחד יוצא דופן הוא לימוד חצאי מרחבים, שם קיים אלגוריתם יעיל עבור המקרה הפריד (כפי שראיתם בתרגול). אבל אפילו עבור חצאי מרחבים, ללא הנחת הפרידות, הבעיה הופכת למאד קשה: ככל הנראה, לא קיים אלגוריתם יעיל המסוגל להחזיר היפותזה עם שגיאה הקטנה מ־0.49999 אפילו אם מובטח לנו ש־$L_{\mathcal{D}}(\mathcal{H}) < 0.00001$. כלומר, אפילו אם קיים חצי מרחב עם שגיאה כמעט מושלמת, עדיין לא ניתן להחזיר היפותזה עם שגיאה לא טריוויאלית.

לאור הנ"ל, אנו לא נפתח אלגוריתמים שפותרים את הבעיה, או אפילו משיגים פתרון מקורב, כי פשוט אין כאלו (ככל הנראה). ניתן לחלק את האלגוריתמים שבכל זאת נראה לשני סוגים:

- **תחליפים קמורים Convex Surrogates.** שיטות בהן אנו "מחליפים" את את הבעיה בבעיה קמורה, כך שלבעיה החליפית יש את התכונות הבאות:
  - (יעילות) התחליף ניתן לפתרון ביעילות
  - (קשר למקור) אם משיגים שגיאה טובה בתחליף, ניתן לקבל שגיאה טובה גם במקור
- **יוריסטיקות.** המשפחה השנייה מכילה אלגוריתמים הפועלים ע"פ כלל טבעי ואינטואיטיבי, אך אין להם, לפחות כיום, בסיס תיאורטי מוצק. למרות החוסר בבסיס תיאורטי, הרבה פעמים יוריסטיקות עובדות בצורה טובה בפועל.

היום ושבוע הבא נתרכז בתחליפים קמורים. בשבועיים שאח"כ, נדבר על יוריסטיקות.