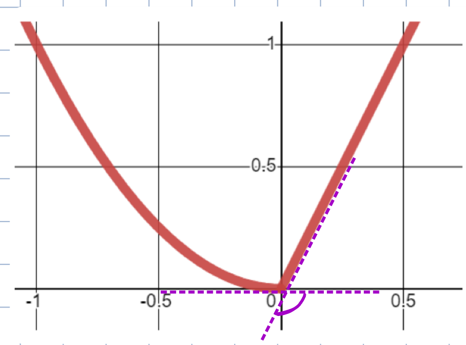
Part A – Optimization

Definition: The set of **Subgradients** of $f: V \to \mathbb{R}$ at point $u \in V$ is:

$$\partial f(u) \triangleq \{ g \in V \mid \forall v \in V: f(v) \geq f(u) + g^T(v-u) \}$$



1) Let $f(x) = \begin{cases} x^2 & x<0 \\ 2x & x \geq 0 \end{cases}$

1.1) Yes, $f(x)$ is convex

1.2) a sub derivative function g for f

- $g(x) = \begin{cases} 2x & x<0 \\ 2 & x>0 \\ 0 & x=0 \end{cases}$

- $\partial f(x) = \{ g \in V \mid \forall v \in V: f(v) \geq f(x) + g^T(v-x) \}$

$g(x) \in \partial f(x) \iff \forall v \in V: \; f(v) \geq f(x) + g_{(x)}^T(v-x)$

- **for $x<0$:** $\quad g(x) = 2x, \; f(x) = x^2$
  - **for $v<0$:** $\quad f(v) = v^2$

    $g(x) \in \partial f(x) \iff \forall v \in V: \; v^2 \geq x^2 + 2x(v-x)$

    $\iff \forall v \in V: \; v^2 \geq x^2 + 2xv - 2x^2$

    $\iff \forall v \in V: \; v^2 - 2xv + 2x^2 \geq 0$

    $\iff \forall v \in V: \; (v-x)^2 \geq 0$ ✓

    תמיד מתקיים

  - **for $v \geq 0$:** $\quad f(v) = 2v$

    $g(x) \in \partial f(x) \iff \forall v \geq 0: \; 2v \geq x^2 + 2x(v-x)$

    $\iff \forall v \geq 0: \; 2v \geq x^2 + 2xv - 2x^2$

    $\iff \forall v \geq 0: \; x^2 - 2xv + 2v \geq 0$ ✓

    חיובי → x^2, חיובי → $-2xv$, חיובי → $2v$

- for $x > 0$: $g(x) = 2$, $f(x) = 2x$
  - for $v \geq 0$: $f(v) = 2v$

$$g(x) \in \partial f(x) \iff \forall v \geq 0 : 2v \geq 2x + 2(v-x)$$

$$\iff \forall v \geq 0 : 2v \geq 2x + 2v - 2x$$

$$\iff \forall v \geq 0 : 0 \geq 0$$

תמיד מתקיים

  - for $v < 0$: $f(v) = v^2$

$$g(x) \in \partial f(x) \iff \forall v < 0 : v^2 \geq 2x + 2(v-x)$$

$$\iff \forall v < 0 : v^2 \geq 2x + 2v - 2x$$

$$\iff \forall v < 0 : v^2 - 2v \geq 0 \quad \checkmark$$

חיובי   חיובי

- for $x = 0$: $g(0) = 0$, $f(0) = 0$
  - for $v \geq 0$: $f(v) = 2v$

$$g(x) \in \partial f(x) \iff \forall v \geq 0 : 2v \geq 0 + 0(v-x)$$

$$\iff \forall v \geq 0 : 2v \geq 0 \quad \checkmark$$

  - for $v < 0$: $f(v) = v^2$

$$g(x) \in \partial f(x) \iff \forall v < 0 : v^2 \geq 0 + 0(v-x)$$

$$\forall v < 0 : v^2 \geq 0 \quad \checkmark$$

לסיכום: קיבלנו כ $g(x) \in \{g \in V \mid \forall v \in V : f(v) \geq f(x) + g^T(v-x)\}$

1.3) $\eta = 0.25$, $x_0 = -1.5$, $f(x) = \begin{cases} x^2 & x<0 \\ 2x & x\geq 0 \end{cases}$, $g(x) = \begin{cases} 2x & x<0 \\ 2 & x>0 \\ 0 & x=0 \end{cases}$

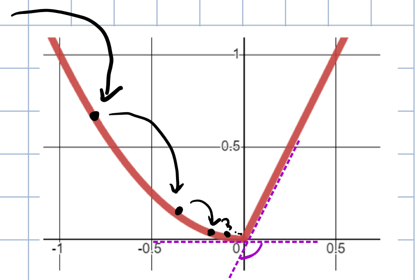
i	x_i	$f(x_i)$	$\frac{\partial f(x_i)}{\partial x} = g(x_i)$
0	-1.5	9/4	-3
1	-0.75	9/16	-1.5
2	-0.375	9/64	-0.75
3	-0.1875	9/256	-0.375

$x_i = x_{i-1} - \eta g(x_{i-1})$

$x_1 = -1.5 - 0.25\cdot(-3)$

$x_2 = -0.75 - 0.25(-1.5)$

$x_3 = 0.375 - 0.25(-0.75)$



Yes, the algorithm will converge to minimum.

($x_i = x_{i-1} - \frac{1}{4} 2x_{i-1} \rightarrow x_i = \frac{1}{2} x_{i-1}$ בכל איטרציה שווה חצי מהקודמת
וכן $x \to 0^-$ לכן x מתכנס ל-0 וכן $f(x) \to 0$.)

1.4) $\eta = 1$, $x_0 = -1.5$, $f(x) = \begin{cases} x^2 & x<0 \\ 2x & x\geq 0 \end{cases}$, $g(x) = \begin{cases} 2x & x<0 \\ 2 & x>0 \\ 0 & x=0 \end{cases}$

i	x_i	$f(x_i)$	$\frac{\partial f(x_i)}{\partial x} = g(x_i)$
0	-1.5	9/4	-3
1	1.5	3	2
2	-0.5	1/4	-1
3	0.5	1	2
4	-1.5	9/4	-3

loop

$x_i = x_{i-1} - \eta g(x_{i-1})$

$x_1 = -1.5 - 1\cdot(-3)$

$x_2 = 1.5 - 1(2)$

$x_3 = -0.5 - 1(-1)$

$x_4 = 0.5 - 1(2)$

No, the algorithm will not converge to minimum.

Part B- Regression

2.

X	Y
0	0
1	3
2	12

$, \mathcal{H} = \{ w_0 + w_1 x + w_2 x^2 + w_3 x^3 \mid (w_0, w_1, w_2, w_3) \in \mathbb{R}^4 \}$

2.1) $w_0 = 0,\ w_1 = 2,\ w_2 = 0,\ w_3 = 1 \Rightarrow h(x) = 2x + x^3$

$h(0) = 0 + 0 = 0$ ✓

$h(1) = 2 + 1 = 3$ ✓

$h(2) = 4 + 8 = 12$ ✓

2.2) A simple quadratic polynomial that fits the data perfectly:

$h(x) = 3x^2$

$h(0) = 0$ ✓

$h(1) = 3$ ✓

$h(2) = 12$ ✓

2.3) MSE: $L(w) = \frac{1}{m} \|Xw - y\|_2^2$

בסעיפים הקודמים הראנו ש- $Xw = y$ ולכן $\|Xw - y\| = \vec{0}$ כך

ש- $L(w) = 0$ עבור שני המקרים.

2.4) Calculate $L(w)$ for $y = 6x - 1$

$w = \begin{pmatrix} -1 \\ 6 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad y = \begin{pmatrix} 0 \\ 3 \\ 12 \end{pmatrix}$

$$L(w) = \frac{1}{3} \left\| \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 6 \end{pmatrix} - \begin{pmatrix} 0 \\ 3 \\ 12 \end{pmatrix} \right\|_2^2 = \frac{1}{3} \left\| \begin{pmatrix} -1 \\ 5 \\ 11 \end{pmatrix} - \begin{pmatrix} 0 \\ 3 \\ 12 \end{pmatrix} \right\|_2^2$$

$$= \frac{1}{3} \left\| \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix} \right\|_2^2 = \frac{1}{3} \cdot \left((-1)^2 + (2)^2 + (-1)^2 \right) = \frac{1}{3}(1 + 4 + 1) = 2$$

2.5) $L_\lambda(w)=\frac{1}{m}\|Xw-y\|_2^2+\lambda\|w\|_2^2$, כאשר

2.1) $\|Xw-y\|_2^2=0$, $\|w\|_2^2=\left\|\begin{pmatrix}0\\2\\0\\1\end{pmatrix}\right\|_2^2=1^2+2^2=5 \Rightarrow L_\lambda(w)=5\lambda$

2.2) $\|Xw-y\|_2^2=0$, $\|w\|_2^2=\left\|\begin{pmatrix}0\\0\\3\end{pmatrix}\right\|_2^2=3^2=9 \Rightarrow L_\lambda(w)=9\lambda$

2.3) $\frac{1}{m}\|Xw-y\|_2^2=2$, $\|w\|_2^2=\left\|\begin{pmatrix}-1\\6\end{pmatrix}\right\|_2^2=1^2+6^2=37$

$\Rightarrow L_\lambda(w)=2+37\lambda$

2.6) בדוגמה הנתונה, y גדל משמעותית כתלות ב- X.
רגרסור ליניארי צריך להתאים את עצמו למגמה זו. יש שני פתרונות
שיכולים לגשר את הגדילה הזו - הפיצ'רים המשמשים (x^i) או המקדמים שלהם:
עבור מונום עם i קטן יותר (quadratic בהשוואה ל- cubic), המעריך של x^i
קטן יותר ולכן המקדמים צריכים להיות גדולים יותר.
מקדמים גדולים יותר $\Leftarrow$ $L_\lambda(w)$ גדול יותר שכן הביטוי הרגולרי $\lambda\|w\|_2^2$
תלוי ישירות בגודל המקדמים.

- ריבוי המקדמים משפיע על התוצאה __פחות__ ממקדמים גדולים אך בעצם יותר.

2.7) בהינתן שאנחנו משתמשים מודל פשוט יותר כמודל עם פחות פיצ'רים/
במקרה הזה - החזקה יותר קטנה במימדי של X, ניתן להשתמש בתיקון הבא:
במקום רגולריזציה שמסתכלת על כל המקדמים כשקולים, רגולריזציה
שתקנוס יותר את פיצ'רים מדרגה גבוהה יותר.

- ניתן לבצע זאת למשל ע"י החלפת מקדם הרגולריזציה, מסקלר לוקטור

$\lambda=\begin{pmatrix}1\\2\\3\\\vdots\\i\end{pmatrix}$ כך של $L_\lambda(w)=\|Xw-y\|_2^2+\|\vec{\lambda}w\|_2^2$

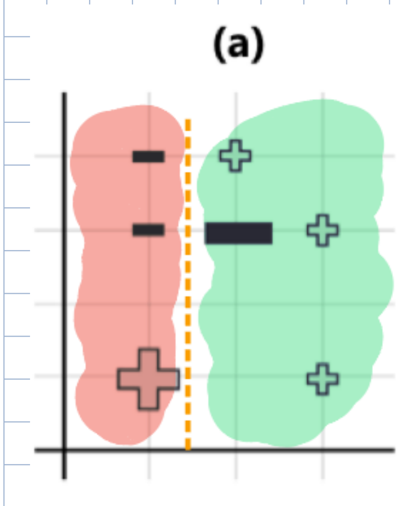
- דרך נוספת יכולה להיות החלפת הנורמה עצמה.

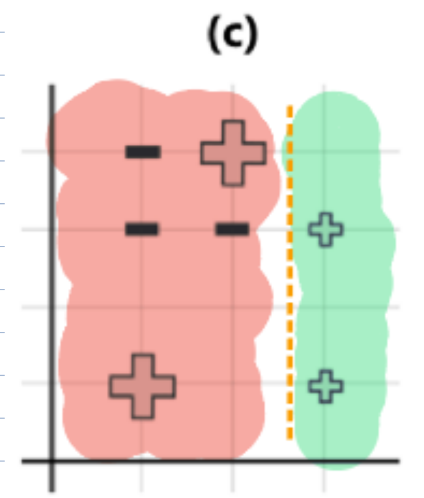
Part C – Boosting

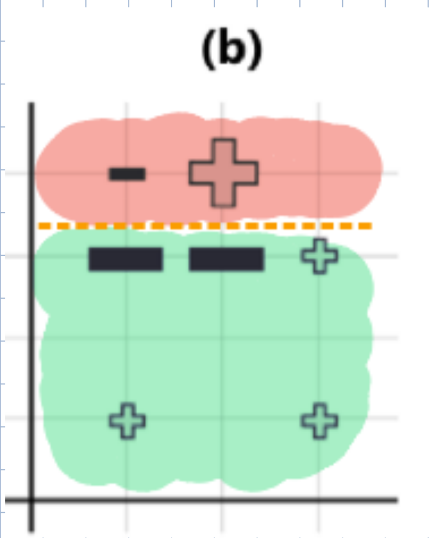
③ אחרי איטרציה יחידה מסווג ע"ב Decision stumps יהיה קו אופקי/אנכי באיזור שמפריד בין סוגי ה-⊕ לסוגי ה-⊖. מתכונות האלג', עבור דגימות שטעה בסיווג שלהן, נגדיל את ההסתברות שלהן לאיטרציה הבאה.
נקבע אם גרף הוא אפשרי או לא לפי תכונות אלו.

• המסווגים האפשריים בעלי האנטרופיה האופטימלית

(בכתום – הקו המפריד
באדום – סיווג כשלילי
בירוק – סיווג כחיובי)







הערה: נבחין כי עבור גרף b קיים מסווג חלש שמחזיר אותו. עם זאת, בהנחת אלג' שמחזיר שגיאה נמוכה, b לא "מבחר" כי יש בו שלושה סיווגים שגויים, בהשוואה לשניים שב-a וב-c.

4) Show: $\forall t \in [T]$ & $\sum_{i=1}^{m} D_i^{t+1} \mathbb{I}[y_i \neq h_t(x_i)] = \frac{1}{2}$

$\sum_{i=1}^{m} D_i^{t+1} \mathbb{I}[y_i \neq h_t(x_i)]$ / def of indicators

$= \sum_{i: y_i \neq h_t(x_i)} D_i^{t+1}$ / def of D_i^{t+1}

$= \sum_{i: y_i \neq h_t(x_i)} \frac{D_i^t \cdot \exp(-\alpha_t y_i h_t(x_i))}{\sum_j D_j^t \cdot \exp(-\alpha_t y_j h_t(x_j))}$ / for $y_i \neq h_t(x_i)$: $y_i \cdot h_t(x_i) = -1$

$= \sum_{i: y_i \neq h_t(x_i)} \frac{D_i^t \cdot \exp(\alpha_t)}{\sum_j D_j^t \cdot \exp(-\alpha_t y_j h_t(x_j))}$ / independent of i

$= \frac{\exp(\alpha_t)}{\sum_j D_j^t \cdot \exp(-\alpha_t y_j h_t(x_j))} \cdot \left(\sum_{i: y_i \neq h_t(x_i)} D_i^t \right)$ / def of ϵ_t

$= \frac{\exp(\alpha_t) \cdot \epsilon_t}{\sum_j D_j^t \cdot \exp(-\alpha_t y_j h_t(x_j))}$ /

$\sum_j D_j^t \cdot \exp(-\alpha_t y_j h_t(x_j)) = \sum_{y_j \neq h_t(x_j)} D_j^t \cdot \exp(-\alpha_t \underbrace{y_j h_t(x_j)}_{=-1}) + \sum_{y_j = h_t(x_j)} D_j^t \cdot \exp(-\alpha_t \underbrace{y_j h_t(x_j)}_{=1})$

$= \sum_{y_j \neq h_t(x_j)} D_j^t \cdot \exp(\alpha_t) + \sum_{y_j = h_t(x_j)} D_j^t \cdot \exp(-\alpha_t)$

$= \exp(\alpha_t) \cdot \epsilon_t + \exp(-\alpha_t)(1 - \epsilon_t)$ / $\alpha_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$

$= \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \cdot \epsilon_t + \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \cdot (1 - \epsilon_t)$

Back to our Expression:

$= \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \cdot \epsilon_t$

$\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \cdot \epsilon_t + \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \cdot (1 - \epsilon_t) = \frac{\sqrt{\epsilon_t(1-\epsilon_t)}}{\sqrt{\epsilon_t(1-\epsilon_t)} + \sqrt{\epsilon_t(1-\epsilon_t)}} = \frac{1}{2}$

Part 0 - Perceptron

5) If the data is separable, by theorem from class it will converge.

Let there be $\eta_1, \eta_2 > 0$.

We'll denote the w^t at the t update as

$$w^1_t = \begin{cases} w_{t-1} + \eta_1 y_t x_t & t > 0 \\ 0 & t = 0 \end{cases}$$

* $(x,y)_t$ is the updated t sample! t is not(!) an iterator over the trainset $\in [1,m]$

$$\Rightarrow w^1_t = \sum_{n=1}^{t} \eta_1 y_n x_n = \eta_1 \sum_{n=1}^{t} y_n x_n$$

likewise, for $w^2_t = \eta_2 \sum_{n=1}^{t} y_n x_n$

טענת עזר

נוכיח באינדוקציה בכל נקודה t, המסווגים המתקבלים מ-w_1, w_2 מצביעים לאותו הכיוון ומסכימים זה עם זה.

בסיס האינדוקציה: עבור $t=0$ מתקיים $w^1_0 = w^2_0 = 0$ ולכן המסווגים המתקבלים מהם מצביעים לאותו הכיוון ומסכימים זה עם זה.

הנחת האינדוקציה: בצעד ה-$t-1$ המסווגים המתקבלים מ-w_1, w_2 מצביעים לאותו הכיוון ומסכימים זה עם זה. מטענת העזר מתקיים

$$w^1_{t-1} = \eta_1 \sum_{n=1}^{t-1} y_n x_n \quad , \quad w^2_{t-1} = \eta_2 \sum_{n=1}^{t-1} y_n x_n$$

צעד האינדוקציה: בעדכון ה-t:

מאחר ושני המסווגים עד כה, הסכימו זה עם זה, (והסדר של מעבר ה-samples זהה), העדכון יתבצע עבור אותה דוגמה (x_t, y_t).

ב) מהגדרת העדכון באלגו':

$$w^1_t = w^1_{t-1} + \eta_1 x_t y_t = \eta_1 \sum_{n=1}^{t} y_n x_n$$

$$w^2_t = w^2_{t-1} + \eta_2 x_t y_t = \eta_2 \sum_{n=1}^{t} y_n x_n$$

$$\Rightarrow \operatorname{sign}(w^{1T}_t x_i) = \operatorname{sign}\left(\left(\eta_1 \sum_{n=1}^{t} y_n x_n\right) x_i\right) = \operatorname{sign}\left(\left(\sum_{n=1}^{t} y_n x_n\right) x_i\right) =$$

$$sign\left(\left(\eta_2 \sum_{n=1} y_n x_n\right) x_i\right)$$

שכן כפל בסקלר לא משפיע על כיוון הוקטור.

⟸ מתקבל שבעבור ה-z המסווגים מצביעים לאותו הכיוון ומסכימים זה עם זה.

• מאחר ולכל z התקבל שהמסווגים מסכימים זה עם זה, מס' האיטרציות הינם להתכנסות זהה.