

# מבוא למערכות לומדות

סיכום הרצאות - סמסטר ב' 2020 - 2021

סוכם ע"י טל ברט

لتיקונים והערות - [tal.beradt@mail.huji.ac.il](mailto:tal.beradt@mail.huji.ac.il)

## תוכן העניינים

5	<b>1 הרצתה 1 - מבוא</b>	1.1 אומדים . . . . .
5	התפלגיות במימד גובה . . . . .	1.2
5	טרנספורמציה ליניארית על דאטא . . . . .	1.3
8	אי-שוויונות וחסמים . . . . .	1.4
10	חיזוי מטבע . . . . .	1.5
10		
13	<b>2 הרצתה 2 - רגרסיה ליניארית</b>	2.1 בעיה . . . . .
13	פונקציית Loss . . . . .	2.2
14	מיעור הסיכון האמפירי (ERM) . . . . .	2.3
14	איך נלמד? . . . . .	2.4
15	מציאת המינימום . . . . .	2.5
15	המשוואות הנורמליות . . . . .	2.6
16	פתרת המשוואות הנורמליות . . . . .	2.7
16	פירוק SVD . . . . .	2.8
18	סיכום מקרה ללא רוש . . . . .	2.9
19	המקרה הרועש . . . . .	2.10
19	עקרון הנראות המירבית . . . . .	2.11
20	התאמת פולינומית Polynomial Fitting . . . . .	2.12
22	<b>3 הרצתה 3 - Clasification</b>	3.1 דיווק . . . . .
22	ניתוח מסובג חדש . . . . .	3.2
23	מסובג Half-Space . . . . .	3.3
23	מסובג SVM - Support Vector Machines . . . . .	3.4
25	מסובג רוגסיה לוגיסטי . . . . .	3.5
26	מסובג k - שכנים קרובים . . . . .	3.6
29		
30	<b>4 הרצתה 4 - מודל PAC</b>	4.1 חזרה - בעיית סיווג . . . . .
30	מסגרת תיאורטית עבור למידה . . . . .	4.2
30	מושגים והגדרות מרכזיים . . . . .	4.3
31	המשפט היסודי של הלמידה הסטטיסטי . . . . .	4.4
31	נתחיל לשחק . . . . .	4.5
32		

32	גרסה ראשונה   Learning Game 1.0	4.6
34	אלגוריתם למידה PAC	4.7
34	גרסה שנייה   Learning Game 2.0	4.8
35	אין ארכוות חינוך	4.9
36	מחלקות היפותזות	4.10
36	גרסה שלישית   Learning Game 3.0	4.11
37	פונקציות סף	4.12
38	מחלקות היפותזות סופיות	4.13
39	מצוער הסיכון אמפירי ERM	4.14
41	מחלקות היפותזות סופיות - סיכום	4.15
41	מידד VC	4.16
42	הגדרות פורמלאיות	4.17
<b>43</b>	<b>הרצאה 5 - המשך מודל PAC</b>	<b>5</b>
43	המשפט היסודי של הלמידה הסטטיסטית	5.1
43	הרחבת המודל	5.2
45	למידות Agnostic-PAC	5.3
46	מצוער השגיאה האמפירית	5.4
46	זיכרון - החוק החלש של המספרים גדולים	5.5
47	המשפט היסודי של הלמידה הסטטיסטית עם Agnostic PAC	5.6
47	חלק ראשון	5.7
47	חלק שני	5.8
<b>51</b>	<b>הרצאה 6 - שיטות ועידה</b>	<b>6</b>
51	הטיה ושותנות	6.1
51	החלotas ועודה	6.2
52	עודות ב- IML	6.3
52	שיטת Bootstrap	6.4
53	שיטת Bagging	6.5
54	די-קוראלציה + Random Forest	6.6
55	שיטת Boosting	6.7
56	שיטת Adaboost	6.8
58	שיטת Boosting מנוק' מבט של מודל PAC	6.9
59	השוואה Bagging & Boosting	6.10
<b>60</b>	<b>הרצאה 7 - רגולרייזציה ובחירה מודל</b>	<b>7</b>
60	רגולרייזציה	7.1

61	עיצים מסויימים . . . . .	7.2
62	בחזקה לרגרסיה . . . . .	7.3
64	אלג' Ridge Regression . . . . .	7.4
65	אלג' Lasso Regression . . . . .	7.5
65	אלג' רגרסיה לוגיסטיבית $\ell_1$ -regularized . . . . .	7.6
66	בחירה מודל והערכתו . . . . .	7.7
<b>70</b>	<b>הרצאה 8 - למידה Unsupervised</b>	<b>8</b>
70	בעיה שונה לגמרי . . . . .	8.1
70	הורדת מימד - PCA . . . . .	8.2
77	קלאסטרינג Clustering . . . . .	8.3
79	קלאסטרינג ספקטRALי . . . . .	8.4
<b>81</b>	<b>הרצאה 9 - Kernels Methods</b>	<b>9</b>
81	רעיון כללי . . . . .	9.1
81	דוגמא ראשונה - Polynomial Fitting . . . . .	9.2
82	ה- Kernel Trick . . . . .	9.3
85	אלגוריתמי Kernel . . . . .	9.4
88	מספר Kernels מפורטים . . . . .	9.5
89	אפיון לפונקציות גרעין . . . . .	9.6
<b>90</b>	<b>הרצאה 10 - אופטימיזציה קמורה 1 - Gradient Descent</b>	<b>10</b>
90	הקדמה . . . . .	10.1
90	הגדרות . . . . .	10.2
95	תת-גרדיינטים Sub-gradients . . . . .	10.3
96	אופטימיזציה קמורה . . . . .	10.4
97	בעיות למידה קמורות . . . . .	10.5
98	אלג' Gradient Descent . . . . .	10.6
100	אלג' Sub-gradient Descent . . . . .	10.7
<b>102</b>	<b>הרצאה 11 - SGD ולמידה عمוקה (רשתות נוירונים)</b>	<b>11</b>
102	ה- SGD . . . . .	11.1
104	רשת נוירונים קטנה . . . . .	11.2

# 1 הרצאה 1 - מבוא

## 1.1 אומדיים

יהי משתנה מקרי  $X$  המתפלג  $p_X$ , ודוגמאות דגימות  $\{X_i\}_{i=1}^m$  המקיימות  $X$  נרצה לשערך כל מיני תכונות של ההתפלגות. לדוגמה תוחלת, שונות. נסמן את האומדיים לתוחלת ולשונות ע"י  $\hat{\mu}$ ,  $\hat{\sigma}$ .

### 1.1.1 אומד לתוחלת Sample Mean

$$\text{האומדן לתוחלת מוגדר על ידי } \hat{\mu} = \frac{1}{m} \cdot \sum_{i=1}^m x_i.$$

### 1.1.2 אומד לשונות Sample Variance

$$\text{האומדן לשונות מוגדר על ידי } \hat{\sigma}_X^2 = \frac{1}{m-1} \cdot \sum_{i=1}^m (x_i - \hat{\mu})^2.$$

- אם התוחלת של האומד שווה לערך אותו הם אומד, הוא יקרא **בלתי מוטה Unbiased**.

## 1.2 התפליגות בממד גובה

### 1.2.1 וקטור אקראי \(\mathbf{X}\) משתנה מקרי וקטורי

אוסף של משתנים מקרים  $X_1, \dots, X_n$  מסודרים באקראי בוקטור.  $\mathbf{X} = (X_1, \dots, X_n)^T$

### 1.2.2 התפליגות משותפת

בහינתן אוסף משתנים מקרים  $X_1, \dots, X_n$ , נגידר את ההתפליגות המשותפת כהתפליגות שכל משתנה מקרי יפול בטוחח \(\backslash\) סט מסוים.

### 1.2.3 צפיפות משותפת

לשתי משתנים מקרים יש פונקציית צפיפות משותפת אם קיימת פונקציה  $f_{XY} : \mathbb{R}^2 \rightarrow \mathbb{R}$  כך שלכל  $A \in \mathbb{R}^2$  מתקיים:

$$\mathcal{D}((X, Y) \in A) = \int_A f_{XY}(x, y) dx dy$$

פונקציה זו נקראת **פונקציית הצפיפות המשותפת**.

### 1.2.4 התפליגות נורמלית

משתנה מקרי  $X$  מתפלג נורמלי עם תוחלת  $\mu$  ושונות  $\sigma^2$  אם פונקציית הצפיפות שלו מוגדרת ע"י:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ונסמן  $X \sim \mathcal{N}(\mu, \sigma^2)$

**1.2.5 התפלגות נורמלית רב-מימדית**

וקטור מ"מ  $X$  מתפלג נורמלי רב מימדי עם תוחלת  $\mu$  ומטריצת שוניות משותפות  $\Sigma$  אם פונקציית הצפיפות שלו מוגדרת ע"י:

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{(x - \mu)^T \cdot \Sigma^{-1} \cdot (x - \mu)}{2}\right)$$

$$\text{ונסמן } X \sim \mathcal{N}(\mu, \Sigma)$$

**1.2.6 התפלגות שלילית**

בהתאם התפלגות משותפת  $p_{X,Y}$ , נוכל לחשב את הצפיפות של  $X$  או של  $Y$  ע"י:

$$p_X(x) = \int_y p_{X,Y}(x, y) dy$$

**דוגמה**

תהי  $X \sim \mathcal{N}(\mu, \Sigma)$  עבור  $\mu = (\mu_1, \mu_2)$ . נרצה למצוא את הצפיפות השולית של  $x_1$ .

$$\begin{aligned} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{(x - \mu)^T \Sigma^{-1} (x - \mu)}{2}\right) &= \frac{1}{\sqrt{(2\pi)^n \left| \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right|}} \exp\left(-\frac{(x - \mu)^T \left[ \begin{array}{cc} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{array} \right] (x - \mu)}{2}\right) \\ &= \frac{1}{\sqrt{(2\pi)^n \sigma_1^2 \sigma_2^2}} \exp\left(-\frac{(x - \mu)^T \left[ \begin{array}{cc} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{array} \right] (x - \mu)}{2}\right) \\ &= \frac{1}{\sqrt{(2\pi)^n \sigma_1^2 \sigma_2^2}} \exp\left(-\frac{1}{2} \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 - \frac{1}{2} \left( \frac{x_2 - \mu_2}{2\sigma_2} \right)^2\right) \\ &= \frac{1}{\sqrt{(2\pi)^n \sigma_1^2}} \exp\left(-\frac{1}{2} \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2\right) \frac{1}{\sqrt{(2\pi)^n \sigma_2^2}} \exp\left(-\frac{1}{2} \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2\right) \end{aligned}$$

$$\begin{aligned}
& \text{כעת, נבחן כי האינטגרל } 1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^n \sigma_2^2}} \exp\left(-\frac{1}{2} \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right) dx_2 = 1 \\
& \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{(x - \mu)^T \Sigma^{-1} (x - \mu)}{2}\right) dx_2 \\
& = \frac{1}{\sqrt{(2\pi)^n \sigma_1^2}} \exp\left(-\frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^n \sigma_2^2}} \exp\left(-\frac{1}{2} \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right) dx_2 \\
& = \frac{1}{\sqrt{(2\pi)^n \sigma_1^2}} \exp\left(-\frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right) \\
& .x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)
\end{aligned}$$

### 1.2.7 מטריצת השוניות המשותפת

יהי וקטור משתנים מקרים  $X = (X_1, \dots, X_d)^T$ . נגדיר את מטריצת השוניות המשותפת  $\Sigma$  כמטריצה  $d \times d$  המקיים כי לכל  $i, j$   $\Sigma_{i,j} = Cov(X_i, X_j)$  בambilים אחרות נקבל:

$$\Sigma = \begin{bmatrix} \mathbb{E}(X_1 - \mathbb{E}(X_1)) \mathbb{E}(X_1 - \mathbb{E}(X_1)) & \cdots & \mathbb{E}(X_1 - \mathbb{E}(X_1)) \mathbb{E}(X_d - \mathbb{E}(X_d)) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(X_d - \mathbb{E}(X_d)) \mathbb{E}(X_1 - \mathbb{E}(X_1)) & \cdots & \mathbb{E}(X_d - \mathbb{E}(X_d)) \mathbb{E}(X_d - \mathbb{E}(X_d)) \end{bmatrix}$$

בכתב וקטוריים נקבל  $\Sigma = \mathbb{E}((X - \mathbb{E}(X)) \cdot (X - \mathbb{E}(X))^T)$ . נבחן כי האיברים על האלכסון של  $\Sigma$  הם  $Var(X_i)$  וכי  $\Sigma$  היא סימטרית.

### 1.2.8 אומד למטריצת השוניות המשותפת

יהי  $X = (X_1, X_2)$  וקטור משתנים מקרים. ניקח דוגמה מ- $m$  אנשים שונים, ונסמן את המידע בצורה הבאה:  
הדוגמאות של  $X_1$  יסומנו ע"י  $x_{1,1}, \dots, x_{1,m}$  בעוד הדוגמאות של  $X_2$  יסומנו ע"י  $x_{2,1}, \dots, x_{2,m}$ .  
לחילופין, נוכל לייצג את המידע במטריצה  $X \in \mathbb{R}^{d \times m}$ .  
לפי הגדרת האומד לשונות שהגדכנו מקודם, נקבל כי:

$$\forall i \in \{1, 2\} : \hat{\sigma}_{X_i}^2 = \frac{1}{m-1} \cdot \sum_{j=1}^m (x_{i,j} - \hat{\mu}_i)^2$$

כעת, נגדיר את האומד לשונות המשותפת  $\hat{\sigma}(X_1, X_2)$  באופן הבא:

$$\hat{\sigma}(X_1, X_2) = \frac{1}{m-1} \cdot \sum_{j=1}^m (x_{1,j} - \hat{\mu}_1)(x_{2,j} - \hat{\mu}_2)$$

האומד למטריצת השוניות המשותפת  $\hat{\Sigma}$  מוגדרת ע"י  $\hat{\Sigma}_{i,j} = \hat{\sigma}(X_i, X_j)$ . היא מטריצה סימטרית ואיברי האלכסון שלה שווים לאיברי  $Var(X_i)$ .

בכתיב וקטורי מטריציוני נקבל:

$$\hat{C} = \frac{1}{m-1} \cdot \sum_{j=1}^m (\mathbf{x}_i - \bar{\mathbf{X}}) (\mathbf{x}_i - \bar{\mathbf{X}})^T$$

עבור  $\bar{\mathbf{X}} \in \mathbb{R}^{d \times 1}$  הוא וקטור האומדים לתוחלת  $(\hat{\mu}_1, \hat{\mu}_2)$ .

אם  $\hat{C} = \frac{XX^T}{n-1}$  מטריצת תוחלות כך ש-  $\bar{\mathbf{X}}_{i,j} = \hat{\mu}_i$  אזי ניתן לחשב את המטריצת השוניות המשותפות ע"י

### 1.2.9 דוגמא

дані висота та маса 3 осіб.  $X^T = \begin{bmatrix} 150 & 45 \\ 170 & 74 \\ 184 & 79 \end{bmatrix}$

матки:  $\bar{\mathbf{X}}^T = \begin{bmatrix} 168 & 66 \\ 168 & 66 \\ 168 & 66 \end{bmatrix}$ , і отже  $\hat{\mu}_2 = \frac{45+74+79}{3} = 66$  і  $\hat{\mu}_1 = \frac{150+170+184}{3} = 168$ , і отже

$$X_{\text{centered}}^T = X^T - \bar{\mathbf{X}}^T = \begin{bmatrix} 150 & 45 \\ 170 & 74 \\ 184 & 79 \end{bmatrix} - \begin{bmatrix} 168 & 66 \\ 168 & 66 \\ 168 & 66 \end{bmatrix} = \begin{bmatrix} -18 & -21 \\ 2 & 8 \\ 16 & 13 \end{bmatrix}$$

і отже ми отримали:

$$\hat{C} = \frac{X_{\text{centered}} X_{\text{centered}}^T}{3-1} = \frac{\begin{bmatrix} 150 & 170 & 184 \\ 45 & 74 & 79 \end{bmatrix} \cdot \begin{bmatrix} -18 & -21 \\ 2 & 8 \\ 16 & 13 \end{bmatrix}}{2} = \frac{\begin{bmatrix} 584 & 602 \\ 602 & 674 \end{bmatrix}}{2} = \begin{bmatrix} 292 & 301 \\ 301 & 337 \end{bmatrix}$$

### 1.3 טרנספורמציה ליניארית על דatas

נתבונן במטריצת שוניות משותפות במימד 2:  $d = 2$

$$C = \begin{bmatrix} \sigma(X_1, X_1) & \sigma(X_1, X_2) \\ \sigma(X_2, X_1) & \sigma(X_2, X_2) \end{bmatrix}$$

נרצה להראות איך טרנספורמציה ליניארית משפיעה על הדאטא וכותזאה על המטריצה  $C$ .

ניקח נקודות אקראיות ממשתנים מקריים בעלי תוחלת אפס  $\bar{X}_1 = \bar{X}_2 = 0$  ושונות זהה  $\sigma_{X_1}^2 = \sigma_{X_2}^2 = \sigma^2$  המשמעות היא ש-  $X_1, X_2$  הם בלתי מתואמים, ולכן המטריצה  $C$  נראה כך:

$$C = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

### 1.3.1 טרנספורמציה ע"י מתיחה במטריצה סקלרית

כעת, נבצע טרנספורמציה ליניארית על הדadata שלנו בעזרת המטריצה המתיחה  
cut S =  $\begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix}$  בעזרת המטריצה המתיחה

cut מטריצת השוניות המשותפות של הדadata החדש היא:

$$\frac{SX(SX)^T}{n-1} = S \cdot \frac{XX^T}{n-1} \cdot S^T = SCS^T = \begin{bmatrix} (s_1\sigma)^2 & 0 \\ 0 & (s_2\sigma)^2 \end{bmatrix}$$

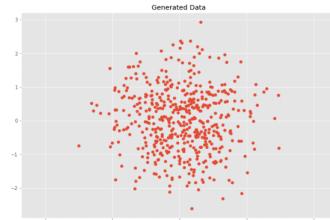


Figure 3: Uncorrelated Variables



Figure 4: Uncorrelated Scaled Variables

המשתנים המקררים  $X_1, X_2$  עדין בלתי מתאימים, אך cut הם מוגדים ע"י  $s_1, s_2$  בהתאם.  
ניתן לראות כי  $s_1 > s_2$  מכיוון שהשונות קטנה יותר לאורך ציר ה- $x$  מאשר לאורך ציר ה- $y$ ,  
והפיזור שהוא בצורת מעגל הפך להיות בצורת אליפסה.

### 1.3.2 טרנספורמציה ע"י סיבוב

נבצע טרנספורמציה ליניארית על הדadata שלנו בעזרת המטריצה המתיחה כמפורט,

ול-  $R$  מטריצת הסיבוב הבאה  $R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$  עבור  $\theta$  זווית הסיבוב.

הdataset החדש מתקבלת ע"י  $RSX$  ומטריצת השוניות המשותפות היא cut:

$$\begin{aligned} \frac{RSX(RSX)^T}{n-1} &= RS \cdot \frac{XX^T}{n-1} \cdot S^T R^T = RSCS^TR^T \\ &= R \cdot \begin{bmatrix} (s_1\sigma)^2 & 0 \\ 0 & (s_2\sigma)^2 \end{bmatrix} \cdot R^T = \sigma^2 \times \begin{bmatrix} s_1^2\cos^2\theta + s_2^2\sin^2\theta & \sin\theta\cos\theta(s_1^2 - s_2^2) \\ \sin\theta\cos\theta(s_1^2 - s_2^2) & s_1^2\sin^2\theta + s_2^2\cos^2\theta \end{bmatrix} \end{aligned}$$

נבחן כי כאשר מوطחים בצורה סימטרית, כך ש-  $s_1 = s_2$ , נקבל כי האיברים מחוץ לאלכסון הראשי עדין מתאפסים,  
ועל כן המשתנים נשארים בלתי מתאימים. אחרת, ביצוע הטרנספורמציה הנ"ל הופכת את המשתנים למתאימים, ונקבל:

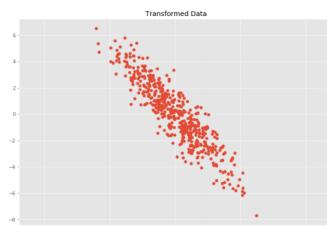


Figure 5: Correlated Scaled Variables

## 1.4 א-שוויונות וחסמים

### 1.4.1 א-שוויון מركוב

עבור משתנה מקרי א-שלילי  $X$  בעל תוחלת סופית וסקלר  $t > 0$  מתקיים:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$$

### 1.4.2 א-שוויון צ'בישב

עבור משתנה מקרי  $X$  בעל תוחלת ושונות סופיים, ו-  $0 < t >$  מתקיים:

$$\mathbb{P}(X - \mathbb{E}(X) \geq t) \leq \frac{\text{Var}(X)}{t^2}$$

### 1.4.3 מסקנה

יהיו  $X_1, \dots, X_m$  משתנים מקרים בלתי תלויים ושווי התפלגות בעלי שונות סופית.

אם נסמן  $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$  נקבל כי לכל  $a > 0$  מתקיים:

$$\mathbb{P}(|\bar{X} - \mathbb{E}(\bar{X})| \geq a) = \mathbb{P}(|\bar{X} - \mathbb{E}(X)| \geq a) \leq \frac{\text{Var}(\bar{X})}{a^2} = \frac{\text{Var}(X)}{ma^2}$$

## 1.5 חיזוי מטבע

יהי משתנה מקרי  $Z$  כלשהו המתפלג  $Ber(p)$  עם סטייה כלשהי  $.p - \frac{1}{2}$ .

יהיו  $Z_1, \dots, Z_m$  רצף של משתנים מקרים בלתי תלויים ושווי התפלגות מההתפלגות  $\mathcal{D}_Z$ .

נסמן מוגן המתקבל מהמשתנים המקרים הנ"ל ע"י  $S$ , כך שמתקיים  $|S| = m$ .

נדיר אלגוריתם למידה  $A$  אשר קיבל דוגמה  $S$  ויחזר את השعروץ שלו לערך של  $p$ , אותו נסמן ע"י  $\hat{A}(S)$  או  $\hat{p}$ .

מכיוון שהדוגמה הינה סופית ולא יכולה לתאר באופן מלא את ההתפלגות, נגידר פרמטר דיוק  $\varepsilon > 0$ .

ונאפשר  $-\hat{p}$  לצאתו מ-  $p$  ב-  $\varepsilon$ . בנוסף, קיימת האפשרות שנתקבל מוגן לא מייצג, נגידר פרמטר בטיחות  $\delta > 0$

ונאפשר למוגן  $\hat{A}(S)$  להתרחש בהסתברות של לכל היותר  $\delta$ .

### 1.5.1 באוף פורמלי

אלגוריתם Mistake Bound Learning עבור חיזוי מטבע הוא פונקציה  $\hat{A}(S)$  המחזיר  $\hat{p} \in [0, 1]$  עבור  $S \in \{0, 1\}^m$

ומקיים את התנאים הבאים:

- לכל  $\delta, \varepsilon \leq 0$  קיים מספר שלם א-שלילי  $m_A(\varepsilon, \delta)$  כך שבгинטן מוגן  $S$  של  $m$  הגרלות הנוגן מההתפלגות

עבור  $0 \leq p \leq 1$ , הסתברות לכך ש-  $\hat{A}(S)$  חסום ע"י  $[\hat{p} - \varepsilon, \hat{p} + \varepsilon] \leq \delta$  היא לפחות  $m \geq m_A(\varepsilon, \delta)$ .

- אם מתקיים  $m < m_A(\varepsilon, \delta)$  אז קיים  $0 \leq p \leq 1$  כך ש-  $\hat{A}(S)$  מוגן  $|\hat{p} - p| > \varepsilon$ .

הfonktsia  $m_A(\varepsilon, \delta) : [0, 1] \times [0, 1] \rightarrow \mathbb{N}$  נקראת Sample Complexity.

### 1.5.2 בחירת האלגוריתם

נתבונן באלגוריתם הבא העומד בדרישות ההגדרה הנ"ל.  
בහינתן דוגמה  $S = (z_1, \dots, z_m)$ , הדריך היירה ביותר לשערץ את  $p$  הוא ממוצע האמפירי:

$$\hat{p}(S) = \frac{1}{m} \cdot \sum_{i=1}^m z_i$$

בעצם "נספור" את כמות האחדות שקיבלו, ונחלק במספר הטילות הכלול.

נבחן כי האומד שלנו הוא בעצם האומד לתוחלת ולכן הוא אומד חסר הטיה ועבור דוגמה כלשהי  $S$ , נוכל לקבל  $0 = |p - \hat{p}|$ .  
עת נרצה לבדוק את טיב האומד. כמה הטילות מטבע נזדקק על מנת להבטיח כי  $\hat{p}$  בהסתברות גבוהה מאוד קרוב ל-  $p$ ?

### 1.5.3 שערוך ה- Sample Complexity בעזרת אי-שוויון מרקוב

על מנת להפעיל את אי-שוויון מרקוב ישירות, משתמש ב-  $|\hat{p} - p|$  כמשתנה המקרי שלנו.

$$\mathcal{D}_p^m (|\hat{p} - p| \geq \varepsilon) \leq \frac{1}{\varepsilon} \cdot \mathbb{E}(\hat{p} - p) = \frac{1}{\sqrt{4m\varepsilon^2}}$$

אם נבחר  $\lceil \frac{1}{4\varepsilon^2} \cdot \frac{1}{\delta^2} \rceil \leq m$  נקבל כי הצד הימני קטן שווה ל-  $\delta$ .  
 $\mathcal{D}_p^m (|\hat{p} - p| \geq \varepsilon) \leq m_A(\varepsilon, \delta) \leq \lceil \frac{1}{4\varepsilon^2} \cdot \frac{1}{\delta^2} \rceil$  אם נדגם  $(\varepsilon, \delta) \in (0, 1)$  דוגמאות האלגוריתם שלנו ישיג  $\delta$

### 1.5.4 שערוך ה- Sample Complexity בעזרת אי-שוויון צ'בישוב

השונות של משתנה מקרי ברנולי מוגדרת ע"י  $p(1-p)$ , כאשר ביטוי זה חסום מלמעלה ע"י  $\frac{1}{4}$ .  
לפי מסקנה 1.4.3 נקבל כי:

$$\mathcal{D}_p^m (|\hat{p} - p| \geq \varepsilon) = \mathcal{D}_p^m (|\hat{p} - \mathbb{E}(\hat{p})| \geq \varepsilon) \leq \frac{p(1-p)}{m\varepsilon^2} \leq \frac{1}{4m\varepsilon^2}$$

**מסקנה ביןים:** ה- Sample Complexity של חיזוי מטבע חסום מלמעלה ע"י  $m(\varepsilon, \delta) \leq \lceil \frac{1}{4\varepsilon^2} \cdot \frac{1}{\delta} \rceil$ .

### 1.5.5 שערוך ה- Sample Complexity בעזרת אי-שוויון הופדיינג

#### הגדרה

יהיו  $a_i \leq X_i \leq b_i$  משתנים מקרים בלתי תלויים וחסומים כך ש-  
נסמן  $X_1, \dots, X_m$  מתקיים:  $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$

$$\mathbb{P} (|\bar{X} - \mathbb{E}(\bar{X})| \geq \varepsilon) \leq 2 \exp \left( \frac{-2m^2\varepsilon^2}{\sum_{i=1}^m (b_i - a_i)^2} \right)$$

**מסקנה**

יהיו  $a \leq X_i \leq b$  משתנים מקרים בלתי תלויים ושווי התפלגות, בעלי תוחלת  $\mathbb{E}(X)$  וחסומים כך ש-  
נסמן  $X_1, \dots, X_m$ . מתקיים:  $\overline{X} = \frac{1}{m} \sum_{i=1}^m X_i$ .

$$\mathbb{P}(|\overline{X} - \mathbb{E}(X)| \geq \varepsilon) \leq 2 \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$$

**הוכחה**

מתקיים  $\sum_{i=1}^m (b_i - a_i)^2 = m(b-a)^2$  והצבה בהגדרה מקיימת את המסקנה.

שימוש בא-שוויון הופding בבעית חיזוי המטבע יתן לנו את אי-השוויון הבא:

$$\mathcal{D}_p^m (|\hat{p} - p| \geq \varepsilon) \leq 2 \exp\left(\frac{-2m\varepsilon^2}{1^2}\right) = 2 \exp(-2m\varepsilon^2)$$

כעת, על ידי לקיחת  $\lceil \frac{1}{2\varepsilon^2} \cdot \log\left(\frac{2}{\delta}\right) \rceil \geq m$  דוגמאות נקבל כי ההסתברות חסומה מלמעלה ע"י  $\delta$  כפי שדרوش.

**1.5.6 לsicום**

אלגוריתם חיזוי המטבע ( $\hat{p}$ ) המשערק את  $p$  ע"י מספר האחדות שהתקבלו חלקו הסך הכלל של ההצלות  
.  $m_A(\varepsilon, \delta) \leq \lceil \frac{1}{2\varepsilon^2} \cdot \log\left(\frac{2}{\delta}\right) \rceil$  חסום מלמעלה ע"י Sample Complexity 1.5.1

## 2 הרצאה 2 - רגרסיה ליניארית

### 2.1 בעיה

אנו עובדים עבור חנות אונליין, ורוצים לחזות ערך פוטנציאלי מהלcool, בהינתן סט של דגימות עבור  $m$  לקוחות. נאוסף  $d$  תכונות על כל לקוח, ובסימונים שלנו  $\mathcal{X} = \mathbb{R}^d$ , וקטור בעל  $d$  תכונות.  $\mathcal{Y} = \mathbb{R}$ , ערך פוטנציאלי כספי.

הדאטה שיש לנו היא מהצורה  $S = \{(x_i, y_i)\}_{i=1}^m$ ,  $x_i \in \mathbb{R}^d$ ,  $y_i \in \mathbb{R}$ , עבור  $x_1, \dots, x_m$  ו-  $y_1, \dots, y_m$  נארגן את המידע בצורה הבאה:

$$X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ \vdots & & \\ - & x_m & - \end{bmatrix}_{m \times d}$$

ו-  $y^\top = (y_1, \dots, y_n) = y$  וקטור הערכים הפוטנציאליים.

**הנחת חסר רוש:** נניח כי קיימת  $\mathcal{Y} \rightarrow \mathcal{X}$  המקיים בדיקת  $y = f(x)$ , עבור דגימות  $(y, x)$ , ללא "רוש"  $z$  כלשהו. מטרתנו היא למצוא פונקציה  $f$  מתוך הדאטה שקיבלו נכך שנוכל לחזות ערך פוטנציאלי של לקוח מוקטור התכונות שלו.

#### 2.1.1 מחלוקת היפוטזות

אנחנו חייבים להניח כי קיימת מחלוקת היפוטזית  $\mathcal{H}$  וכלל למידה  $\mathcal{H} \in h$  כלשהו. נניח וגילינו כי הערך  $y$  של לקוח הוא פונקציה ליניארית של ערכי הווקטור  $x$ , אז נגדיר את מחלוקת היפוטזות הליניארית:

$$\mathcal{H}_{\text{lin}} = \left\{ (x_1, \dots, x_d) \mapsto w_0 + \sum_{i=1}^d x_i w_i \mid w_0, \dots, w_d \in \mathbb{R} \right\}$$

כל פונקציה  $h$  במחלוקת  $\mathcal{H}$  מוגדרת ע"י "משקלות" Intercept- $w_0$ ,  $\dots$ ,  $w_d$  Intercept- $w_1, \dots, w_d$  "משקלות". על מנת להיפטר מה Intercept-, נוסיף לכל דגימה  $x_i \in \mathbb{R}^d$  כורדינאט אפס שווה ל- 1. בעת נגדיר את  $(w_0, w_1, \dots, w_d)^\top = w$  ונקבל כי:

$$\mathcal{H}_{\text{lin}} = \{x \mapsto x^\top w \mid w \in \mathbb{R}^{d+1}\}$$

#### 2.1.2 מקרה ריאליובי מול לא ריאליובי

- **המקרה הריאליובי.** במקרה זה אנו מניחים כי  $f \in \mathcal{H}_{\text{lin}}$ , ונוכל לקוות לשחזר אותה. במקרה זה מתקיים  $w_i = x_i^\top w$  עבור  $i = 1, \dots, m$  ועבור  $w \in \mathbb{R}^{d+1}$ . במקרה זה יש לנו מערכת מטריצית  $X$  וקטור  $w$ . במקרה זה יש לה לפחות פתרון אחד.
- **המקרה הלא ריאליובי.** במקרה זה  $f$  אינו ליניארי, ולכן לא נוכל למצוא את  $f$  מכיוון ש-  $f \notin \mathcal{H}_{\text{lin}}$ . נוכל לקוות למצוא כלל  $h_S$  שմקורב את  $f$ . במקרה זה המערכת  $y = w$  אין בהכרח פתרון.

### 2.1.3 פירוש גיאומטרי

נסמן  $\varphi_i \in \mathbb{R}^m$  העמודה ה- $i$  של המטריצה  $X$ . אזי מתקיים:

$$X = \begin{bmatrix} | & | & & | \\ \varphi_0 & \varphi_1 & \cdots & \varphi_d \\ | & | & & | \end{bmatrix}$$

- במקרה הריאלי למערכת  $y = w \in \text{Im}(X)$  יש פתרון ולכון  $y \in \text{Im}(X)$ . למערכת יש פתרון יחיד או אין סוף פתרונות.
- אם  $y = w$ , אז  $y$  הינו צירוף ליניארי של העמודות  $\varphi_0, \dots, \varphi_d$ .
- ← אם קיימים פתרון יחיד אזי  $\varphi_0, \dots, \varphi_d$  בלתי תלויים, ומהווים בסיס למרחב העמודות של  $X$ , שהוא  $\text{Im}(X)$ .
- ← אם קיימים אינסוף פתרונות אזי  $\varphi_0, \dots, \varphi_d$  תלויים ליניארית.
- במקרה הלא ריאלי למערכת  $y = w \notin \text{Im}(X)$  אין פתרון ולכון  $y \notin \text{Im}(X)$ .

### 2.2 פונקציית Loss

פונקציית  $L(y, h_S(\mathbf{x}))$  הינה פונקציה שמודדת את השגיאה בין החיזוי לתగית האמיתית.

- פונקציית  $L(y, h_S(\mathbf{x})) = |y - h_S(\mathbf{x})|$ . מקיימת Absolute Value Loss.
  - פונקציית  $L(y, h_S(\mathbf{x})) = (y - h_S(\mathbf{x}))^2$ . מקיימת Squared Loss.
- בבעה זו נבחר להשתמש ב-Squared Loss.**

### 2.3 מזעור הסיכון האמפירי (ERM)

מכיוון שבחרנו למדוד את הביצועים ע"י  $L(y, h_S(\mathbf{x})) = (y - h_S(\mathbf{x}))^2$ , הגיוני שנרצה למצוא כלל  $h_S$  שմזער את הביטוי זהה עבור הדאטה שניתנה לנו.  $\sum_{i=1}^m L(y_i, h_S(\mathbf{x}_i))$  ועבור  $h_S(\mathbf{x}_i) = \{\mathbf{x}_i, y_i\}_{i=1}^m$  כלשהו מוגדר ע"י הסיכון האמפירי עבור הדאטה. במקרה שלנו, הסיכון האמפירי של הפונקציה הליניארית  $w \mapsto \mathbf{x}^\top w$  הוא:

$$\sum_{i=1}^m (y_i - \mathbf{x}_i^\top w)^2 = \|\mathbf{y} - Xw\|^2 = (\mathbf{y} - Xw)^\top \cdot \mathbf{y} - Xw$$

מזעור הסיכון האמפירי במקרה שלנו אומר למזער את סכום הריבועים של הסטייה בין ערכי  $y$  לבין הפונקציה הליניארית. בambilים אחרות, נבחר את הפונקציה הליניארית  $h_S \in \mathcal{H}_{\text{lin}}$  שהכי קרובה לערכי  $y$  למרחק ריבוע הטועות.

הסטייה  $w_i - y_i$  נקראת **היתר** \ Residual ה- $i$ , RSS( $w$ ) =  $\|\mathbf{y} - Xw\|^2$  נקרא **יתרת סכום הריבועים** \ Residual Sum of Squares ומסומן  $\mathcal{H}_{\text{lin}}$ .

## 2.4 איז איז נלמד?

על מנת ללמוד את הפונקציה הליניארית לפי מינימול הסיכון האמפירי, علينا למצוא את המינימום

$$\operatorname{argmin}_{w \in \mathbb{R}^d} \text{RSS}(\mathbf{w}) = \operatorname{argmin}_{w \in \mathbb{R}^d} \|y - Xw\|^2$$

עלינו למצוא  $\|w\|^2$ .  $h_S(x) = x^T y - Xw = \hat{w}$  ואז להגדיר  $\hat{w}$  שיטה זו תעבוד עבור שני המקרים שציינו:

- **במקרה הריאליiziabi** מכיוון ש-  $(X) \in \text{Im}(y)$  אנו יודעים שקיים לפחות פתרון אחד  $\hat{w}$  המקיים  $y = X\hat{w}$ . פתרון זה ישיג ערך מינימלי של אפס, ומכיון שהפונק' RSS חסומה מלמטה ע"י אפס נקבל פתרון מינימלי.
- **במקרה הלא ריאליiziabi** מכיוון ש-  $(X) \notin \text{Im}(y)$  לא קיים פתרון  $\hat{w}$  המקיים  $y = X\hat{w}$  ולכן נקווה למצוא וקטור מספק טוב במנוחי מעור סכום הריבועים.

בעת נשאלת השאלה, **איך נמצא את  $w$ ?**

## 2.5 מציאת המינימום

תנאי הכרחי לכך שוקטור  $w$  ימנמל את הביטוי  $\|y - Xw\|^2$  הוא שיטקיים:

$$\begin{aligned} \frac{\partial}{\partial w_j} \text{RSS}(w) &= \frac{\partial}{\partial w_j} \|y - Xw\|^2 \stackrel{\text{הגדרת נורמה}}{=} \frac{\partial}{\partial w_j} \sum_{i=1}^m (y_i - x_i^T w)^2 \\ &\stackrel{\text{לייניארית הנגזרת}}{=} \sum_{i=1}^m 2 \cdot (y_i - x_i^T w) \cdot (-x_i) = -2 \sum_{i=1}^m (x_i)_j \cdot (y_i - x_i^T w) = 0 \end{aligned}$$

עבור כל  $j \leq d$ . נוכל לכתוב את  $d + 1$  המשוואות הנ"ל בכתב מטריציוני באופן הבא:

$$\nabla \text{RSS}(w) = -2X^T \cdot (y - Xw) = 0$$

מכיוון שאנו ממערים תבנית בי-لينיארית בעלת מינימום, כל פתרון למשוואת הנ"ל בהכרח יהיה נקודת מינימום.

## 2.6 המשוואות הנורמליות

ראינו כי תנאי הכרחי לכך שוקטור  $w$  ימזרע את  $(w)$  RSS הוא שיהיה פתרון למשוואת  $X^T \cdot (y - Xw) = 0$ . משוואות אלה נקראות **משוואות נורמליות**.

### 2.6.1 בחזרה לגיאומטריה

נזכיר בסימון  $\varphi_i \in \mathbb{R}^m$  העמודה ה-  $i$  של המטריצה  $X$ . המקיים:

$$X = \begin{bmatrix} | & | & & | \\ \varphi_0 & \varphi_1 & \cdots & \varphi_d \\ | & | & & | \end{bmatrix}$$

נוכל להבחן כי המשוואות הנורמליות יכולות להיכתב באופן שקול ע"י  $\langle \varphi_i | y - Xw \rangle = 0$  עבור  $0 \leq i \leq d$ . מכיוון שראינו כי  $\varphi_0, \dots, \varphi_d$  פורשים את תת-המרחב  $(X)^\perp$ , פתרו שמשוואות הנורמליות חיבר לקיטים  $(X)^\perp$ . אם נסמן  $\hat{z} = y - \hat{y} \in \text{Im}(X)^\perp$  ונגדיר את **קטור היתרונות** \ Residual Vector ע"י נקבל כי  $\hat{z} \in \text{Im}(X)^\perp$  הטענה אורתוגונלית של הוקטור  $y$  על  $\text{Im}(X)$ .

## 2.7 פתרת המשוואות הנורמליות

ראשית, נניח כי מספר הדגימות גדול ממספר התכונות, כלומר  $d+1 \geq m$ . נרצה לפתור את  $w$ .

### 2.7.1 מקרה ראשון - העמודות של $X$ בלתי תלויות ליניארית

מקרה זה מתקיים אם ורק אם  $\dim(\text{Ker}(X)) = 0$ , אם ורק אם  $\dim(\text{Ker}(X)) = 1$ . במקרה זה קיים ייחיד למשוואות הנורמליות:

$$w' = (X^\top X)^{-1} \cdot X^\top y$$

בחינה גיאומטרית, יש דרך ייחודית לכתוב את  $y$  (המקרה הריאלי-בלוי) או את הטליה של  $y$  על  $\text{Im}(X)$  (המקרה הלא ריאלי-בלוי), כצירוף ליניארי של העמודות של המטריצה  $X$ .

### 2.7.2 מקרה שני - העמודות של $X$ תלויות ליניארית

עבור מקרה זה ישנו אינסוף פתרונות למשוואות הנורמליות. בבחינה גיאומטרית, יש אינסוף דרכים לכתוב את  $y$  (המקרה הריאלי-בלוי) או את הטליה של  $y$  על  $\text{Im}(X)$  (המקרה הלא ריאלי-בלוי), כצירוף ליניארי של העמודות של המטריצה  $X$ . כיצד נפתר את המשוואות הנורמליות  $w$ ?

## 2.8 פירוק SVD

### 2.8.1 עובדות ותכונות

1. המטריצה  $X_{m \times (d+1)}$  שלנו ניתנת לפירוק ע"י  $X = U\Sigma V^\top$ . עבור  $U_{m \times m}$  אורתוגונלית,  $\Sigma_{m \times (d+1)}$  אלכסונית ו-  $V_{(d+1) \times (d+1)}$  אורתוגונלית.

2. איברי האלכסון של המטריצה  $\Sigma$  הם **הערכיות הסינגולריות** של המטריצה  $X$  ומסודרים  $0 \geq \sigma_1 \geq \dots \geq \sigma_{d+1} \geq \dots$ .

3. עמודות המטריצה  $U$  הם הוקטוריים העצמיים של המטריצה  $XX^\top$ . הם נקראים **הוקטוריים הסינגולריים השמאליים** של  $X$ .

4. עמודות המטריצה  $V$  הם הוקטוריים העצמיים של המטריצה  $X^\top X$ . הם נקראים **הוקטוריים הסינגולריים הימניים** של  $X$ .

5.  $\sigma_1^2, \dots, \sigma_{d+1}^2$  הם הערכיות העצמיים המשותפות של המטריצות  $XX^\top$  ו-  $X^\top X$ .

6. סדר העמודות של המטריצות  $V, U$  הוא כך שהעמודה ה- $i$  מתאימה לערך העצמי  $\sigma_i^2$ .

7. **כל מטריצה ניתנת לפירוק SVD.** פירוק זה איננו יחיד.

8. (**חשוב**) מתקיים כי  $\dim(\text{Ker}(X)) = 0$  אם ורק אם  $\sigma_{d+1} > 0$ , כלומר כל הערכים הסינגולריים הם חיוביים ממש. פירוק ה-SVD נותן באופן מיידי את המימדים של הגרעין והתמונה של  $X$ , ובכך גם האם  $X$  הינה הפיכה או לא.

## 2.8.2 המטריצה $\Sigma^\dagger$

המטריצה  $\Sigma$  ופירוק ה-SVD של  $X \in \mathbb{R}^{m \times (d+1)}$  צורת Moonre-Penrose Pseudoinverse  $X^\dagger = V\Sigma^\dagger U^\top$  היא עבור  $\Sigma^\dagger$  מטריצה אלכסונית ממימד  $(d+1) \times m$  בעלת אלכסון המוגדר באופן הבא:

$$\Sigma_{i,i}^\dagger = \begin{cases} \frac{1}{\sigma_i} & \sigma_i > 0 \\ 0 & \text{else} \end{cases}$$

ונגיד  $\mathbf{y} = \hat{\mathbf{w}}$ . **עובדת:** הוקטור  $\hat{\mathbf{w}}$  הוא תמיד פתרון למשוואות הנורמליות.

○ אם המטריצה  $X^\top X$  הופכית, אז

$$\begin{aligned} \hat{\mathbf{w}} &= (X^\top X)^{-1} \cdot X^\top \mathbf{y} = \left( (U\Sigma V^\top)^\top \cdot U\Sigma V^\top \right)^{-1} \cdot (U\Sigma V^\top)^\top \cdot \mathbf{y} \\ &= (V\Sigma^\top U^\top \cdot U\Sigma V^\top)^{-1} \cdot V\Sigma^\top U^\top \cdot \mathbf{y} = (V\Sigma^\top \Sigma V^\top)^{-1} \cdot V\Sigma^\top U^\top \cdot \mathbf{y} \\ &= (V^\top)^{-1} \tilde{\Sigma}^{-1} V^{-1} \cdot V\Sigma^\top U^\top \cdot \mathbf{y} = V\tilde{\Sigma}^{-1} \Sigma^\top U^\top \cdot \mathbf{y} = V\Sigma^\dagger U^\top \mathbf{y} = X^\dagger \mathbf{y} \end{aligned}$$

ומתקיים השוויון שהגדכנו לעיל. (המטריצה  $\tilde{\Sigma}$  היא מסדר  $(d+1) \times (d+1)$  עם  $(\tilde{\Sigma}_{i,i}) = \Sigma_{i,i}^2$ )

○ אם המטריצה  $X^\top X$  אינה הופכית, אז נסמן  $r = \text{rank}(X)$ , המקיימים  $1 \leq r < d+1$  מכיוון שהגרעין של  $X$  אינו טריוויאלי במקרה זה. הערכים הסינגולריים של המטריצה  $\Sigma$  מקיימים  $\sigma_r > \dots > \sigma_1 \geq 0$ .  
יהי  $X = U\Sigma V^\top$  פירוק SVD של  $X$ . העמודות של  $U, V$  מהווים בסיס אורטוגונלי לאربעת תתי המרחב הבאים:

$$\begin{array}{lll} U_{\mathcal{R}} \in \mathbb{R}^{m \times r} & \mathcal{R}(X) & = \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\} \\ V_{\mathcal{R}} \in \mathbb{R}^{(d+1) \times r} & \mathcal{R}(X^\top) & = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_r\} \\ V_{\mathcal{N}} \in \mathbb{R}^{(d+1) \times (d+1-r)} & \mathcal{N}(X) & = \text{span}\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_{d+1}\} \\ U_{\mathcal{N}} \in \mathbb{R}^{m \times (m-r)} & \mathcal{N}(X^\top) & = \text{span}\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_m\} \end{array}$$

נגיד בנוסף את  $\mathcal{S} \in \mathbb{R}^{r \times r}$  מטריצה אלכסונית בעלת  $r$  הערכים הסינגולריים החיוביים על האלכסון.

נגיד את הצורה הקומפקטיבית של  $X$  בעזרת הסימונים הנ"ל:

$$X := U\Sigma V^\top = \begin{bmatrix} U_{\mathcal{R}} & U_{\mathcal{N}} \end{bmatrix} \cdot \begin{bmatrix} \mathcal{S} & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} V_{\mathcal{R}}^\top \\ V_{\mathcal{N}}^\top \end{bmatrix} = U_{\mathcal{R}} \cdot \mathcal{S} \cdot V_{\mathcal{R}}^\top = \tilde{U} \tilde{\Sigma} \tilde{V}^\top$$

כעת מתקיים במשואה  $X^\top \mathbf{y} = X^\top X \mathbf{w}$

$$\tilde{V} \tilde{\Sigma}^\top \tilde{U}^\top \mathbf{y} = (\tilde{V} \tilde{\Sigma}^\top \tilde{U}^\top) \tilde{U} \tilde{\Sigma} \tilde{V}^\top \mathbf{w} \Rightarrow \tilde{V} \tilde{\Sigma}^\top \tilde{U}^\top \mathbf{y} = \tilde{V} \tilde{\Sigma}^\top \tilde{\Sigma} \tilde{V}^\top \mathbf{w}$$

נכפול את שני האגפים משמאל במטריצה  $\tilde{V}^\top$  ונקבל:

$$\tilde{\Sigma}^\top \tilde{U}^\top \mathbf{y} = \tilde{\Sigma}^2 \tilde{V}^\top \mathbf{w}$$

מכיוון ש- $\tilde{V}$  אורתוגונלית ולכן גם  $\tilde{V}^\top \cdot \tilde{V} = I$ , ומתקיים  $I \cdot \tilde{\Sigma}^{-2} = \tilde{\Sigma}^{-2}$ . בicut נכפול את שני האגפים משמאל ב-

$$\tilde{V} \tilde{\Sigma}^{-2} \cdot \tilde{\Sigma}^\top \tilde{U}^\top \mathbf{y} = \tilde{V} \tilde{\Sigma}^{-2} \tilde{\Sigma}^2 \tilde{V}^\top \mathbf{w} \Rightarrow \mathbf{w} = \tilde{V} \tilde{\Sigma}^{-1} \tilde{U}^\top \mathbf{y}$$

icut נרצה להרחב חזרה את הצורה הקומפקטיבית של SVD ע"י שימוש בהגדרת Moonre-Penrose Psuedoinverse :

$$\hat{\mathbf{w}} = \tilde{V} \tilde{\Sigma}^{-1} \tilde{U}^\top \mathbf{y} = V \Sigma^\dagger U^\top \mathbf{y} = X^\dagger \mathbf{y}$$

מתקיים כי  $\hat{\mathbf{w}}$  הוא הפתרון **בעל הנורמה המינימלית**:  $\|\hat{\mathbf{w}}\| = \min \{\|\mathbf{w}\| \mid X\mathbf{w} = \mathbf{y}\}$ .

### 2.8.3 למה לשימוש בפירוק SVD בשיביל ללמידה את ש?

1. עובד תמיד. גם במקרה הריאלייזבילי וגם במקרה הלא ריאלייזבילי. גם עבור פתרון יחיד וגם עבור אינסוף.

2. יציב מבחינת נומריאת.

### 2.8.4 איך נהפוך את הפתרון בעזרת SVD ליציב מבחינה נומריאת?

לפעמים המטריצה  $X^\top X$  היא הפיכה מבחינה פורמאלית, אך בפועל היא מאוד קרובה להיות סינגולרית (לא הפיכה). במקרה זה קורה כאשר העמודות של  $X$  הם כמעט co-linear או אם עמודה אחת של  $X$  היא **במעט צירוף** ליניארי של השאר. במקרה זה, יהיו ערכי סינגולרים של  $X$  בעלי ערך קטן מאוד, ובכך  $\frac{1}{\sigma_i}$  עלול להיות גדול מאוד, ולאבד מהדיוק בגלל שיטת ה- floating point שבאורטה מייצגים המחשבים את המספרים. **כיצד נפתר בעיה זו?** נגידר סוף כלשהו, ונגידר:

$$\Sigma_{i,i}^{\dagger, \varepsilon} = \begin{cases} \frac{1}{\sigma_i} & \sigma_i > \varepsilon \\ 0 & \text{else} \end{cases}$$

ובכך לא נתמודד עם ערכים סינגולרים שייתר קטנים ממה שחצנו לעצמנו להתמודד איתו.

## 2.9 סיכום מקרה ללא רעש

- בהינתן נתונים  $(x, f(x))$  אנו יודעים איך ללמידה פונקציה ליניארית  $\mathcal{H}_{\text{lin}} \in \mathcal{H}$ .
- **אנו זוקים לפחות  $d+1$  דגימות** על מנת לפתור את המערכת  $\mathbf{y} = X\mathbf{w}$ .
- נלמד באמצעות פתרת המשוואות הנורמליות, המנסה לכתוב את  $\mathbf{y}$  (במקרה הריאלייזבילי) או את ההטלה האורתוגונלית של  $\mathbf{y}$  על  $\text{Im}(x)$  (במקרה הלא ריאלייזבילי) כצירוף ליניארי של עמודות  $X$ .
- אם העמודות של  $X$  **בלתי תלויות ליניארית** - למשוואות הנורמליות פתרון יחיד.
- אם העמודות של  $X$  **תלויות ליניארית** - למשוואות הנורמליות יש **אינסוף פתרונות**.

- ווכל לפטור את המשוואות הנורמליות בצורה יציבה נומրית בעזרת **פירוק SVD**, ללא קשר למספר הפתרונות שיש למערכת (יחיד \ אין סופי).

- אם למערכת **פתרון יחיד**, פירוק SVD יאפשר למצוא אותו.  
אם למערכת **אין סוף פתרונות**, פירוק SVD יאפשר למצוא את הפתרון בעל הנורמה המינימלית.

## 2.10 המקרה הרועש

הנחה שהנחנו בתחלת בדרכ בה ערכיו  $y$  הם **בדיקות** ליניארים בדגימות של  $x$  אינה ריאלית. בפועל, לערכי  $y$  עלול להיות "רעש" כלשהו. על כן כעת נניח כי הדטא אוטוanno מקבלים הוא מהצורה:

$$\forall 1 \leq i \leq m : (x_i, f(x_i) + z_i)$$

עם  $(0, \sigma^2) \sim z_m$ ,  $z_1, \dots, z_m$ , ככלומר מתפלגים באופן אחיד ובלתי תלוי מהתפלגות בעלת תוחלת 0 ושונות  $\sigma^2$ .

### 2.10.1 איך נלמד כעת?

נמשיך להניח כי ישנו מספיק דגימות, ככלומר  $m \geq d+1$ . נניח כי קיים  $f \in \mathcal{H}_{\text{lin}}$  כלומר קיים  $w$  כלשהו עבורו  $y_i = x_i^\top w + z_i$ . נסמן את וקטור הרעש ע"י  $(z_1, \dots, z_m)^\top = z$  ונקבל בכתב מטריציוני כי  $z = Xw + u$ . כמעט בוודאות מתקיים כי  $u \notin \text{Im}(X)$  ועל כן למערכת  $w = u$  אין פתרונות.

### 2.10.2 הזרך למקרה ללא הרעש יפעל גם כאן!

ווכל להשתמש שוב בפונקציית ה- Square Loss כמוקדם, ונלמד ע"י מינימול הסיכון האמפירי:

$$w_S := \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \|y - Xw\|^2$$

כלומר, שוב נלמד את  $h_S \in \mathcal{H}_{\text{lin}}$  ע"י פתרת המשוואות הנורמליות.

דרך זו מאד הגיונית - יש לנו וקטור  $(X) \notin u$  אשר הרעש "הוציא" מתח המרחב  $(X) \text{Im}$ . Least Squares כפי שראינו מוקדם, פתרת המשוואות שקופה להטיל את  $u$  חזרה על  $(X) \text{Im}$ , ובכך למצוא קירוב

## 2.11 עקרון הנראות המירבית

אם נניח כי הרעש שלנו מתפלג גאוסיאני, ככלומר  $(x_i^\top w, \sigma^2) \sim \mathcal{N}(y_i, \sigma^2)$ , נקבל כי  $y_i \sim \mathcal{N}(x_i^\top w, \sigma^2)$  באופןבלתי תלוי. נניח כעת כי אנחנו יודעים את וקטור המשקלות  $w$ . נוכל לשאול כעת מה היא ההסתברות לקבל וקטור  $u$  כלשהו? התשובה לכך היא הצפיפות המשותפת של ההתפליגות הגaussיאניות:

$$\mathbb{P}(y | w) = \prod_{i=1}^m \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i^\top w - y_i)^2}{2\sigma^2}} \right)$$

זהה בעצם שאלת בהסתברות: אנו יודעים מה הוא  $w$ , מה היא ההסתברות לקבל את  $u$ ?

אנחנו מעוניינים בשאלת ההפוכה. דגモノ  $y$  כלשהו. מה הערך  $w$  בעל הסתברות הגבוהה ביותר? עקרון **הנראות המירבית** מציין לבחור את ה- $w$  שנותן את הסתברות הגבוהה ביותר לקל את הוקטור  $y$ . נכתוב את **פונקציית הנראות**, כתע עבור  $w$  כאשר  $y$  הוא הקבוע:

$$L(w | y) = \frac{1}{(\sqrt{2\pi\sigma^2})^m} \prod_{i=1}^m e^{-\frac{(x_i^\top w - y_i)^2}{2\sigma^2}}$$

אומד הנראות המירבית (MLE) עבור  $w$  הוא:

$$\begin{aligned} \hat{w} &:= \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmax}} L(w | y) = \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \log(L(w | y)) = \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \log \left( \frac{1}{(\sqrt{2\pi\sigma^2})^m} \prod_{i=1}^m e^{-\frac{(x_i^\top w - y_i)^2}{2\sigma^2}} \right) \\ &= \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \underbrace{\log \left( \frac{1}{(\sqrt{2\pi\sigma^2})^m} \right)}_{\text{קבוע}} + \sum_{i=1}^m -\frac{(x_i^\top w - y_i)^2}{2\sigma^2} = \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \sum_{i=1}^m -\frac{(x_i^\top w - y_i)^2}{2\sigma^2} \\ &= \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmax}} -\frac{1}{2\sigma^2} \cdot \sum_{i=1}^m (x_i^\top w - y_i)^2 = \boxed{\underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^m (x_i^\top w - y_i)^2} \end{aligned}$$

בזומה לאומד סכום הריבועים שקיבלנו ע"י הסיכון האמפירי.

## 2.12 התאמת פולינומית Polynomial Fitting

נרצה להרחיב את המודל של רגרסיה ליניארית. נניח שבמקום היחס הליניארי בין  $\mathcal{X}$  ל- $y$  כ- $w + \varepsilon$  עבור  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ , נציג פונקציות  $h_1, \dots, h_d : \mathbb{R}^d \rightarrow \mathbb{R}$  ונציג את הקשר:

$$y_i = \sum_{j=1}^d h_j(x)^\top w$$

כאשר הפונקציות  $h_1, \dots, h_d$  מוגדרות כפונקציות בסיס, ומאפשרות לנו להגדיר יחס שהוא ליניארי בפרמטרים של  $w$ , אך לא בהכרח ליניארים ב- $x$ . דוגמא ספציפית היא **התאמת פולינומית**. יהיו  $x_1, \dots, x_m \in \mathbb{R}$  ו-  $y_1, \dots, y_m \in \mathbb{R}$ . נרצה להגדיר יחס פולינומי בין  $\mathcal{X}$  ל- $y$ , מדרגה לכל היותר  $N$ . מחלוקת היפותיות תוגדר ע"י:

$$\mathcal{H}_{\text{poly}}^d = \left\{ x \mapsto p_w(x) = \sum_{i=1}^d w_i x^i \mid w \in \mathbb{R}^{d+1} \right\}$$

במקרה זה נגדיר את פונקציות הבסיס ע"י  $h_j(x) = x^j$  לכל  $j \in \{0, \dots, d\}$ .

$$\text{נסמן } h(x) := (h_0(x), \dots, h_d(x))^T = (1, x, x^2, \dots, x^d)^T$$

$$p_{\mathbf{w}}(x) = \sum_{j=0}^d h_j(x) \mathbf{w}_j = \langle h(x) \mid \mathbf{w} \rangle = h(x)^T \mathbf{w}$$

כאשר 1 הוא Intercept  $h_0(x) = x^0 = 1$

כמו מוקדם, נרצה למצוא את וקטור המקדמים  $\mathbf{w}$ . בהינתן דוגמאות  $S = \{(x_i, y_i)\}_{i=1}^m$ , נחפש אותם ל-

ונפתר את בעיית Least Squares הבאה:

$$\hat{\mathbf{w}} := \underset{\mathbf{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^m \left( h(x_i)^T \mathbf{w} - y_i \right)^2$$

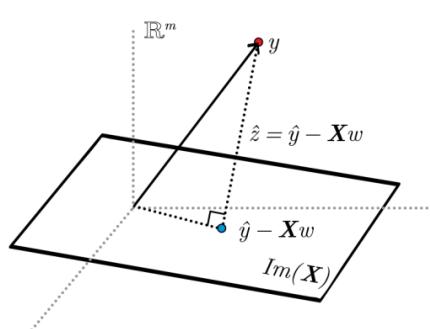
נבחן כי כעת המטריצה  $\mathbf{X}$  על הדאטה החדש היא מטריצת **ונדרומונד** מדרגה מלאה, עם:

$$\mathbf{X} = \begin{bmatrix} - & h(x_1) & - \\ - & h(x_2) & - \\ \vdots & & \\ - & h(x_m) & - \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^d \end{bmatrix}_{m \times (d+1)}$$

הפרמטר  $d$  הוא לבחירתנו. עבור  $d$  קטן יחסית, מחלוקת ההיפוטזות  $\mathcal{H}_{\text{poly}}^d$  תהיה קטנה. פולינום מדרגה גבוהה מ- $d$  לא יכול להיות מתואר אפילו ע"י ההסתrema הטובה ביותר ביוטר מ- $\mathcal{H}^d$ . באופן לא פורמלי, **הטיה** מתארת כמה "נכון" הפונקציה  $f$  יכולה להיות משוערכת ע"י מחלוקת ההיפוטזות שלנו. ככל שהחלוקת גדולה יותר  $\leftarrow$  **הטיה** נמוכה יותר.

כאשר מתעסקים בדגימות בעלות רעש, מתייחסים למשג השונות. בהינתן דוגמאות אקרניות, הפולינום המשוערך שלנו יהיה מושפע מהן. ככל שהרעש גבוה יותר, כך השונות של השערוך  $\hat{y}_S$  שלנו תהיה גדולה יותר.  
 ככל שחלוקת ההיפוטזות גדולה יותר  $\leftarrow$  השונות גדולה יותר. **למה?**  
 כי אלגוריתם הלמידה "רודף" אחר הנקודות, ככל שמעלת הפולינום גדולה יותר - הוא יכול לתפוס יותר נקודות.

**מבחינה גיאומטרית**, בהינתן  $z_i = \mathbf{x}_i^T \mathbf{w} + y_i$  לכל  $i$ , אנו מעריכים את  $\mathbf{w}$  ע"י הטלה של  $y$  על  $\text{Im}(\mathbf{X})$ .  
 ככל שהמייד של  $\text{Im}(\mathbf{X})$  גדול יותר ( $d$  גדול יותר), הטלה מנקה **פחות רעש**, ומתקבלים שונות גבוהות יותר.



### ה-Trade-off 2.12.1 Bias-Variance

- חלוקת היפוטזות קטנה לרוב תגרום **הטיה גדולה ושותנות קטנה**.
- חלוקת היפוטזות גדולה לרוב תגרום **הטיה קטנה ושותנות גדולה**.

## 3 הרצאה 3 - Classification

בבואה סיווג נתמקד במרחב דגימות  $\mathcal{X} = \mathbb{R}^d$  וסיווג ביןארי  $\{-1, 1\}$ . כמוון שניתן להקליל לסיווג בעל  $k$  איברים.

### 3.1 דיק

כיצד נמדד את רמת הדיק של מסווג  $h$  כלשהו? הדרך פשוטה ביותר - נמדד את כמה השגיאות:

$$L_s(h) = \sum_{i=1}^m \mathbb{1}_{y_i \neq h(x_i)} = |\{i \mid y_i \neq h(x_i)\}|$$

שיטה זאת נקראת Misclassification Error, והבעיה בה היא שהיא מתייחס לשגיאות בצורה זהה. בפועל שני סוגי השגיאות שהמסווג יכול לעשות הן בעלות השלכות ומהירות שונות.

#### 3.1.1 שני סוגי שגיאות

שני סוגי השגיאות אותם יכול מסווג  $h$  לעשות הן:

	$y = -1$	$y = 1$
$\hat{y} = -1$	✓	Type II Error
$\hat{y} = 1$	Type I Error	✓

1. **שגיאה מסוג ראשון - Type I Error**.

שגיאה זאת מתרחשת כאשר אנו מסוויגים חיובי ובפועל התשובה שלילית.

שגיאה זו נקראת גם False Positive.

2. **שגיאה מסוג שני = Type II Error**.

שגיאה זאת מתרחשת כאשר אנו מסוויגים שלילי והפועל התשובה חיובית.

שגיאה זו נקראת גם False Negative.

- ככל, תמיד נגדיר את הסימונים שלילי וחובי כך ששגיאה מסוג ראשון - סימון שלילי בתור חיובי, תוגדר כשגיאה החמורה יותר.

#### 3.1.2 סדר במושגים

נניח כי  $-1 = y$  משמעו שלילי, וכי  $1 = y$  משמעו חיובי.

- אם  $-1 = \hat{y}$  אז המקרה בו  $1 = \hat{y}$  הוא True Negative והמקרה בו  $-1 = \hat{y}$  הוא False Positive.
  - אם  $1 = \hat{y}$  אז המקרה בו  $1 = \hat{y}$  הוא True Positive והמקרה בו  $-1 = \hat{y}$  הוא False Negative.
- cut-off נסמן ב- P את המספר התוצאות החיוביות, וב-N את מספר התוצאות השליליות.
- True / False Negatives .True / False Positives .TP / FP הוא מספר ה-

- ה-Precision מוגדר ע"י  $\frac{TP}{TP+FP}$

- ה-FNR מוגדר ע"י  $\frac{FN}{P}$

- ה-Recall \ TPR מוגדר ע"י  $\frac{TP}{P}$

- ה-Error Rate מוגדר ע"י  $\frac{FP+FN}{P+N}$

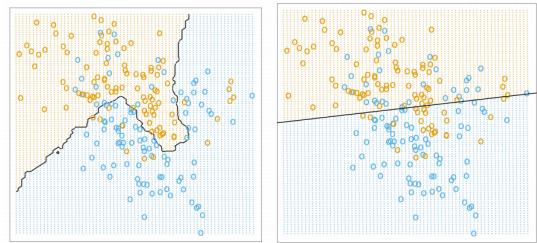
- ה-TNR \ Specificity מוגדר ע"י  $\frac{TN}{N}$

- ה-Accuracy מוגדר ע"י  $\frac{TP+TN}{P+N}$

- ה-FPR מוגדר ע"י  $\frac{FP}{N}$

### 3.1.3 גבול החלטתית Decision Boundary

על מנת להבין טוב יותר מסוג  $h$  כשלחו, דרך שימושית היא להסתכל על החלוקה בין  $\{x \mid h(x) = 1\}$  לבין  $\{x \mid h(x) = -1\}$ .



## 3.2 ניתוח מסוג חדש

בהתנחת מסוג  $h$  כשלחו, נרצה לשאול את השאלות הבאות:

- האם המודל החדש בר פירוש?
- מהי מחלוקת היפותזות  $\mathcal{H}$ ? איך נראה גבול החלוקה?
- מה הוא העיקרונו לפיו אנו בוחרים  $h_S \in \mathcal{H}$  ביןtan DATA  $S$ ?
- האם זהו מודל יחיד או משפחה של מודלים?
- איך נמשח עיקרונו זה בצורה חשובה?
- לאחר האימון, כיצד נחסן את המודל המאומן  $\mathcal{H}$ ?
- מתי נשתמש בו?
- בהינתן מודל מאומן  $h_S \in \mathcal{H}$ , כיצד נחשב חיזוי  $y$  עבור דגימה חדשה  $x$ ?

## 3.3 מסוג Half-Space

### 3.3.1 מחלוקת היפותזות

הfonקצייה  $\langle w | x \rangle + b$  ננתנת ערכי  $1 \pm$  נקודות באחד מצדי ה

על מישור.
 במקרה שלנו כי  $0 = b$ . מחלוקת העל-מישורים היפותזית מוגדרת ע"י  $\mathcal{H}_{\text{half}} = \{h_w \mid w \in \mathbb{R}^d\}$ .

### 3.3.2 עיקרונו הבחירה

נשתמש ב- Error Misclassification. עבור דגימה שחוسبה בצורה נכונה, קיבל כי  $0 > \langle w | x \rangle + b$ . שcn על  $\langle w | x \rangle + b$  להיות בעלי סימן זהה. אם הסימנים שונים, קיבל תוצאה שלילית. מספר הסיווגים (הסיכון האמפירי) שկול ל-  $L_S(h) = \#\{i \mid y_i \neq \langle w | x \rangle + b\}$ .

נניח תחילה כי הדatta שלנו **ניתן לחלוקת ליניארית**, כלומר ניתן להעביר קו כך שכל השיליליים יהיו מצד אחד, וכל החיוביים מהצד الآخر. במקרה זה קיים  $w$  המקיים  $L_S(h_w) = 0$ . מכך שזאת שcola לעקרון ERM - מזעור השגיאה שבחרנו על הדatta הנתון.

### 3.3.3 מימוש חישובי

$w$  המקיימים  $0 = L_S(h_w) = \sum_i y_i \langle x_i | w \rangle$  לכל  $i \in [m]$ .  
 אנו מניחים שאחד כזה קיים, ואם נורמלו ע"י  $\min_i y_i \langle x_i | w \rangle$  נשאף לחפש את  $w_0$  המקיים:  
 $\forall i \in [m] : y_i \langle x_i | w_0 \rangle \geq 1$

כיצד נוכל לחשב זאת ביעילות?

### 3.3.4 אופטימיזציה קמורה

כפי שראינו בתרגול, הבעה הנ"ל היא בעית אופטימיזציה קמורה.  
 היתרונות בעיות אלה הוא **ש殆מיד קיים להם מינימום**. מינימום זה ייחיד, ונitinן לחسابו **ביעילות**.  
 עבור מסוג העל-מישור, מימוש מצור הסיכון האמפירי מתבצע ע"י פתרת הבעה הבאה:

minimize	0
subject to	$\forall i \in [m] : y_i \langle x_i   w \rangle \geq 1$

זהה בעיה **פיזיבילית**, בה אנו רוצים למצער 0.  
 ככלומר, אנו לא מנסים למצער דבר, אלא רק למצוא וקטור העונה על כל התנאים.

- במקרה שהدادטא שלנו **ניתן חלוקה ליניארית**, ניתן לפתור בעיה זו ע"י אלגורתמי תכנון ליניארי \ Perceptron Algorithm.
- במקרה שהدادטא שלנו **לא ניתן חלוקה ליניארית**, פתרת הבעיה הזאת היא חישובית קשה ולא אפשרית.

### 3.3.5 לסייע

- **מחלקת היפוטזות:**  $\mathcal{H}_{\text{half}} = \{h_w \mid w \in \mathbb{R}^d\}$ .
- **עקרון למידה:** על-מישור מפריד עם מצור סיכון אמפירי ע"י חישוב מס' סיוגים שגויים.
- **מימוש חישובי:** תכנון ליניארי, Perceptron Algorithm.
- **איך לאחסן את המודל המאומן:** ע"י שבירת הווקטור  $w$ .
- **מתי להשתמש:** אף פעם.

## 3.4 מסובג SVM - Support Vector Machines

### 3.4.1 מחלקת היפוטזות

אנו נשאים בחלוקת ה

על-משורים
 $\mathcal{H}_{\text{half}} = \{h_w \mid w \in \mathbb{R}^d\}$

### 3.4.2 עקרון למידה

נרצה לשאול את עצמנו - איזה על-משור מחלק הצורה הטובה ביותר?

לשם כך נגדיר מס' מושגים:

בהתנן על-משור המוגדר ע"י  $L = \{v \mid \langle v | w \rangle = 0\}$ , וקטור  $x$ , נגדיר את המרחק בין  $x$  ל- $L$  על ידי:

$$d(x, L) = \min \{\|x - v\| \mid v \in L\}$$

**טענה:** אם  $1 = \|w\|$  אז מתקיים  $|\langle w | x \rangle| \leq d(x, L)$ .

נזכיר כי הגדרנו על משור ע"י וקטור  $w$  המקיים  $0 > \langle w | x_i \rangle$  לכל  $i$ .  
 ה- Margin של על-משור מפריד מוגדר להיות המרחק מהדגם הקרוב ביותר אליו:  
 $\min_i |\langle w | x_i \rangle|$ .  
 הדגימות הקרובות ביותר יקרו **Support Vectors**.

- במקרה שהדעתה שלנו **ניתן לחלוקת ליניארית**,  
 נפעיל לפיה Hard-SVM ונמצא את ה
על-משור המפריד בעל ה- Margin הגדול ביותר:

$$\underset{w: \|w\|=1}{\operatorname{argmax}} \left\{ \min_{i \in [m]} \{ |\langle w | x_i \rangle| \mid \text{s.t. } y_i \langle w | x_i \rangle > 0 \} \right\} \stackrel{\text{תרגיל}}{\iff} \underset{w}{\operatorname{argmin}} \left\{ \|w\|^2 \mid \forall i : y_i \langle w | x_i \rangle \geq 1 \right\}$$

זהה בעיית אופטימיזציה קמורה בעלת **תבנית ריבועית ואילוצים ליניארים** ולכן נקראת גם QP Solver.  
 ניתן לפתור אותה באמצעות QP.

- במקרה שהדעתה שלנו **לא ניתן לחלוקת ליניארית**, ננסה את האילוצים שלנו לצורה הבאה:

$$\begin{aligned} \underset{w, \xi}{\operatorname{argmin}} \quad & \left\{ \lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i \right\} \\ \text{subject to} \quad & \forall i : y_i \langle x_i | w \rangle \geq 1 - \xi_i \quad \wedge \quad \xi_i \geq 0 \end{aligned}$$

זהה גם בעית QP, כאשר המשתנים  $\xi$  מודדים את ההפרות על התנאי  $y_i \langle x_i | w \rangle \geq 1$ .  
 אם לפטורו האופטימלי קיים  $0 > \xi_i$ , אז הדגם ה-  $i$  מפרה את ה- Margin ב-  $\xi_i$ .  
 הפרמטר  $\lambda$  נקרא **פרמטר רגולרייזציה**.

ככל ש- $\lambda$  גדלה, הפרמטר  $\|w\|^2$  יהיה גדול מאוד ומאנץ המזעור יעסוק בו. מתחששות הפרות מאוד גדולות.  
 מצד שני, כאשר  $\lambda$  קטנה, מאנצ המזעור יעסק בהפרות, ולא ב-  $\|w\|^2$  שקובעת את ה- Margin.

### 3.4.3 לסיכום - Soft SMV

- **מחלקת היפוטזות:**  $\mathcal{H}_{\text{half}} = \{h_w \mid w \in \mathbb{R}^d\}$
- **עקרון למידה:** על-מישור מפריד עם Margin מקסימלי.
- **IMPLEMENTATION:** Quadratic Programming
- **חישוב חיזוי:**  $y_i = \langle x_i | w \rangle$
- **איך לאחסן את המודל המאמן:** ע"י שמירת הווקטור  $w$  וה- $b$  Intercept.
- **משפחה:** כנ. כל מודל נקבע ע"י פרמטר הרוגולרייזציה  $\lambda \in [0, \infty]$ .
- **מתי להשתמש:** (i) As a simple baseline (ii) After embedding data in high-dimensional space (Kernerlization)

## 3.5 מסוג רgresיה לוגיסטי

### 3.5.1 הקדמה

נזכיר ברגרסיה ליניארית בעלת רעש נאותני. הנחנו כי  $y_i = \langle x_i | w \rangle + \epsilon_i$  ולכן  $y_i \sim \mathcal{N}(\langle x_i | w \rangle, \sigma^2)$ . על כן האנלוגיה תהיה להניח כי כל  $y_i \sim \text{Ber}(p_i)$  עבור  $p_i$  ליניארי ב- $x_i$ .

נבחר פונקציית לינק  $\phi : \mathbb{R} \rightarrow (0, 1)$ : מונוטונית עולה, חד"ע ועל. נניח כי  $p_i = \phi(\langle x_i | w \rangle)$  עבור וקטור משקלות  $w \in \mathbb{R}^{d+1}$  קלשו. בדומה לרגרסיה ליניארית אנו נכלל את ה- Intercept כך ש-  $x = (1, x_1, \dots, x_d)^\top$  ו-  $w = (w_0, w_1, \dots, w_d)^\top$ .  $\text{Ber}(p_i) = \text{Ber}(\phi(\langle x_i | w \rangle))$  ובכך בעצם נניח מודל הסתברותי המקיים { $y_i$ } משתנים מקרים בלתי תלויים המתפלגים

### 3.5.2 פונקציית לינק

למרות שישנו מס' פונקציות העונות על הדרישות, אנחנו משתמש בפונקציה הלוגיסטי  $\pi(x) = \frac{e^x}{1+e^x}$ . זהה פונקציה חלקה, מונוטונית עולה והפיכה.

### 3.5.3 מחלקה היפוטזות

מחלקה היפוטזות שלנו תוגדר ע"י  $\mathcal{H}_{\text{logistic}}^d = \{x \mapsto \pi(\langle x | w \rangle)\}$

### 3.5.4 עקרון הבחירה

על מנת לבחור  $w \in \mathcal{H}_{\text{logistic}}^d$  נשימוש בעקרון ה- Maximum Likelihood Principle: הצפיפות המשותפת של הווקטור  $y = (y_1, \dots, y_m)^\top$  עבור וקטור משקלות  $w \in \mathbb{R}^{d+1}$  היא:

$$\mathbb{P}(Y = y \mid w) = \prod_{i=1}^m p_i(w)^{y_i} \cdot (1 - p_i(w))^{1-y_i} \quad \left| \quad p_i = \pi(\langle x_i | w \rangle) = \frac{\exp(\langle x_i | w \rangle)}{1 + \exp(\langle x_i | w \rangle)} \right.$$

הנראות היא בעצם הצפיפות המשותפת, כפונקציה של  $w$ , עבור ערכי  $\{(x_i, y_i)\}_{i=1}^m$  קבועים. במקרה שלנו:

$$L(w | y) = \prod_{i=1}^m p_i(w)^{y_i} \cdot (1 - p_i(w))^{1-y_i}$$

$$\underset{\mathbf{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} L(\mathbf{w} | \mathbf{y}) = \underset{\mathbf{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \log(L(\mathbf{w} | \mathbf{y}))$$

מכיוון ש-  $\log$  הינה פונקציה חיובית מונוטונית עולה ממש. נסמן  $\ell$  ונגידר ( $\ell := \log(L(\mathbf{w} | \mathbf{y}))$ ) ונקבל:

$$\begin{aligned} \ell(\mathbf{w} | \mathbf{y}) &= \log \left( \prod_{i=1}^m p_i(\mathbf{w})^{y_i} \cdot (1 - p_i(\mathbf{w}))^{1-y_i} \right) = \sum_{i=1}^m \log(p_i(\mathbf{w})^{y_i} \cdot (1 - p_i(\mathbf{w}))^{1-y_i}) \\ &= \sum_{i=1}^m \log(p_i(\mathbf{w})^{y_i}) + \log((1 - p_i(\mathbf{w}))^{1-y_i}) \\ &= \sum_{i=1}^m y_i \cdot \log(p_i(\mathbf{w})) + (1 - y_i) \cdot \log(1 - p_i(\mathbf{w})) \\ (\text{חוקי לוגים}) \quad &= \sum_{i=1}^m y_i \cdot \log\left(\frac{e^{\beta_i}}{1 + e^{\beta_i}}\right) + (1 - y_i) \cdot \log\left(1 - \frac{e^{\beta_i}}{1 + e^{\beta_i}}\right) \\ &= \sum_{i=1}^m y_i \cdot \log\left(\frac{e^{\beta_i}}{1 + e^{\beta_i}}\right) + (1 - y_i) \cdot \log\left(\frac{1}{1 + e^{\beta_i}}\right) \\ &= \sum_{i=1}^m y_i \cdot \log(e^{\beta_i}) - y_i \cdot \log(1 + e^{\beta_i}) + (1 - y_i) \cdot (0 - \log(1 + e^{\beta_i})) \\ &= \sum_{i=1}^m y_i \cdot \log(e^{\beta_i}) - \log(1 + e^{\beta_i}) = \boxed{\sum_{i=1}^m y_i \cdot \langle \mathbf{x}_i | \mathbf{w} \rangle - \log(1 + e^{\langle \mathbf{x}_i | \mathbf{w} \rangle})} \end{aligned}$$

קיבלו כי לפי עקרון הנראות המקסימלית עליינו לבחור  $w \in \mathbb{R}^{d+1}$  (או לחילופין וקטור  $h \in \mathcal{H}_{\text{logistic}}^d$ ) ע"י מציאת:

$$\hat{\mathbf{w}} := \underset{\mathbf{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \sum_{i=1}^m y_i \cdot \langle \mathbf{x}_i | \mathbf{w} \rangle - \log(1 + e^{\langle \mathbf{x}_i | \mathbf{w} \rangle})$$

זהו פונקציה **קעורה** ב-  $w$ , ועל כן הבעה מוגדרת כבעית אופטימיזה קמורה, ונitinן לפתור אותה בעזרת אלגוריתם מתאים.

- יש אלגוריתם מיוחד המועד לפתור את הפונקציה הספציפית הזאת. החבילת glmnet מומלצת.

### 3.5.5 פרשנות

בהתנן מסובג מאמון ומוכן, לעיתים רבות נשאל:

- אם ניתן להסביר מדוע חיזוי שגא?
- אלו פיצרים היום החשובים ביותר עבור הסיווג?

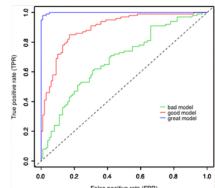
הרגression הלוגיסטי היא מסובג בר פירוש.

nocל להסתכל על וקטור המשקלות  $w$  ולראות אילו פיצרים קיבלו משקל גבוה, ואילו קיבלנו משקל נמוך.

## 3.5.6 חיזוי

חשיבות לשים לב כי הגדרנו את מחלוקת ההיפותזות על ידי  $\mathcal{H}_{\text{logistic}}^d = \{x \mapsto \pi(\langle x | w \rangle)\}$ , ועל כן היא מכילה פונקציות מהצורה  $[0, 1] \rightarrow \mathbb{R}^d$  ולא  $\{0, 1\} \rightarrow \mathbb{R}$ . משמעות הדבר היא שמודל רגסיבי לוגיסטי מאומן נותן לנו הסברות סיווג משוערכות, כלומר בהינתן דוגמה חדשה  $\hat{x}$  אנו מקבלים כי  $\hat{y} \in [0, 1]$  הוא שיעור ההסתברות ש- $x$ .

$$\hat{y} := \begin{cases} 1 & h(x) > \alpha \\ 0 & h(x) \leq \alpha \end{cases}$$



## 3.5.7 בחרית ה-cut-off

נסמן את  $1 = \hat{y}$  בתור **שלילי** ואת  $0 = \hat{y}$  בתור **חיובי**.  
כאשר  $1 \sim \alpha$ , נסוג את רוב הדוגמאות כ**שליליות**, ויהיו מעט **False Positives** ו-**True Positives**.  
כאשר  $0 \sim \alpha$ , נסוג את רוב הדוגמאות כ**חיוביות**, ויהיו הרבה **False Positives** ו-**True Positives**.  
קיים trade-off בין היחס של  $\text{True Positives}$  ו- $\text{False Positives}$  בבחירה של ה-cut-off. דרך ויזואלית להעריך את כל ה-trade-off הוא עקומת ROC, המשמשת לבחירת  $\alpha$  לפי ה-trade-off הרצוי.

## 3.5.8 לסייע

- **מחלקת היפותזות:**  $\mathcal{H}_{\text{logistic}}^d = \{x \mapsto \pi(\langle x | w \rangle)\}$
- **עקרון למידה:** נראות מקסימלית.
- **מימוש חישובי:** אופטימיזציה קמורה למינימום הנראות. קיימים אלגוריתם מיוחד לפונקציה זו.
- **חישוב חיזוי:** ע"י עברו  $0 < \hat{y} < 1 = \mathbf{1}_{h(x) > \alpha}$  משתנה cut-off.
- **איך לאחסן את המודל המאומן:** ע"י שימירת הוקטור המשקלות  $w$ .
- **בר פירוש:** כן, ע"י המשקלות המותאמת לכל פיצר.
- **ספק הסתברות לשיזוד:** כן.
- **משפחה:** רק כאשר אנו מוסיפים או מורידים פיצרים.
- **מתי להשתמש: תמיד.**

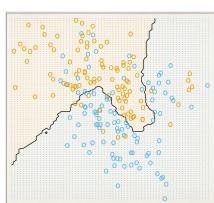
### 3.6 מסווג $k$ - שכנים קרובים

מסווג זה שונה מאשר המסווגים עליהם למדנו. אין לו מחלוקת היפוטזות, ואין לו שלב אימון. הפרמטר היחיד הוא  $k$ , עבור  $m \leq k \leq 1$  (בעדיפות איזוגי).

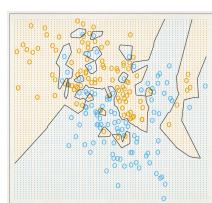
#### 3.6.1 חיזוי

בהתנן דגימה  $\mathbb{R}^d \in x$ , נמצא את  $k$  השכנים הקרובים ביותר (בעזרה פונק' מרחק  $\rho$  כלשהו) ל- $x$ , ונורוך "הצבעה". הערך שיחזר על עצמו יותר פעמים יהיה התוצאה אותה נחיזיר.

$k = 15$



$k = 1$



#### 3.6.2 בחירת $k$

- עבור  $k$  קטן, קיבל כי ההטיה נמוכה והשונות גבוהה.
- עבור  $k$  גדול, קיבל כי ההטיה גבוהה והשונות נמוכה.

#### 3.6.3 מימוש חישובי

- **ברוט-פורס:** נחשב מרחקים ע"י  $(x, x_i) \rho$  ונמיין עבור כל דגימה חדשה  $x$ . נאלץ לשמור את כל הדאטה  $S$ .
- **מבנה נתונים מאוחתל מראש:** במקומות שלב האימון, נוכל להכין מבנה נתונים עבור הדאטה  $S$  שלנו, על מנת לאפשר חיפוש שכנים מהיר ויעיל יותר. בכך נוכל להיפטר מ- $S$  ולשמור רק את מבנה הנתונים החדש.

#### 3.6.4 לסתוקים

- **מחלקת היפוטזות:** אין.
- **עקרון למידה:** אין.
- **מימוש חישובי:** ברוט-פורס או מבנה נתונים ייעודי.
- **איך לאחסן את המודל המאומן:** אין מודל. שמירת הדאטה  $S$  או מבנה הנתונים.
- **בר פירוש:** לא.
- **ספק הסתברות לשיזוק:** לא, אלא אם כן משתמשים במספר רב של שכנים.
- **משפחה:** כן, לפחות  $k$  מספר השכנים.
- **מתי להשתמש:** תמיד לנסות כאשר אפשר ממש.

## 4 הרצאה 4 - מודל PAC

### 4.1 חזרה - בעית סיווג

התחום שלנו הוא  $\mathcal{X}$ , קבוצת האובייקטים אותם נרצה לסווג. הטווח שלנו הוא  $\mathcal{Y}$ , קבוצת כל התוצאות האפשריים. לדוגמה,  $\{0, 1\}$  עבור הצלחה או כישלון. כלל החלטה שלנו הוא  $\mathcal{Y} \rightarrow \mathcal{X}$ :  $h$ , המשתמש לתיאוג דגימות עתידיות. נקרה גם חזה, היפוטזה, מסועג.

#### 4.1.1 אינפוט אוטופוט של לומד

- **אינפוט:** דוגימות DATA,  $.S = (x_1, y_1), \dots, (x_m, y_m) \in (\mathcal{X} \times \mathcal{Y})^m$

- **אוטופוט:** חיזוי,  $h : \mathcal{X} \rightarrow \mathcal{Y}$

מכאן שלומד הוא מיפוי  $\mathcal{A} : S \mapsto h : \mathcal{X} \rightarrow \mathcal{Y}$ , המפה  $h$  מAPPING הינה אינטואיטיבית,  $h$  צריכה לצדוק על דוגימות עתידיות. על מנת לכמת זאת, נזקק לfrmול מתמטי לגבי אופן יצרת הדוגימות העתידיות.

### 4.2 מסגרת תיאורטית עבור למידה

#### 4.2.1 למה צריך מסגרת?

נרצה לענות על השאלות הבאות:

- מה ניתן ללמוד ומה לא?

- כאשר ניתן ללמוד, כמה דוגימות נדרש?

- כאשר ניתן ללמוד, איך ניתן ללמוד?

במסגרת הרצאה זו נלמד תיאורית למידה שספקת תשובה מלאה עבור Batch Supervised Learning.

#### 4.2.2 מודל ייצור הדטא

אנו נדרש להנחות מתמטיות פורמליות על איך נציג דוגימות הדוגימות.

נניח כי קיימת התפלגות  $\mathcal{D}$  מעל הדוגימות  $\mathcal{X}$ .

אנו עוסק במקרה של **סיווג**, אך ניתן להכליל הכל לביעות רגסיביה. אנו נניח כי:

- כל דוגמה  $x_i$  נדגמה מההתפלגות  $\mathcal{D}$ .

- כל הדוגימות נדגו באופן בלתי תלוי אחת בשנייה.

בנוסף נניח כי בהכרח קיים  $f$  כלשהו כך ש-  $f(x) = y$ .

**סט הדוגימות** שלנו מכיל זוגות  $(x_i, f(x_i))_{i=1 \dots m}$  כאשר כל  $x_i \sim \mathcal{D}$ .

כל דוגמה עתידית  $x \in \mathcal{X}$  תדגם מההתפלגות  $\mathcal{D}$  גם היא, באופן בלתי תלוי בכל שאר הדוגימות.

### 4.2.3 הערכת ביצועים

נדיר את ממד השגיאה באופן הבא:

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x \sim \mathcal{D}}(h(x) \neq f(x)) \stackrel{\text{def}}{=} \mathcal{D}(\{x \in \mathcal{X} \mid h(x) \neq f(x)\})$$

נקרא גם **הסיכון Risk** או **שגיאת הסיווגים השגוים Misclassification Error**.

מדד השגיאה הוא בעצם ההסתברות לשגיאות על דגימה עתידית כלשהי. נשים לב כי  $f, \mathcal{D}$  אינם ידועים. חשוב לציין כי לרוב נרצה להתייחס לשגיאות בצורה יותר ביקורתית, להבדיל בין סוגים השגיאות ולתת משקל שונה לכל אחת.

## 4.3 מושגים והגדרות מרכזיים

### 4.3.1 למידות PAC

נאמר כי מחלוקת היפוטזות  $\mathcal{H}$  היא **למידה PAC** אם קיימת פונקציה  $\mathbb{N} \rightarrow (\tilde{m}_{\mathcal{H}}, 1)^2$  ואלגוריתם למידה  $\mathcal{A}$  כך שלכל  $(\delta, \varepsilon) \in (0, 1)^2$ , לכל התפלגות  $\mathcal{D}$  מעל  $\mathcal{X}$  ולכל פונקציית תיוג  $\{\pm 1\} \rightarrow \mathcal{X}$ :  $f : \mathcal{X} \rightarrow \mathcal{D}$  המקיימת  $L_{\mathcal{D},f}(h^*) = 0$  על  $m_{\mathcal{H}}(\varepsilon, \delta) \geq \tilde{m}_{\mathcal{H}}(\varepsilon, \delta)$  דגימות המתפלגות באופן בלתי תלוי מ- $\mathcal{D}$  עבורו  $h \in \mathcal{H}$  כלשהו, כאשר מרכיבים את האלגוריתם  $\mathcal{A}$  על  $m_{\mathcal{H}}(\varepsilon, \delta) + 1$  מתיויגות ע"י הפונקציה  $f$ , האלגוריתם מחזיר היפוטזה  $h_S = \mathcal{A}(S)$  כך שבהתברור של לפחות  $\delta - 1$  (מעל הסט  $S$ ) מתקיים כי  $L_{\mathcal{D},f}(h_S) \leq \varepsilon$ .

### 4.3.2 סיבוכיות המודל

עבור מחלוקת היפוטזיות למידה PAC, נגידר את **סיבוכיות המודל** של  $\mathcal{H}$  עבור  $\delta, \varepsilon$  ספציפיים כמספר המינימלי של דגימות  $(\delta, \varepsilon)_{\mathcal{H}}$  הנחוצות בהגדלת למידות PAC ביחס ל-  $\delta, \varepsilon$ .  
**פונקציית סיבוכיות המודל** של מחלוקת היפוטזות  $\mathcal{H}$  מסומנת ע"י  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$ .

### 4.3.3 מימד VC

תהיה  $\mathcal{X} \subseteq \mathcal{H}$  מחלוקת היפוטזות. עבור תת-קובוצה  $\mathcal{X} \subseteq C \subseteq \mathcal{H}_C$  להיות ההגבלה של  $\mathcal{H}$  לקובוצה  $C$ , כלומר הקבוצה  $\mathcal{H}_C = \{h_C \mid h \in \mathcal{H}\}$  כאשר עבור  $y \in \mathcal{Y} : h : \mathcal{X} \rightarrow \mathcal{Y}$  הפונקציה  $y \in \mathcal{Y}$  מקיימת כי  $h(x) = y$  לכל  $x \in C$ . נגידר את **מימד VC** של  $\mathcal{H}$  ע"י:

$$\text{VCdim}(\mathcal{H}) = \max \{|C| \mid C \subseteq \mathcal{X} \text{ and } |\mathcal{H}_C| = 2^{|C|}\}$$

ונבחן כי  $\text{VCdim}(\mathcal{H}) \leq \infty$ .

## 4.4 המשפט היסודי של הלמידה הסטטיסטי

אם נאץ את למידות PAC כתפיסה שלנו ללמידה, נקבל **אפיקון מלא** ללמידה:

- מה ניתן ללמידה
- כמה דגימותanno זוקקים
- כיצד ללמידה כאשר הלמידה אפשרית

תוצאה זו נקראת לעיתים המשפט היסודי של הלמידה הסטטיסטי:

- מחלוקת היפותזות  $\mathcal{H}$  היא **למידה PAC** אם ורק אם  $\epsilon < \text{VCdim}(\mathcal{H})$ .
- **סיבוכיות המודול** של מחלוקת היפותזות בעלי ממד VC סופי היא בערך:
$$m_{\mathcal{H}}(\epsilon, \delta) \sim \frac{\text{VCdim}(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\epsilon}$$
**אי אפשר בפחות, לא צריך יותר.**
- כלל ה- ERM תמיד מושג את המינימום. כלומר, כאשר הלמידה היא אפשרית, כלל ה- ERM לומד עם מס' מינימלי של דוגמאות.

## 4.5 נתחליל לשחק

נרצה לדמיין משחק כלשהו ביןנו - הלומדים - לבין הטבע, עם **תשולם אקראי**. אנחנו מחליטים החלטה (בוחרים אלגוריתם  $\mathcal{A}$ ) בתנאים עמודים: לא יודעים את  $f$ ,  $\mathcal{D}, f : \mathcal{X} \rightarrow \mathcal{Y}$ .

## 4.6 גרסא ראשונה | Learning Game 1.0

- מספר הדוגמאות  $m$  הינו קבוע.
- אנו מתחילהם, ובוחרים אלגוריתם למידה  $\mathcal{A} : (\mathcal{X}, \mathcal{Y})^m \rightarrow \mathcal{A}$ .
- הטבע **מכיר את הבחירה שלנו**, ובוחר התפלגות  $\mathcal{D}$  מעל  $\mathcal{X}$ , ופונקציית תיוג  $\gamma : \mathcal{Y} \rightarrow \mathbb{R}$ .
- נדגם סט  $S$  של  $m$  דוגמאות באופן תלוי מ-  $\mathcal{D}$ , ומתיויג ע"י  $f$ .
- האלגוריתם  $\mathcal{A}$  מקבל סט  $S$  זה ומחזיר כלל החלטה  $h_S = \mathcal{A}(S)$ .
- **התשלום האקראי** מוגדר להיות  $(h_S)_{\mathcal{D}, f}$ . ככלר החלק היחסי של הסיווגים הטעילים שככל ההחלטה יעשה על דוגמאות שידגמו באופן תלוי מ-  $\mathcal{D}$  ויתוויג ע"י  $f$ . התשלום הוא **אקראי** מכיוון שהסט  $S$  הינו אקראי ובעקבות כך גם  $h_S$ .

### 4.6.1 שאלת ראשונה

האם קיים אלגוריתם כלשהו שיבטיח חסם עליון  $1 - \epsilon \leq 0$  על השגיאה בהסתברות של **1**?

התשובה לכך היא **לא**. בהינתן  $1 - \epsilon \leq 0$  הלומד לא יכול לקוטר להפיק כלל החלטה  $h$  עם  $\epsilon \leq L_{\mathcal{D}, f}(h)$  ללא כל קשר לאיך שייפעל הטבע. **למה?** ישנו תמיד סיכוי קטן לקבל סט דוגמאות פתולוגית למחרי, שלא מייצג כלל את  $\mathcal{D}$ . כלל ההחלטה  $h_S$  יכול לטעות על רובו של  $\mathcal{A}$ , ואם הסט  $S$  ממש רע,  $h_S$  יכול להיות בעל שגיאה גבוהה בהרבה מכל  $\epsilon$ .

**דוגמא.** ניקח  $\{x_1, x_2\} = \mathcal{X}$ . נקבע  $0 > \gamma$ . אלגוריתם הלמידה שלנו  $A$  חייב לקבוע מה לחזות עבור נקודה שלא נראית בסט הדגימות  $S$ . בה"כ נבחר לחזות + על נקודה שלא נראית בסט האימון. הטעו בוחר לשחק עם התפלגות  $\mathcal{D}$  כך  $\text{ש-}\gamma = \mathcal{D}(\{x_1\}) = 1 - \gamma$ , ועם פונקציית תיוג  $f(x_1) = f(x_2) = -\gamma$ . בנסיבות של  $\gamma^m$ , סט האימון  $S$  יכול  $x_2$  בלבד. כיצד יסוג כלל הלמידה שלנו  $h_S$  את  $x_1$  במקרה זה? מכיוון ש- $x_1$  לא בסט האימון, אנו נחזה  $+ = h_S(x_1)$ . כתוצאה לכך, קיבל כי בדגם נקודה כלשהי מ- $\mathcal{D}$ , הכלל שלנו יטעה בהסתברות של  $\gamma - 1$ . מכאן ש- $\gamma - 1 > \delta$ , ונקבל סתירה.

#### 4.6.2 אלגוריתם למידה Probably Correct

נרצה לעדן במעט את הדרישה שלנו. נסכים לאלו' הלמידה להיכשל בהסתברות  $(0, 1) \in \delta$ , המוגדרת בתנאי המשחק.

הגדרה

כאשר לאלגוריתם למידה  $A$  יש שגיאה גדולה בהסתברות של לכל היותר  $\delta$ , עברו  $(0, 1) \in \delta$ , נאמר כי אלגוריתם הלמידה  $A$  **בכל הנראה Probably צודק**, עם **בטיחות confidence**.

בעצם נרצה לאלגוריתם שלנו להיכשל לחלוטין, בהסתברות של לכל היותר  $\delta$  קבוצה מראש. נוכל רק ל��ות לחסום עליון כל השגיאה (שייעמוד בכל דרך פעולה של הטעו) בהסתברות של לפחות  $\delta - 1$ .

#### 4.6.3 שאלה 2

במקרה של דגימה לא פתולוגית (המארע עם הסתברות של לפחות  $\delta - 1$ ), האם נוכל לקבל **דיקוק מושלם**? כמובן האם קיים אלג' למידה  $A$  בעל **שגיאה אפסית** (בהסתברות של לפחות  $\delta - 1$ ), ללא כל קשר לאיך שייפעל הטעו?

גם כאן התשובה היא **לא**. בהינתן  $0 > \delta$  אלג' הלמידה לא יכול ל��ות להפיק כלל  $h_S$  כך שהסתברות של לפחות  $\delta - 1$  ישג  $0 = L_{\mathcal{D},f}(h)$  ללא כל קשר לאיך שייפעל הטעו. **למה?** קיום  $0 = L_{\mathcal{D},f}(h)$  משמעו כי  $1 = (\{h_S(x) = f(x)\})$ , כלומר  $h_S$  עשויים לא טועה. אם הטעו יחליט לבחור  $\mathcal{D}$  שיתן משקל נמוך מאוד ל-  $\mathcal{X} \in x$  כלשהו בהסתברות גבוהה, אז  $x$  לא כולל ב-  $S$ , ול-  $h_S$  לא יהיה כל ידע כיצד לסוג אותו.

**דוגמא.** ניקח  $\{x_1, x_2\} = \mathcal{X}$ . נקבע  $0 > \delta$ . לאלגוריתם הלמידה שלנו  $A$  יהיה מותר להיכשל בהסתברות  $\delta$ . כמו קודם, נקבע  $0 > \gamma$ . אלגוריתם הלמידה שלנו  $A$  חייב לקבוע מה לחזות עבור נקודה שלא נראית בסט הדגימות  $S$ . בה"כ נבחר לחזות + על נקודה שלא נראית בסט האימון. הטעו בוחר לשחק עם התפלגות  $\mathcal{D}$  כך  $\text{ש-}\gamma = \mathcal{D}(\{x_1\}) = 1 - \gamma$ , ועם פונקציית תיוג  $f(x_1) = f(x_2) = -\gamma$ . בנסיבות של  $\gamma^m$ , סט האימון  $S$  לא יכול את  $x_2$ . כיצד יסוג כלל הלמידה שלנו  $h_S$  את  $x_2$  במקרה זה? מכיוון ש- $x_2$  לא בסט האימון, אנו נחזה  $+ = h_S(x_2)$ . כתוצאה לכך, קיבל כי בדגם נקודה כלשהי מ- $\mathcal{D}$ , הכלל שלנו יטעה בהסתברות של  $\gamma$ . מכאן ש- $0 > \gamma = L_{\mathcal{D},f}(h_S)$ , ונקבל סתירה.

#### 4.6.4 אלגוריתם למידה Approximately Correct

נרצה לעדן כמעט את הדרישה שלנו. נסכים לכל החלטה שצדק באופן משוער: כל החלטה  $h_S$  בעל  $\epsilon \leq L_{\mathcal{D},f}(h_S)$ , עבור  $\epsilon$  המוגדר בתנאי המשחק.

**הגדרה**

כאשר לאלגוריתם למידה  $A$ , יש חסם עליון  $0 > \epsilon$  על השגיאה (עבור  $\epsilon$  כלשהו), נאמר כי האלגוריתם הלמידה  $A$  **בערך** **צודק**, עם **דיוק**  $\epsilon$ .  
accuracy

#### 4.7 אלגוריתם למידה PAC

כאשר אלגוריתם למידה  $A$  מפיק כל אקראי  $h_S$  בעל חסם עליון  $0 > \epsilon$  על השגיאה בהסתברות של לפחות  $\delta - 1$  (ביחס ל-  $S$ ) עבור  $0 > \delta$  כלשהו, אבל גם בעל שגיאה גדולה מאוד בהסתברות של לפחות  $\delta$ , נאמר כי  $A$  הוא **PAC** (Probably Approximately Correct) עם **דיוק**  $\epsilon$  ובטיחות  $\delta$ .

בנוטציה מתמטית, האלג'  $A$  הוא אלג' למידה PAC (עם דיוק  $\epsilon$  ובטיחות  $\delta$ ) אם לכל  $f : \mathcal{D} \rightarrow h_S$ :

$$\mathbb{P}_{\mathcal{D}^m} \{S \in (\mathcal{X} \times \mathcal{Y})^m \mid L_{\mathcal{D},f}(h_S) \leq \epsilon\} \geq 1 - \delta$$

#### 4.7.1 דיוק מול בטיחות

חשוב שנבין את ההבדל בין **הדיוק**  $\epsilon$  לבין **הבטיחות**  $\delta$ :

- **ראשית** אנו דוגמים סט אימון  $S$  באקריאות. אלג' הלמידה רץ על סט זה, ועל כן החזוי שלו הוא אקראי. אם במקורה הסט  $S$  יצא "מוזר", ולא באמות מייצג את  $\mathcal{D}$ , כלל ההחלטה  $h$  שיופק יהיה שגויה. המספר  $\delta$  הוא ההסתברות לכישלון כתוצאה מסוימת אימון  $S$  מוזר.
- **שנית**, אנו בודקים את כלל ההחלטה  $h_S$  שלנו על דאטא חדש, אקראי גם הוא. המספר  $\epsilon$  מתייחס לדיק של  $h$ .

#### 4.8 גרסא שנייה | Learning Game 2.0

מקבאים ערכי דיוק  $0 > \epsilon$  ובטיחות  $0 > \delta$  רצויים. משחקים שוב נגד הטבע, עם תשלים אקראי.

- אנו מתחילהם, ובוחרים מס' דוגמאות  $m$  ואלגוריתם למידה  $A : (\mathcal{X}, \mathcal{Y})^m \rightarrow \mathcal{Y}^{\mathcal{X}}$ . גם  $m$  וגם  $A$  יכולים להיות תלויים ב-  $(\epsilon, \delta)$ .
- הטבע **מכיר את הבחירה שלנו**, ובוחר אחרינו התפלגות  $\mathcal{D}$  מעל  $\mathcal{X}$ , ופונקציית תיוג  $\mathcal{U} : \mathcal{X} \rightarrow \mathcal{Y}$ . כמובן, בחירתו של הטבע יכולה להיות תלויות ב-  $(\delta, \epsilon)$ , ב-  $m$  וגם ב-  $A$ .

- ו נדגם סט  $S$  של  $m$  דוגמאות באופן בלתי תלוי מ-  $\mathcal{D}$ , ומתוויג ע"י  $f$ .

- האלגוריתם  $\mathcal{A}$  מקבל סט  $S$  זה ומוחזר כלל החלטה  $(S) = \mathcal{A}(S)$ .

- התשלום האקראי** מוגדר להיות  $L_{\mathcal{D},f}(h_S)$ .  
התשלום הוא **אקראי** מכיוון שהסט  $S$  הינו אקראי ובעקבות כך גם  $h_S$ .

על מנת להכريع אם הצלחנו במשחק, נגריל מס' רב של פעמים סט דוגמאות  $S$  (עבור אותן בחירות). נספר ונחשב את ההסתברות המאורה  $\{S \mid L_{\mathcal{D},f}(h_S) \leq \varepsilon\} \leq \delta$  על סמך כל אוסף הדוגמאות  $S$ . אם ההסתברות שנמצאה גדולה מ-  $\delta - \varepsilon$ , כלומר אם אלג' הלמידה הוא אכן PAC עם דיוק  $\varepsilon$  ובטיחות  $\delta$  נגד האסטרטגיה הטובה ביותר ביותר של הטבע, נאמר **שהצלחנו במשחק (ביחס לפרמטרים  $\delta, \varepsilon$ )**.

#### 4.8.1 כמה קטן יכול להיות סט הדוגמאות $m$ ?

מידע עולה כספ. נרצה לבחור  $m$  קטן ככל שניתן, שעדין מספק לנו אלג' למידה PAC. אנו חשים כי חייב להיות trade-off קלשהו בין  $(\delta, \varepsilon, m)$ .

### 4.9 אין ארוחות חיים

לצערנו באופן כללי לא נוכל לנתח במשחק 2.0 נגד הטבע, עבור כל פרמטר  $\delta, \varepsilon$ . **למה?** אנו לא יודעים דבר על  $f, \mathcal{D}$ , ואם יהיה יותר מדי אפשרויות ל- $f$ , לא משנה כמה גדול יהיה מס' הדוגמאות שלנו  $m$ , לא נוכל להיות בטוחים שנמצא כלל החלטה מספיק מדויק.

**דוגמא.** נניח כי  $\mathcal{X} = \infty$ . נזכור כי ראשית אנו בוחרים  $m$ , והטבע מכיר ב-  $m$  שבחרנו. לכן נוכל לקבוע את ערכו. אנו חייבים להחליט מה לחזות על נקודות שלא נמצאות בסט האימון שלנו. נסמן את ההחלטה שלנו ע"י  $(x) = g$ . כמובן, אלג' הלמידה  $\mathcal{A}$  מגדיר כי אם  $x \in \mathcal{X}$  לא אובחנו בסט האימון, אז כלל ההחלטה ש-  $\mathcal{A}$  יchiaר יראה  $(x) = h_S(x)$ .

עתה, הטבע בוחר תת-קובוצת סופית  $C \subseteq \mathcal{X}$  בעלת  $|C| > 2m$ . הנחנו  $\infty = |X|$  ולכן זה אפשרי. הטבע בוחר את  $\mathcal{D}$  להיות התפלגות איחידה מעל  $C$ . עבור פונקיות התיוג  $f$ , הטבע בוחר  $f(x) = -g(x)$  לכל  $x \in \mathcal{X}$  (ההפק ממה שיחזע הכלל  $h_S$  על נקודות שלא ראה).

עתה, יהיו  $S$  סט דוגמאות. נגיד  $C \subseteq \text{supp}(S)$  קבוצת כל הנקודות  $x \in \mathcal{X}$  שהיו חלק מסויט הדוגמאות הnockחי. מכיוון ש-  $m \leq |\text{supp}(S)|$  ו-  $\mathcal{D}$  מתפלג אחד על  $C$ , מתקיים כי  $\frac{1}{2} \geq |\{\{x \in \mathcal{X} \setminus \text{supp}(S)\} \cap C\}| \geq |\{\{x \in \mathcal{X} \setminus \text{supp}(S)\} \cap \text{supp}(S)\}|$ . לא הייתה בסט הדוגמאות היא לפחות חצי. בambilים אחרות, ההסתברות לכך שנקודה חדשה הנדodata מ-  $\mathcal{D}$  לא הייתה בסט הדוגמאות היא לפחות חצי.

עתה לפי חוקי המשחק, הסט  $S$  מזון לתוך  $\mathcal{A}$ , ומתקבל  $\mathcal{A}(S) = h_S$ . מה היא השגיאה של  $h_S$ ?

לא קשר למה שיחזע  $h_S$  על נקודות בסט האימון  $S$ , מכיוון שההסתברות של נקודה חדשה מ-  $\mathcal{D}$  להיות חלק מהקובוצה  $\text{supp}(S)$  הוא לפחות חצי, והכלל יראה על נקודה זו  $(x) = g(x)$  או  $(x) = h_S(x)$ .

בעוד  $f(x) = -g(x)$ , נסיק כי עבור סט דוגמאות זה מתקיים  $L_{\mathcal{D},f}(h_S) \geq \frac{1}{2}$ .

הבעיה היא שמדובר זה קורה לכל סט דוגמאות  $S$ . כמובן הטבע בחר אסטרטגיה לפיה בהסתברות 1 מעל אוסף דוגמאות  $S$  קלשו, תוצאות המשחק תהיה  $L_{\mathcal{D},f}(h_S) \geq \frac{1}{2}$ .

### 4.9.1 מה הבעה?

הטבע יכול לבחור **כל פונקציית תיוג לבחירתו**, ואנו ניסינו ללמידה (לחזות בדיק) את  $f$  מספר דגימות שהיה קטן מידי ביחס למספר הפונקציות האפשריות שהיינו צריכים לבחור מבינם. נלמד לכך שלא יוכל לבנות אלג' למידה PAC אם קבוצת פונקציות התיוג האפשריות גדולה מידי.

זהו בעצם **עקרון אין ארוחות חינם**. ללא מידע כלשהו על פונקציות התיוג אותן אנו מנסים ללמידה, אנו מגלים כי הלמידה בלתי אפשרית.

**חשיבות לשים לב** כי הדוגמא שניתנה למשתמש היא בפועל שגوية. היא לא מספקת הוכחה לעקרון אך מספקת את האינטואיציה הנדרשת לה נזדקק בהמשך.

### 4.9.2 עקרון אין ארוחות חינם פורמלי

יהי  $\mathcal{A}$  מרחב דגימות אינסופי,  $\infty = |\mathcal{A}|$ . לכל  $\frac{1}{\delta} < \varepsilon < 0$  קיים  $0 > \delta$  כך שלכל אלג' למידה  $\mathcal{A}$  וسط אימון מוגדל  $m$  קיימת התפלגות  $(x) \mathcal{D}$  מעל  $\mathcal{A}$  ופונקציית תיוג  $\mathcal{U} \rightarrow \mathcal{A}$ : כך שבהתברות של לפחות  $\delta$  מעל הסט הנבחר  $S$  של  $m$  דגימות שנדרשו באופן תלוי מ-  $\mathcal{D}$ , מתקיים  $\varepsilon \geq L_{\mathcal{D},f}(h_S) \geq L_{\mathcal{D},f}(h_S) \geq \varepsilon$ .

## 4.10 מחלקות היפוטזות

על מנת שנוכל ללמידה, אלג' הלמידה חייב לקבל מידע מקדים על הפונקציה  $f$ . משמעות הדבר לכך שהוא יכול להניח כי  $f$  מגיעה מחלוקת היפוטזות כלשהי,  $\mathcal{U} \subseteq \mathcal{H}$ .

### 4.10.1 ההנחה הריאלייזבילית

נניח כי הוגדרה מחלוקת היפוטזות  $\mathcal{H}$  כלשהי עבור המשחק שלנו. המקרה הריאלייזבילי הוא שהטבע חייב לבחור פונקציה  $\mathcal{H} \in f$ . למעשה לא משנה לנו אם הטבע יבחר  $f$  שלא בהכרח ב- $\mathcal{H}$ , של עוד  $h^* \in \mathcal{H} \approx^{\text{as}} f$  עבור  $\mathcal{H}$ . ההנחה הריאלייזבילית מבחן פורמלית היא שהטבע בוחר פונקציה  $f$  שקיימת עבורה  $\mathcal{H} \in h^*$  בעלת  $0 = L_{\mathcal{D},f}(h^*)$ .

## 4.11 גרסא שלישית | Learning Game 3.0

זהו הגרסה הסופית למשחק. מקרים ערבי דיווק  $0 > \varepsilon$  ובתיוחות  $\delta > 0$  רצויים. מקרים מחלוקת היפוטזות  $\mathcal{U} \subseteq \mathcal{H}$  מושכים שוב נגד הטבע, עם תשלום אקראי.

- אנו מתחילה, ובוחרים מס' דגימות  $m$  ואלגוריתם למידה  $\mathcal{H} \rightarrow (\mathcal{X}, \mathcal{Y})^m$ .

גם  $m$  וגם  $\mathcal{A}$  יכולים להיות תלויים ב-  $(\varepsilon, \delta)$ .

- הטבע **מכיר את הבחירה שלנו**, ובוחר אחרינו התפלגות  $\mathcal{D}$  מעל  $\mathcal{X}$ , ופונקציית תיוג  $\mathcal{H} \in f$  (או פונקציה  $\mathcal{D}$ -כמעט-תמיד זהה לפונקציה מ- $\mathcal{H}$ ).  
כלומר, בחריתו של הטבע יכולה להיות תלואה ב-  $\mathcal{H}, \varepsilon, \delta, m$  וגם ב- $\mathcal{A}$ .

- נדגם סט  $S$  של  $m$  דוגמאות באופן בלתי תלוי מ-  $\mathcal{D}$ , ומתויג ע"י  $f$ .

- האלגוריתם  $\mathcal{A}$  מקבל סט  $S$  זה ומחזיר כלל החלטה ( $S$ )  
 $.h_S = \mathcal{A}(S)$

- התשלום האקראי** מוגדר להיות  $L_{\mathcal{D},f}(h_S)$ .  
התשלום הוא **אקראי** מכיוון שהסט  $S$  הינו אקראי ובעקבות כך גם  $h_S$ .

על מנת להכיריע אם הצלנו במשחק, נגידיל מס' רב של פעמים סט דוגמאות  $S$  (עבור אותם בחירות).  
נספור ונחשב את ההסתברות המאורע  $\{S \sim \mathcal{D}^m \mid L_{\mathcal{D},f}(h_S) \leq \varepsilon\}$  על סמך כל אוסף הדוגמאות  $S$ .  
אם ההסתברות שנמצאה גדולה מ-  $\delta - 1$ , כלומר אם אלג' הלמידה הוא אכן PAC עם דיוק  $\varepsilon$  ובטיחות  $\delta$  נגד האסטרטגייה הטובה ביותר של הטבע, **נאמר שהצלהנו במשחק (ביחס לפרמטרים  $\mathcal{H}, \varepsilon, \delta$ )**.

#### 4.11.1 האם נוכל לקוות ללמידה $\mathcal{H}$ המקיים $\infty = |\mathcal{H}|$ ?

ראינו לעיל כי במקרה בו  $= \infty = |\mathcal{X}|$  ולא נוכל להצליח במשחק הלמידה 3.0.  
אם זה פשוט בלתי אפשרי ללמידה כאשר  $\infty = |\mathcal{X}|$ ?  
למזמן, כאשר מחלוקת ההיפותזות "קטנה מספיק", נוכל להצליח במשחק.  
נתבונן בדוגמה הבאה עבור  $\mathcal{X} = \mathbb{R}$  ומחלוקת היפותזות  $\mathcal{H}$  פשוטה מאוד,  
ונראה כי נוכל להצליח במשחק הלמידה 3.0 עבור כל ערכי  $\delta, \varepsilon$  מוגדרים.

### 4.12 פונקציות סף

נדיר תחום  $\mathcal{X} = \mathbb{R}$  וטוחה תגיות  $\mathcal{Y} = \{0, 1\}$ .  
מחלקת ההיפותזות של פונקציות הסף מעל  $\mathbb{R}$  מוגדרת באופן הבא:

$$\mathcal{H}_{th} = \{x \mapsto h_\theta(x) : \theta \in \mathbb{R} \cup \{\pm\infty\}\}$$

עבור  $0 < x \leq \theta$   $h_\theta(x) = 1$  ו-  $h_\theta(x) = 0$  לכל  $\theta \leq x$ .

#### 4.12.1 הגדרת אלגוריתם הלמידה

נציג את אלג' הלמידה הבא. האלג' מוחזיר את היפוטזה  $(x) \mapsto h_{\theta_{alg}}(x)$  עבור

$$\theta_{alg} = \max_{y_i=0} x_i$$

כאשר אם  $y_i = 1$   $\theta_{alg} = \infty$ , כלומר לכל  $i \in [m]$  נחיזר את כל הנקודות לערך 1.  
באופן דומה, אם  $y_i = 0$   $\theta_{alg} = \infty$ , כלומר לכל שמסוווג את כל הנקודות לערך 0.

### 4.12.2 טענה

נקבע  $0 < \delta, \varepsilon$ . אם  $\frac{\log(\frac{1}{\delta})}{\varepsilon} \geq m$ , אז לכל התפלגות  $\mathcal{D}$  מעל הממשיים, ועבור כל בחירת פונקציית תיוג סף  $f_\theta \in \mathcal{H}_{th}$ , השגיאה  $(h_{\theta_{alg}}(x), f_\theta(x))$  של אלגוריתם למידת פונקציות הסף תהיה לכל היותר  $\varepsilon$ , בהסתברות של לפחות  $\delta - 1$ .

#### הוכחה

נקבע התפלגות  $\mathcal{D}$  מעלה  $\mathbb{R}$  והיפוטזה  $f_\theta \in \mathcal{H}_{th}$ . (זהו הצעד של הטבע במשחק). מתכונת אלג' הלמידה אנו יודעים כי  $\theta_{alg} \leq \theta$ . נבחן כי ככל החלטה שMOVED ע"י האלג' יתייג נכון דגימות המקיימות  $\theta_{alg} < x \leq \theta$  ויטה עבור דגימות  $\theta < x \leq \theta + \varepsilon$ . אם ההסתברות  $\varepsilon < \mathcal{D}(\{x | \theta_{alg} < x \leq \theta\}) < \mathcal{D}(\{x | \theta < x \leq \theta + \varepsilon\})$ , ולכן הטועות האמיתית **תמיד** קטנה מ-  $\varepsilon$  ללא תלות ב-  $S$  או  $m$ . נניח כי  $\varepsilon \geq \mathcal{D}(\{x | \theta' < x \leq \theta\}) = \varepsilon'$  ונדרש כך ש-  $\varepsilon' \leq \theta_{alg} \leq \theta$ . נבחן כי אם קיים  $x_i \in S$  (כי  $\theta' < x_i \leq \theta$ ) עבורו  $y_i = f_\theta(x_i) \neq 0$  (כי  $\theta_{alg} \leq \theta$  ו-

$$L_{D,f_\theta}(h_{\theta_{alg}}) = \mathcal{D}(\{x | \theta_{alg} < x \leq \theta\}) \leq \mathcal{D}(\{x | \theta' < x \leq \theta\}) = \varepsilon$$

ההסתברות לכך שלא קיבל דגימה כזו היא  $(1 - \varepsilon)^m$ .

נשתמש בעובדה כי  $e^{-\varepsilon m} \leq 1 - \varepsilon$ , ונראה כי הביטוי  $e^{-\varepsilon m}$  יהיה קטן מ-  $\delta$  אם

### 4.12.3 סיכום

ראינו כי עבור תחום  $\mathcal{X} = \mathbb{R}$ , טווח תוצאות  $\mathcal{Y} = \{0, 1\}$  ומחלקה היפוטזות  $\mathcal{H} = \mathcal{H}_{th}$  יש לנו אסטרטגיה עבורה **תמיד נצליח** כנגד הטבע, עבור כל ערכי  $\delta, \varepsilon$  נתוניים. בambilים אחרות, עבור כל ערכי  $\delta, \varepsilon$ , האסטרטגיה שלנו היא אלג' למידה PAC ללא קשר לאיך שהטבע משחק.

במקרה בו התחום  $\mathcal{X} = \mathbb{R}$ , טווח תוצאות  $\mathcal{Y} = \mathbb{R}$  ומחלקה היפוטזות  $\mathcal{H} = \mathbb{R}^{\mathbb{R}}$ , ראיינו כי היו ערכי  $\delta, \varepsilon$  עבורם לא יכולנו להצליח לא משנה איך בחרנו לשחק. נסיק מכך שהאופציה עבורה נצליח עבור כל ערך של  $\delta, \varepsilon$  היא כנראה תוכונה של מחלקה היפוטזות. עובדה זו מחייבת אותנו להגדירה של **למידות PAC**.

## 4.13 מחלקות היפוטזות סופיות

על מנת להבין קצת יותר על متى למידות PAC היא אפשרית, נתבונן תחילתה במחלקות היפוטזות סופיות. נבחן כי על אף סופיותן, הן יכולות להיות גדולות מאוד. אנו עלולים לחשב כי אין שהוא ניתן לחישוב כליליות כזו, אך מסתבר שקיים אלג' למידה פשוט **תמיד יעבד** על מחלקות היפוטזות סופיות. הרעיון מאחוריו האלג' פשוט: ננסה להיות כמה שיותר צודקים על סט האימון.

## 4.14 מזעור הסיכון האמפירי ERM

מבחן פורמלית, בהינתן סט אימון  $(x_1, y_1), \dots, (x_m, y_m)$  נגדר את הסיכון האמפירי של כלל החלטה  $h \in \mathcal{H}$  כleshו ע"י:

$$L_S(h) = \frac{1}{m} \cdot |\{i \mid h(x_i) \neq y_i\}|$$

אלגוריתם הלמידה שלנו פשוט. בהינתן סט אימון  $S$ , נרצה להציג את  $h \in \mathcal{H}$  המזער את הסיכון האמפירי בambilim אחרות, ניתן לתאר את האלגוריתם ע"י:

$$\mathcal{A}_{\text{ERM}} : S \mapsto \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$$

המינימום לא חייב להיות יחיד, ובמקרה כזה האלג' שלנו יחזיר את אחד מהם. אלג' זה נקרא Empirical Risk Minimization (ERM) Learner, ונסמן ע"י  $\text{ERM}_{\mathcal{H}}$ .

**איך אנו יודעים כי בהכרח קיים מינימום?**

נבחן כי  $L_S(h) \geq 0$  וכי אנחנו מוחשים מינימום מעלה קבוצה סופית, ולכן בהכרח קיים אחד. בפועל תחת הנחת הריאליות לפיה הטבע בוחר  $\mathcal{H} \in f$ , אנו יודעים כי לכל סט אימון  $S$  מתקיים  $L_S(f) = 0$  עבור אותה פונקציה  $f$  אותה בחר הטבע, ולכן החסם התיכון של 0 הוא יישג. בambilim אחרות, תחת ההנחה  $\mathcal{H} \in f$ , אלג' הלמידה ERM תמיד יציג כלל החלטה עם  $y_i = h(x_i)$  לכל  $i \in [m]$ . חוק כזה נראה **עקבי** Consistent - הוא עקבי עם סט האימון.

### 4.14.1 טענה

יהו  $\mathcal{X}$  מרחב דגימות,  $\mathcal{Y} = \{0, 1\}$  ו-  $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$  עם  $|\mathcal{H}| < \infty$ . יהי  $h_S^{\text{ERM}}$  כלל היסק של אלג' למידה  $\text{ERM}_{\mathcal{H}}$  נקבע  $(0, 1) \in (\delta, \varepsilon)$ . לכל התפלגות  $\mathcal{D}$  ולכל פונקציה  $f$ , מתקיים  $\varepsilon \leq L_{\mathcal{D}, f}(h_S^{\text{ERM}}) \leq 1 - \delta$ .

#### הוכחה

בהינתן סט דגימות  $S$ , "יתכנו מס' כללי הכללה שימזעו את כלל ה- ERM". נסמן את אוסף כל כללי ERM ע"י  $\text{ERM}_{\mathcal{H}}(S)$ . נסמן ב-  $\mathcal{H}_B = \{h \in \mathcal{H} \mid L_{\mathcal{D}, f}(h) > \varepsilon\}$  סט דגימות רנדומי. לפי הגדרה מתקיים כי:

$$\{S \mid L_{\mathcal{D}, f}(h_S^{\text{ERM}}) > \varepsilon\} = \{S \mid h_S^{\text{ERM}} \in \mathcal{H}_B\}$$

נסמן ע"י  $\mathcal{D}^m$  את ההסתברות לדגם את סט הדגימות  $S$ . נרצה להוכיח כי:

$$\mathcal{D}^m(\{S \mid h_S^{\text{ERM}} \in \mathcal{H}_B\}) \leq \delta$$

נזכיר כי תחת הנחת הריאליות, כל כלל  $h$  שיבחר ע"פ עקרון ERM יהיה בעל סיכון אמפירי של אפס ( $0 = 0$ ).

נסמן כעת את אוסף סטי הדגימות שמטיעות אותן ע"י  $M$ :

$$M = \{S \mid \exists h \in \mathcal{H}_B \text{ s.t } L_S(h) = 0\}$$

כלומר, לכל סט דגימות  $S \in M$  קיים כלל החלטה  $h$  שלמרות שהסיכון האמפירי שלו הוא אפס (זיהוי מושלם), ובכך עלול להיבחר ע"י ERM, מקיים כי שגיאת ההכללה שלו גדולה מ- $\varepsilon$ .

כעת נבחן כי  $M \subseteq \{S \mid h_S^{\text{ERM}} \in \mathcal{H}_B\}$ , מכיוון שאלו הדגימות עבורם האלג' שלנו בוחר כלל רע (שמקיים  $\varepsilon > L_{\mathcal{D},f}(h)$ ), ומכיון שהאלג' שלנו פועל לפי עקרון ERM אז בהכרח מתקיים  $L_S(h) = 0$ . מכאן נבין כי על מנת להוכיח את מה שרצינו, מספיק לבחור  $m$  כך ש-  $\delta < \mathcal{D}^m(M)$ . נרצה לכתוב את  $M$  בצורה אחרת:

$$M = \bigcup_{h \in \mathcal{H}_B} \{S \mid L_S(h) = 0\}$$

נפעיל כעת חסם האיחוד ונקבל:

$$\mathcal{D}^m(\{S \mid h_S^{\text{ERM}} \in \mathcal{H}_B\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S \mid L_S(h) = 0\})$$

נבחן כי לכל כלל  $\mathcal{H} \in h$  מתקיים כי  $\mathcal{D}^m(\{S \mid L_S(h) = 0\})$  היא בדיקת הסתברות לדגום סט דגימות  $S$  עליו  $h$  צודקת בכל התוצאות. מכיוון שה-  $x_i$  נדגמים באופן בלתי תלוי, ההסתברות זהה לכך ש-  $h$  תצדק עבור כל  $x_i$  בנפרד. נזכיר כי הגדכנו את ההסתברות לכך ש-  $h$  **טעעה** על  $x$  אקראי היא בדיקת  $L_{\mathcal{D},f}(h)$ .

מכאן שההסתברות שתצדק היא  $(1 - L_{\mathcal{D},f}(h))^m$ , וההסתברות שהיא תצדק על  $m$  ככל שהיא היא  $(1 - L_{\mathcal{D},f}(h))^m$ . מכאן שלכל  $h$  מתקיים:

$$\mathcal{D}^m(\{S \mid L_S(h) = 0\}) = (1 - L_{\mathcal{D},f}(h))^m$$

בפועל אם  $h \in \mathcal{H}_B$  אז הגדכנו כי  $\varepsilon < (1 - \varepsilon)^m$  ולכן  $L_{\mathcal{D},f}(h) > \varepsilon$  לבסוף נקבל:

$$\mathcal{D}^m(\{S \mid h_S^{\text{ERM}} \in \mathcal{H}_B\}) \leq \sum_{h \in \mathcal{H}_B} (1 - \varepsilon)^m \leq |\mathcal{H}_B| \cdot (1 - \varepsilon)^m \leq |\mathcal{H}| \cdot (1 - \varepsilon)^m$$

נשתמש באינטואיטיבי הבא:  $\mathcal{D}^m(\{S \mid h_S^{\text{ERM}} \in \mathcal{H}_B\}) < |\mathcal{H}| \cdot e^{-\varepsilon m}$  ונקבל כי  $1 - \varepsilon \leq e^{-\varepsilon}$ . האגף הימני יהיה קטן יותר שווה  $\delta$  אם  $L_{\mathcal{D},f}(h_S^{\text{ERM}}) > \varepsilon$  קטנה מ- $\delta$ , ועל כן מתקיים כי  $\varepsilon \leq L_{\mathcal{D},f}(h_S^{\text{ERM}})$  בהסתברות של לפחות  $\delta - 1$ ,-CNDR.

**4.14.2 היחס בין  $m, \delta, \varepsilon$**

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{\varepsilon} \quad \left| \begin{array}{l} \varepsilon \leq \frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{m} \\ \delta \leq \frac{|\mathcal{H}|}{e^{m\varepsilon}} \end{array} \right.$$

## 4.15 מחלקות היפותזות סופיות - סיכום

הראינו כי מחלוקת היפותזות סופית  $\mathcal{H}$  היא למידה PAC ע"י שימוש באלו' למידה ERM, בעל סיבוכיות מודל  $\frac{\log(\frac{|\mathcal{H}|}{\delta})}{\varepsilon} m_{\mathcal{H}}(\varepsilon, \delta) \leq$ .

- האם החסם  $\frac{\log(\frac{|\mathcal{H}|}{\delta})}{\varepsilon} m_{\mathcal{H}}(\varepsilon, \delta)$  הדוק? האם אלו' הלמידה ERM או אלו' למידה אחר יכול להיות אלו' למידה PAC עם מס' דוגמאות קטן יותר?
- מה קורה כאשר יש רעש והtagיות  $y$  לא נקבעות בצורה דטרמיניסטיבית ע"י  $x$ ?
- מה קורה כאשר מחלוקת ההיפותזות היא אינסופית?

## 4.16 VC מימד

נניח שקיבלנו סט אימון  $(x_1, y_1), \dots, (x_m, y_m) = S$ , והצלחנו למצוא פונקציה  $h \in \mathcal{H}$  בעלת סיכון אמפירי 0.  $L_S(h) = 0$ .  
 נניח בעת כי חיבלנו בסט  $S$  ע"י שינוי תגיות  $y$ . נקרא לסט המוחובל  $S'$ .  
 נניח כי הצלחנו למצוא  $h' \in h$  שונה מ-  $h$  גם עבורי'  $S'$ , המכילה 0  $L_{S'}(h') = 0$ .  
 אם נוכל לעשות זאת, ללא תלות באיך שנחבל בתגיות, משמעות הדבר היא שימושו לא בסדר:  
 איך נוכל לקוות להכليل בהסתמך על סט אימון  $S$  אם ללא קשר לתגיות של  $S$  אנו יכולים למצוא  $h \in \mathcal{H}$  עם  $h = 0$ ?

### 4.16.1 הגדרה

תהי  $\mathcal{X} \subseteq C$  תת קבוצה של מרחב המדגם ו-  $\mathcal{Y} : h \rightarrow \mathcal{Y}$  היפותזה כלשהי.  
 נגידר את  $h_C$  להיות ההגבלה של  $h$  בתחום  $C$ , ונסמן  $h_C : C \rightarrow \mathcal{Y}$  ע"י  $x \in C$  לכל  $h_C(x) = h(x)$ .

### 4.16.2 אבחנה חשובה

נניח ו-  $\mathcal{H}$  מכילה את כל הפונקציות מעל  $\mathcal{X} \subseteq C$  כלשהו מגודל  $m$  כך ש-  $(\mathcal{Y} \rightarrow \mathcal{Y} : h \rightarrow h_C)$  במרקחה זה לא נוכל למצוא אלו' למידה PAC שימושה ב-  $\frac{m}{2}$  או פחות דוגמאות.  
 טענה זו מtabסת על עקרון "אין ארכות חינס" שראינו קודם, וניתן לבנות דוגמא באופן זהה לחולティין.

ניתן לראות כי כל עוד  $\mathcal{H}$  תכיל קבוצה כלשהי מגודל  $2m$  עם התכונה כי  $\{h_C \mid h \in \mathcal{H}\} = (C \rightarrow \mathcal{Y})$  לא נוכל ללמידה עם סט אימון מגודל  $m$ .

משמעות הדבר היא **הגודל המקסימלי** של קבוצה  $C$  צו הוא קרייטי:

1. הוא מביא לנו חסם תחתון על  $m_{\mathcal{H}}$ , מספר הדוגמאות המינימלי הנחוץ.
2. אם הגודל המקסימלי הוא  $\infty$ , כלומר  $\forall m \in \mathbb{N}$  התחום  $\mathcal{X}$  מכיל קבוצה  $C$  מגודל  $m > |C|$  המקיימת  $\{h_C \mid h \in \mathcal{H}\} = (C \rightarrow \mathcal{Y})$ , המחלוקת  $\mathcal{H}$  אינה למידה PAC.

**4.17 הגדרות פורמליות****4.17.1 ניטוץ**

תהי  $\mathcal{X}$  ותהי  $\mathcal{H}_C = \{h_C \mid h \in \mathcal{H}\}$  ההגבלת של  $\mathcal{H}$  ל- $C$ , כלומר  $C = \{x_1, \dots, x_{|C|}\} \subseteq \mathcal{X}$ .  
 נקבעו עם  $\mathcal{Y} = \{\pm 1\}$ , על מנת שנוכל לייצג כל  $h_C$  כוקטור  $(h_C(x_1), \dots, h_C(x_{|C|})) \in \{\pm 1\}^{|C|}$ .  
 מספר הוקטוריים האפשריים מהצורה זו הוא  $2^{|C|}$ , ועל כן  $|\mathcal{H}_C| \leq 2^{|C|}$ .  
 נאמר כי  $\mathcal{H}$  מנטצת את  $C$  אם  $|\mathcal{H}_C| = 2^{|C|}$ .

**4.17.2 מימד VC**

מימד VC של מחלקת היפותזות  $\mathcal{H}$  כלשיי מוגדרת ע"י:

$$\text{VCdim}(\mathcal{H}) = \max \{|C| \mid \mathcal{H} \text{ shatters } C\}$$

או במלילים אחרות, מימד VC הוא גודל תת-הקובוצה המקסימלית  $\mathcal{X} \subseteq \mathcal{X}$  כך ש- $(C \rightarrow \mathcal{Y})$  מנטצת את  $\mathcal{H}$ .

**4.17.3 חישוב  $\text{VCdim}(\mathcal{H})$** 

לפי ההגדרה לעיל, על מנת להראות כי  $\text{VCdim}(\mathcal{H}) = d$  علينا להראות כי:

1. קיימת קבוצה  $C$  בגודל  $d$  כך ש- $\mathcal{H}$  מנטצת אותה.

2. לכל קבוצה  $C$  בגודל  $k$  עבור  $k \geq d + 1$ ,  $\mathcal{H}$  אינה מנטצת אותה.

## 5 הרצאה 5 - המשך מודל PAC

### 5.1 המשפט היסודי של הלמידה הסטטיסטיות

נציג כעת את המשפט בניסוחו הפורמלי.

תהי  $\mathcal{H}$  מחלקה היפותזות של מסובגים ביןaries עם מימד VC המקיים  $\infty < d$ . במקרה זה,  $\mathcal{H}$  היא למידה PAC אם ורק אם  $\infty < d$ .

1. קיימים קבועים אבסולוטיים  $C_1, C_2$  כך שסיבוכיות המודל של  $\mathcal{H}$  מקיימת:

$$C_1 \cdot \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \cdot \frac{d \cdot \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)}{\varepsilon}$$

2. בנוסף, החסם העליון על סיבוכיות המודל מושג ע"י אלג' למידה ERM.

### 5.2 הרחבת המודל

#### 5.2.1 הקדמה

המסגרת התיאוריתית ללמידה של מידות PAC היא בעייתית. נציג מס' בעיות:

1. **לא מדלת דאטא מושג**. במציאות, למורות שתגית  $-\lambda \in \mathcal{X}$  כלשהו הייתה אמרה להיות 1 (לדוגמא), היא יכולה להימדד כמינוס 1, מסיבות כמו טעות, רעש וכו'.

2. **התנחה הריאלייזבילית היא אינה ריאלית**. אנו מגדירים את מחלקה היפותזות. זה לא ריאלי להגביל את "הטבע" לבחור מחלקה היפותזות הספציפית שאנו בחרנו.

3. **מוגבלים ל-** Misclassification loss. יכולנו למדוד את ביצועי המסוג שלנו רק ע"י ה- loss או בשמו loss 1-0. נרצה לאפשר מדידת ביצועים בעזרת כל פונקציית loss שנבחר.

#### 5.2.2 נאפשר רעש

נגידր מעתה את ההתפלגות  $\mathcal{D}$  כהתפלגות מעל  $\mathcal{X} \times \mathcal{A}$ . משמעות הדבר היא שדגימה מ-  $\mathcal{D}$  היא מהצורה  $(x, y) \sim \mathcal{D}$ . שני משתנים מקרים מקבלים ערכים מ-  $\mathcal{X} \times \mathcal{A}$  בהתפלגות  $\mathcal{D}$ .

#### 5.2.3 פקטורי $x$

נדגום  $x$  מההתפלגות השולית של  $X$  עם הסתברות  $\mathbb{P}(X = x)$ .

נבחר כעת את התגית המתאימה ע"י ההסתברות המותנית  $\mathbb{P}(Y = y | X = x)$ .

מכיוון שההתפלגות השולית של המשתנה המקרי  $Y$  היא ברנולי,

קיימת פונקציה  $\mathcal{X} \rightarrow [0, 1]$  כך ש-  $\mathbb{P}(Y = +1 | X = x) = p(x)$ .

אם קיבל כי  $p(x) = 0$  או  $p(x) = 1$  אנחנו שוב במצב דטרמיניסטי עבור התגית של  $x$ . כך בעצם אנו ממלדים רעש במדידה.

### 5.2.4 פקטורי $\mathbb{P}(X = \mathbf{x}, Y = y) = \mathbb{P}(Y = y) \mathbb{P}(X = \mathbf{x} | Y = y)$

נדגום ראשית תגית ע"י משתנה ברנולי אקראי (הערכים האפשריים הם  $\{\pm 1\}$ ).  
לאחר מכן לכל מחלוקת יהיה את ההסתברות שלה עבור הדגימות  $x$ .  
נדגום את  $x$  מ-  $\mathbb{P}(X = x | Y = -1)$  או מ-  $\mathbb{P}(X = x | Y = 1)$  על סמך התגית שנבחרה.

### 5.2.5 הגדרת פונקציית ה- loss מחדש

נשאר לבינתיים עם ה- loss 0-1. בהינתן  $\mathcal{Y} \rightarrow \mathcal{X}, h$ , נגיד:

$$L_{\mathcal{D}}(h) := \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}}(h(\mathbf{x}) \neq y) = \mathcal{D}\{(x, y) \mid h(x) \neq y\}$$

כאשר נבחן כי אין יותר "אמות מוחלטת".  
לא קיים יותר  $f$  לתיגג לפיו, והדבר הכיר קרוב לאמת הוא ההסתברות המותנית  $\mathbb{P}(Y = y | X = x)$ .

### 5.2.6 כבר לא ריאלייזבילי

במקרה הדטרמיניסטי, תחת ההנחה הריאלייזבילית, יכולנו למצוא שגיאה כללית אפשרית:

$$\min_{h \in \mathcal{H}} L_{\mathcal{D}, f}(h) = L_{\mathcal{D}, f}(f) = 0$$

כעת, במקרה הלא דטרמיניסטי, אין אמת מוחלטת. ההנחה הריאלייזבילית כבר לא הגיונית.  
בגלל רעש המדידה, לא נוכל למצוא שגיאה אפשרית. נאלץ כעת לשנות את הגדרת הדיק שולנו:  
נצפיה מאלגוריתם הלמידה שלנו לפולוט כל החלטה עם שגיאת הכללה **גדולה** לכל יותר ב-  $\varepsilon$   
מהשגיאה המינימלית האפשרית  $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ . אך נשאלת השאלה האם מינימום זה קיים? מה נוכל להגיד לגבי?

### 5.2.7 הגדרה - מסوغ בייס אופטימי

בהינתן הסתברות כלשהי  $\mathcal{D}$  מעל  $\mathcal{Y} \times \mathcal{X}$ , נגיד ר את מסوغ בייס האופטימי ל-  $\mathcal{D}$  ע"י:

$$f_{\mathcal{D}}(\mathbf{x}) = \begin{cases} 1 & \mathbb{P}(y = 1 | \mathbf{x}) > \frac{1}{2} \\ 0 & \text{else} \end{cases}$$

כאשר הכלל  $\mathcal{Y} \rightarrow \mathcal{X}$  תלוי ב-  $\mathcal{D}$  הלא ידוע. הוא נקרא גם "כוח עליון" שהיה מספק לנו את  $\mathcal{D}$ , הינו יכולים לסוג ע"פ  $f_{\mathcal{D}}$ .

**תרגיל.** מסوغ בייס אופטימי הוא בעל שגיאת ההכללה הטובה ביותר. ( $\forall g : \mathcal{X} \rightarrow \mathcal{Y} : L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$ )

מכאן נסיק כי לכל מחלוקת היפוטזיות  $\mathcal{Y}^{\mathcal{X}} \subseteq \mathcal{H}$  מתקיים:

$$L_{\mathcal{D}}(f_{\mathcal{D}}) = \min_{h \in \mathcal{Y}^{\mathcal{X}}} L_{\mathcal{D}}(h) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$$

ולכן הקבוצה  $\{L_{\mathcal{D}}(h) \mid h \in \mathcal{H}\}$  חסומה מלרע ונוכל לכתוב

### 5.2.8 הגדרה - Approximately Correct

יהי  $\varepsilon > 0$ . נאמר כי כלל היסק  $h \in \mathcal{H}$  הוא  $\varepsilon$  ביחס להתפלגות  $\mathcal{D}$  על  $\mathcal{X} \times \mathcal{Y}$  אם:

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$$

### 5.2.9 פונקציית הפסד כללית

נגידר **פונקציית loss כללית ע"י פונקציה** ( $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, \infty)$ ,  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ) נרשות  $(x, y, z) = \ell(h(x), y)$  עבור  $\ell$  על  $\mathcal{H}$ . הדוגמא הכי נפוצה לפונקציית הפסד עבור סיוג היא loss 0-1:

$$\ell_{0,1}(h, (x, y)) := \begin{cases} 1 & h(x) \neq y \\ 0 & h(x) = y \end{cases}$$

כאשר ההפסד איתנו עבדנו עד כה יכול להיות מתואר ע"י  $L_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}(\ell_{0,1}(h, (x, y)))$  מעתה ואילך נניח פונקציית הפסד כללית אך לעיתים נחזר ל-  $\ell_{0,1}$ .

### 5.2.10 שגיאת הכללה

לכל התפלגות  $\mathcal{D}$  מעל  $\mathcal{Y} \times \mathcal{X} = \mathcal{Z}$ , היפותזה  $\mathcal{Y} \rightarrow \mathcal{X}$  ופונקציית הפסד כללית  $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, \infty)$  נגידר את שגיאת הכללה של  $h$  המושרת על ידי  $\ell$  ביחס להתפלגות  $\mathcal{D}$  על ידי:

$$L_{\mathcal{D}}(h) := \mathbb{E}_{z \sim \mathcal{D}}(\ell(h, z))$$

## 5.3 למידות PAC

### 5.3.1 אלגוריתם למידה Agnostic-PAC

יהיו  $\varepsilon \in (0, 1)$ ,  $\delta \in (0, 1)$ . נאמר כי אלג' למידה  $\mathcal{A}$  הוא אלג' למידה Agnostic PAC עם בטיחות  $\delta$  ודיק  $\varepsilon$  ביחס לפונק' הפסד  $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, \infty)$ , מחלוקת היפותזות  $\mathcal{H}$  והתפלגות  $\mathcal{D}$  מעל  $\mathcal{Y} \times \mathcal{X}$  אם ההסתברות לקבל סט דוגמאות  $S$  עבורו יפיק  $\mathcal{A}$  כלל היסק  $h_S \in \mathcal{H}$  עם הפסד  $L_{\mathcal{D}}(h_S) \leq \varepsilon$  מההפסד הטוב ביותר שנייה להציג ע"י כל היפותזה ב-  $\mathcal{H}$  היא לפחות  $\delta - 1$ . בכתיבה מתמטית:

$$\mathcal{D}^m \left( \left\{ S \mid L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon \right\} \right) \geq 1 - \delta$$

### 5.3.2 למידה PAC Agnostic

מחלקת היפותזות  $\mathcal{H}$  היא **למידה PAC Agnostic** ביחס לפונק' הפסד  $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, \infty)$  אם קיימת פונקציה  $\mathbb{N} \rightarrow \mathbb{N}$  ואלג' למידה  $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$  עם התכונה הבאה: לכל התפלגות  $\mathcal{D}$  מעל  $\mathcal{Y} \times \mathcal{X}$  ולכל  $\varepsilon, \delta \in (0, 1)$  מתקיים:

$$\mathcal{D}^m \left\{ S_m \mid L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon \right\} \geq 1 - \delta$$

עבור  $h_S = \mathcal{A}(S)$  ו-  $D$  דוגמאות בלתי תלויות מההתפלגות  $S_m = (x_1, y_1), \dots, (x_m, y_m)$

## 5.4 מזעור השגיאה האמפירית

### 5.4.1 הגדרה - שגיאה אמפירית ביחס לפונק' הפסד כללית

יהי כלל החלטה  $\gamma \rightarrow \mathcal{X} : h$ . השגיאה אמפירית של  $h$  ביחס לפונק' הפסד  $\ell$  ודוגמאות  $S = \{z_i\}_{i=1}^m$  מוגדרת ע"י:

$$L_S(h) := \frac{1}{m} \sum_{i=1}^m \ell(h, z_i)$$

### 5.4.2 הגדרה - אלגוריתם למידה ERM עבור Agnostic-PAC

נגידר את אל' הלמידה ERM ע"י  $\mathcal{A}_{\text{ERM}} : S \mapsto \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$ . נבחן כי גם במקרה זה הכלל המתקבל לא בהכרח ייחיד.

## 5.5 תזכורות - החוק החלש של המספרים גדולים

לפי חוק זה, בהינתן  $X_i$  משתנים מקרים בלתי תלויים בעלי תוחלת  $(X_i) = \mathbb{E} = \mu$  מתקיים:

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m X_i = \mu$$

כאשר ההתכונות היא בהסתברות. כלומר, לכל  $\delta > 0$  מתקיים:

$$\lim_{m \rightarrow \infty} \mathbb{P} \left( \left| \frac{1}{m} \sum_{i=1}^m X_i - \mu \right| > \delta \right) = 0 \quad \left| \begin{array}{l} \exists m_0 \in \mathbb{N} \text{ s.t } \forall m > m_0 : \mathbb{P} \left( \left| \frac{1}{m} \sum_{i=1}^m X_i - \mu \right| > \delta \right) < \varepsilon \end{array} \right.$$

### 5.5.1 אבחנה

עבור כל  $h$  מתקיים כי  $(L_S(h)) = L_{\mathcal{D}}(h)$ . לפי החוק החלש של המספרים גדולים, אם  $S$  סט דוגמאות בגודל  $m$  הנדגם באופן בלתי תלוי, אז  $L_S(h) = L_{\mathcal{D}}(h)$  מתכנס בהתפלגות ל-  $L_{\mathcal{D}}(h)$  מכיוון שלכל  $\delta > 0$  קיים  $\varepsilon \in \mathbb{N}$  כך שלכל  $m > m_0$  מתקיים:

$$\mathbb{P}(|L_S(h) - L_{\mathcal{D}}(h)| > \delta) < \varepsilon$$

אך האם משמעות הדבר היא ש-  $\underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_{\mathcal{D}}(h)$

## 5.6 המשפט היסודי של הלמידה הסטטיסטיות עם Agnostic PAC

תהי  $\mathcal{H}$  מחלקה היפותזות של מסוגים ביןaries עם מימד VC המקיים  $\infty < d = \text{VCdim}(\mathcal{H}) \leq \infty$ . איזי,  $\mathcal{H}$  היא למידה Agnostic PAC אם ורק אם  $\infty < d$ . במקרה זה,

קיימים קבועים אבסולוטיים  $C_1, C_2$  כך שסיבוכיות המודל של  $\mathcal{H}$  מקיימת:

$$C_1 \cdot \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \cdot \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$$

בנוסף, החסם העליון על סיבוכיות המודל מושג ע"י אלג' למידה ERM. נרצה לגעת בחלוקת מההוכחה, בעיקר על מנת לקבל אינטואציה על הקשר בין מימד VC לממדות.

## 5.7 חלק ראשון

חלוקת הראשון של המשפט אומר לנו למוד את  $\mathcal{H}$  אם ורק אם מימד VC שלה סופי. באופן ספציפי, אם  $\infty = \text{VCdim}(\mathcal{H})$  איזי לא ניתן לבנות אלג' למידה Agnostic PAC עבור  $\mathcal{H}$  עבור כל אלג' למידה  $\mathcal{A}$  וכל מספר דגימות  $m$ .

דיברנו על נושא זה באופן לא פורמלי - אם קיימת  $\mathcal{X} \subseteq C$  אותה  $\mathcal{H}$  מנתצת, אף אלג' למידה לא יוכל להיות אלג' למידה PAC אם הוא מסתמך על פחות מ-  $\frac{|C|}{2}$  דגימות. אם מימד VC אינסופי, קיימת קבוצה  $C$  מכל גודל ועל כן אף מס' דגימות לא יספיק.

## 5.8 חלק שני

חלוקת השני של המשפט אומר שאלג' למידה ERM הוא אלג' למידה אוניברסלי לכל מחלקה היפותזות  $\mathcal{H}$  מימייד VC סופי. באופן ספציפי, המשפט טוען כי אם  $\mathcal{H}$  היא מחלקה היפותזות מימייד VC סופי, איזי  $\mathcal{H}$  למידה Agnostic PAC עם כל אלג' למידה ERM (כלומר עם כל כלל המציג  $L_S(h)$ ).

### 5.8.1 עקרון למידה ERM

אלג' למידה ERM בוחר כלל  $L_S(h)$  המציג את  $(S)$  (השגיאה האמפירית) עבור סט דגימות  $S$ . אנו מוקוים שהוא חוק:

$$(1) \quad \mathcal{D}^m \{S \in (\mathcal{X} \times \mathcal{Y})^m \mid |L_D(h_S) - L_S(h_S)| < \varepsilon\} \geq 1 - \delta$$

מקרה זה יכול לקרות רק כאשר נדגום  $S$  "מיוחד" -

אחד כזה שעבור כל  $h \in \mathcal{H}$  השגיאה האמפירית  $(h)$   $L_S(h)$  קרבה לשגיאת הכללה  $(h)$ . מדוע זה קשה להוכיח?

נבחן כי לכל  $h \in \mathcal{H}$  מתקיים  $L_S(h) = L_D(h) \rightarrow \mathbb{E}(L_S(h)) = L_D(h)$  וזו מהחוק החלש  $L_S(h)$  בהסתברות.

משמעות הדבר הכי שלכל  $\delta, h, \varepsilon$  קיים  $m_0$  כך שלכל  $m > m_0$  מתקיים  $\mathbb{P}\{|L_D(h_S) - L_S(h_S)| < \varepsilon\} > 1 - \delta$ . נרצה  $m_0$  במידה שווה בתפלגות  $D$  ובהיפותזה  $H$ .

אך  $m_0$  תלוי גם ב-  $D$  וגם ב-  $h$ .

### 5.8.2 התכנסות במידה שווה

סדרת פונקציות  $f_n : X \rightarrow \mathbb{R}$  מתכנסת במידה שווה ל-  $f : X \rightarrow \mathbb{R}$  אם ורק אם:

$$\forall \varepsilon > 0 \quad \exists m_0 \in \mathbb{N} \quad \text{s.t.} \quad \forall x \in X : |f_n(x) - f(x)| < \varepsilon$$

אנו נרצה ש-  $L_S(h)$  תתכנס במידה שווה ל-  $L_{\mathcal{D}}(h)$  ב-  $\mathcal{D}$  וב-

### 5.8.3 סט דוגמאות אפסילון מייצג

סט דוגמאות  $S$  הוא  $\varepsilon$ -מייצג עבור  $\mathcal{D}, \mathcal{H}, \ell$ , אם ורק אם:

תנאי זה יבטיח כי מיעור  $L_S(h)$  מעל  $\mathcal{H}$  (ERM) יהיה מאוד קרוב למיעור  $L_{\mathcal{D}}(h)$  מעל  $\mathcal{H}$ . אם יהיה לנו סט דוגמאות  $S$  שהוא  $\varepsilon$ -מייצג, אז  $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$  "כמעט" ישיג  $\text{ERM}_{\mathcal{H}}(S)$ . באופן פורמלי:

למה. יהיו  $S$  סט דוגמאות  $\frac{\varepsilon}{2}$ -מייצג עבור  $\mathcal{D}, \mathcal{H}, \ell$ . ה-  $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$  פلت של  $\text{ERM}_{\mathcal{H}}(S)$  (כלומר  $L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$ )

הוכחה. נסמן  $h^* = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ . מתקיים:

$$L_{\mathcal{D}}(h_S) \stackrel{i}{\leq} L_S(h_S) + \frac{\varepsilon}{2} \stackrel{ii}{\leq} L_S(h^*) + \frac{\varepsilon}{2} \stackrel{i}{\leq} L_{\mathcal{D}}(h^*) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = L_{\mathcal{D}}(h^*) + \varepsilon$$

$$(i) \text{ לפי הגדרה } h_S \in \arg\min_{h \in \mathcal{H}} L_S(h) \quad |L_S(h) - L_{\mathcal{D}}(h)| < \frac{\varepsilon}{2} \Rightarrow -\frac{\varepsilon}{2} < L_S(h) - L_{\mathcal{D}}(h) < \frac{\varepsilon}{2}$$

### 5.8.4 תכונות ההתכנסות במידה שווה

נאמר כי מחלוקת היפותזות  $\mathcal{H}$  בעלת תכונות ההתכנסות במידה שווה אם קיים  $m_{\mathcal{H}}^{\text{UC}}$  כך שלכל  $1 < \varepsilon, \delta < 0$  ולכל התפלגות  $\mathcal{D}$  מעל  $\mathcal{X} \times \mathcal{Y}$  מתקיים:

$$(2) \quad \mathcal{D}^m \{S \in (\mathcal{X} \times \mathcal{Y})^m \mid S \text{ הוא } \varepsilon\text{-מייצג}\} \geq 1 - \delta$$

ובכך קיבלנו כי על מנת להוכיח את חלקו השני של המשפט, علينا להוכיח את (2).

נגדיר פונקציה  $F_m^{\mathcal{D}}(S) := \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$  באופן הבא:

פונקציה זו מיפה סט דוגמאות  $S$  מגודל  $m$  למקרה הגרוע ביותר -

הבדל המקסימלי מעל  $\mathcal{H}$  בין השגיאה האמפירית לשגיאת החכללה. נבחן כי  $F$  היא משתנה מקרי,

כפונקציה של סט דוגמאות  $S$  רדוני, אשר התפלגותו תלויה ב-  $m, \mathcal{D}$ . נרצה להראות כי בהסתברות גבוהה  $F_m^{\mathcal{D}}$  הוא קטן. פורמלית, נרצה להראות כי לכל  $1 < \delta < 0$  קיים  $N \in \mathbb{N}$  כך שלכל התפלגות  $\mathcal{D}$  מעל  $\mathcal{X} \times \mathcal{Y}$  מתקיים:

$$(3) \quad \mathcal{D}^m \{F_m^{\mathcal{D}}(S) > \varepsilon\} < \delta$$

**5.8.5 מקרה ראשון |  $\mathcal{H}$  סופית**

טענה. נקבע  $\delta, \varepsilon$ . קיימים  $m_0 > m_0(\delta, \varepsilon)$  מתקיים לכל  $\mathcal{D}$  כי  $\delta < \varepsilon$  מוגדרת מתקיים כי:

$$\begin{aligned} \mathcal{D}^m \{F_m^{\mathcal{D}}(S) > \varepsilon\} &= \mathcal{D}^m \{S \mid \exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \\ &\leq \sum_{h \in \mathcal{H}} \mathcal{D}^m \{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \quad (\text{חסם האיחוד}) \\ &\leq |\mathcal{H}| \cdot \max_{h \in \mathcal{H}} \mathcal{D}^m \{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \end{aligned}$$

ולכן עליינו לחסום את  $\mathcal{D}^m \{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}$  במידה שווה עבור  $\mathcal{D}, h$ . מהחוק החלש נקבל כי מכיוון  $L_S(h) - L_{\mathcal{D}}(h)$  הוא הממוצע האמפירי של משתנים מקרים בלתי תלויים בעל תוחלת  $\mathbb{E}(L_S(h) - L_{\mathcal{D}}(h)) = 0$  וידועים כי לכל  $0 < \varepsilon$  מתקיים  $\mathcal{D}^m \{|L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \xrightarrow{m \rightarrow \infty} 0$ .

אך זה לא מספיק, מכיוון שנרצה שתלויה ב-  $\mathcal{D}$ .

נרצה להשתמש באי-שוויון הופדייג. נגדיר  $\ell(h, (x_i, y_i)) = \ell_i$ . נבחן כי  $\theta_i := \ell_i(h, (x_i, y_i))$  וכי  $L_S(h) = \frac{1}{m} \sum_{i=1}^m \theta_i$ . נקבע:  $L_{\mathcal{D}}(h) - L_S(h)$  במידה שווה ב-  $h$  וב-  $\mathcal{D}$ .

$$\mathcal{D}^m \{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \leq 2 \exp(-2m\varepsilon^2)$$

וכעת:

$$\begin{aligned} \mathcal{D}^m \{S \mid \exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} &\leq \sum_{h \in \mathcal{H}} \mathcal{D}^m \{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \\ &\leq |\mathcal{H}| \cdot \max_{h \in \mathcal{H}} \mathcal{D}^m \{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \\ &\leq 2|\mathcal{H}| \exp(-2m\varepsilon^2) \\ \mathcal{D}^m \{F_m^{\mathcal{D}}(S) > \varepsilon\} &\leq 2|\mathcal{H}| \exp(-2m\varepsilon^2) < \delta \quad \text{נקבל } m \geq \frac{\log(\frac{2|\mathcal{H}|}{\delta})}{2\varepsilon^2} \end{aligned}$$

**5.8.6 מקרה שני |  $\mathcal{H}$  אינסופית**

כמובן שלא ניתן להשתמש בטיעונים הקודמים במקרה זה. נזכיר כי עבור קבוצה סופית  $\mathcal{X}$  נסמן ב-  $\mathcal{H}_C$  את קבוצת ההיפותזות המוגבלות לקבוצה  $C$ . המפתח להוכחה הוא להבין מהר הגבלה  $\mathcal{H}_C$  גדלה עם  $|C|$ . אם  $\text{VCdim}(\mathcal{H}) \leq |C|$  יתכן  $\mathcal{H}$  מוגבלת את  $C$ . איזי יתכן  $|\mathcal{H}_C| = 2^{|C|}$ ? איזי  $|\mathcal{H}_C| = 2^{|C|}$ ? כמה גדול יכול להיות  $|\mathcal{H}_C|$ ? איזי  $|\mathcal{H}_C| > \text{VCdim}(\mathcal{H})$ ?

עבור כל מחלקה היפותזות  $\mathcal{H}$  נגדיר  $\tau_{\mathcal{H}}(m) = \max \{|\mathcal{H}_C| \mid C \subseteq \mathcal{X}, |C| = m\}$  על ידי  $\tau_{\mathcal{H}}(m) = 2^m$ . זה מס' הפונקציות המקסימלי שיכולים להתקיים ע"י הגבלת  $\mathcal{H}$  לכל תת-קבוצה של  $\mathcal{X}$  מוגדל  $m$ .

כל ש- $\mathcal{H}$  גדולה ומסובכת יותר, נצפה ש-  $\tau_{\mathcal{H}}(m)$  יגדל.

במילים אחרות,  $\tau_{\mathcal{H}}(m)$  מודדת כמה מהר - לכל היותר - יכולה  $\mathcal{H}_C$  להיות גדולה יחד עם  $|C|$ . לדוגמה ראיינו כי אם  $\text{VCdim}(\mathcal{H}) = \infty$  אז  $\tau_{\mathcal{H}}(m) = 2^m$  כלומר  $\mathcal{H}_C$  יכולה להיות אקספוננציאלי ב-  $|C|$ .

**הגדרה**

יהי  $\mathcal{H} \subseteq \mathcal{C}$ . נניח שקיים  $b > 0$ ,  $m_0 \in \mathbb{N}$ ,  $m > m_0$  כך שלכל  $m > m_0$  מתקיים: אזי נאמר כי  $\mathcal{H}_C$  גdal פולינומיאלית ב-  $|C|$ . ההוכחה כעת מתחלקת לשני חלקים:

1. אם  $H_C$  גdal פולינומיאלית ב-  $|C|$ , אזי  $\mathcal{H}$  בעלת תוכנות ההתכנסות במידה שווה, ועל כן למידת Agnostic PAC.

2. אם  $\text{VCdim}(\mathcal{H}) < \infty$ , אזי  $\mathcal{H}_C$  גdal פולינומיאלי ב-  $|C|$ .

**הוכחת חלק 1**

טעינה. אם  $H_C$  גdal פולינומיאלית ב-  $|C|$ , אזי  $\mathcal{H}$  בעלת תוכנות ההתכנסות במידה שווה. הוכחה. נזכר כי אנו רוצים להוכיח כי  $\delta < \varepsilon$ . נסמן  $\mathcal{D}^m = \{F_m^{\mathcal{D}}(S) : S \in \mathcal{D}\}$ . מכיוון ש-  $F_m^{\mathcal{D}}$  הוא משתנה מקרי אי-שלילי, נוכל לחסום אותו ע"י "א"ש מרקוב. נרצה למצוא סדרה של מספרים  $\alpha_m$  כך ש-

$$\mathbb{E}_{\mathcal{D}^m}(F_m^{\mathcal{D}}(S)) \leq \alpha_m$$

הקסם מתחבא בעובדה כי  $\alpha_m$  תהיה תלולה ב-  $\mathcal{H}$  אך לא ב-  $\mathcal{D}$ , ובכך נחסום את התוחלת במידה שווה מעל  $\mathcal{D}$ . אם נמצא סדרה כזו אזי נקבל כי:

$$\mathbb{P}_{\mathcal{D}^m} \left\{ \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| > \varepsilon \right\} = \mathbb{P}_{\mathcal{D}^m} \{ F_m^{\mathcal{D}}(S) > \varepsilon \} \leq \frac{\mathbb{E}_{\mathcal{D}^m}(F_m^{\mathcal{D}}(S))}{\varepsilon} \leq \frac{\alpha_m}{\varepsilon}$$

במילים אחרות, בהסתברות של לפחות  $\frac{\alpha_m}{\varepsilon} - 1$ , סט דוגמאות  $S$  באורך  $m$  יהיה  $\varepsilon$ -מייצג!

הצלחנו להשיג במידה שווה מעל  $\mathcal{H}$   $h \in \mathcal{H}$  ומעל  $\mathcal{D}$ .

אם בנוסף נצליח למצוא סדרה  $\alpha_m$  גם יורדת לאפס, אז לכל  $\delta, \varepsilon$  נוכל לקבוע  $m_0$  כזה שלכל  $m > m_0$  מתקיים כי  $\delta < \frac{\alpha_m}{\varepsilon}$ . משמעות הדבר היא ש-  $\mathcal{H}$  בעלת תוכנות ההתכנסות במידה שווה. נמצא סדרה  $\alpha_m$  כזו:

**лемה פרציית - לא נוכיה.** יהיו  $F_m^{\mathcal{D}}$  ממשוואה (3). אזי לכל  $\mathcal{D}$  מתקיים:

$$\mathbb{E}_{\mathcal{D}^m}(F_m^{\mathcal{D}}(S)) \leq O\left(\frac{\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}\right) + o(m)$$

לפי הנחה  $H_C$  גdal פולינומיאלית ב-  $|C|$ , ולכן קיים  $m > m_0$  כך שלכל  $m > m_0$  מתקיים  $O\left(\frac{\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}\right) + o(m)$  מכאן שהסדרה  $(\frac{\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}})$  יורדת לאפס.

**הוכחת חלק 2**

טעינה. אם  $\text{VCdim}(\mathcal{H}) < \infty$ , אזי  $\mathcal{H}_C$  גdal פולינומיאלי ב-  $|C|$ . הוכחה. אנו נוכיח את הלמה הבאה: אם  $\text{VCdim}(\mathcal{H}) < \infty$  אזי  $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$ .

## 6 הרצאה 6 - שיטות ועידה

היום נראה **מטא-אלגוריתמים**. שיטות לשיפור הביצועים של אלג' למידה קיימים. אנו נעבור על אלג' למידה לביעית סיווג, אך ניתן להרחיב לביעות שונות. הביצועים יכולים לקבל שינויים אדריכליים.

### 6.1 הטייה ושונות

- **הטiya** היא חלק מsgivingת הכללה שמתתקבל ע"י ההיפוטזה היא "טובה" מ-  $\mathcal{H}$ . אם חשוב על פונקציית תיוג לא ידועה הנבחנה ע"י הטבע, הטiya מודדת כמה טוב ניתן לשערד את פונקציית התיוג ע"י ההיפוטזה "הכי קרובה" ב-  $\mathcal{H}$ . כמובן שככל ש-  $\mathcal{H}$  גדולה יותר היא בעלת יותר כוח לתיאור פונקציות  $f$  מסובכות יותר, ולכן הטiya שלה נמוכה יותר.
- **השונות** היא חלק מsgivingת הכללה שמתתקבל ע"י העובדה כי סט דוגמאות האימון הוא אקראי, וכן כלל ההחלטה שלנו  $h$  אקראי גם הוא. ככל ש-  $\mathcal{H}$  גדולה יותר, לאלג' הלמידה יותר חופש לדודו. אחרי נקודות אקראיות בסט דוגמאות האימון, שלא בהכרח מיצגות את פונק' התיוג האמתי אותה אנו מנסים ללמוד. ככל שהמודל מסובך יותר, כך קטנה הטiya וגילה השונות. באופן לא פורמלי, שגיית הכללה היא הסכום (או "השלוב") של השניים.

### 6.2 החלטות ועדות

נחשב על עדות  $T$  חברים, שצרכיה לקבל החלטה של כן או לא. כל אחד מבעלי הוועדה מצביע. לכל אחד מהם יש הסתברות של  $p$  להיות צודקים ו-  $1-p$  להיות טוענים, וכל חברי הוועדה "חכמים באופן שווה". לאחר הצביעות של כל חברה הוועדה, דעת הרוב מתתקבלת.  $\bar{X} = \text{sign} \left( \sum_{t=1}^T X_t \right) \stackrel{\text{i.i.d}}{\sim} \text{Bin}(p)$ . ההחלטה הוועדה היא  $\{1\pm1\}$ .

#### 6.2.1 רמת דיק הועדה

אם כל חבר צודק בד"כ ( $p > 0.5$ ), אז ההסתברות שהועדה תהיה צודקת גבוהה בהרבה מהסתברות שאדם בודד יצדיק, והוא תגדל ככל ש-  $T$  גדל.

#### 6.2.2 שונות ההחלטה הוועדה

נתה המצביע שתקבל החלטות החלטות באופן עקבי. כלומר, אם הוא מצביע מספר פעמים, באופן בלתי תלוי, מה הסיכוי שתקבל את אותה ההחלטה?

#### 6.2.3 מציאות

במציאות ועדות לא מצביעות באופן בלתי תלוי. לכן נניח כי לכל 2 חברים ועודה יש קוראלציה  $\rho \leq 0$ , וicut כל חבר צודק בהסתברות (זהה עבור colum) של  $p$  וכל זוג חברים בעל קוראלציה (זהה עבור colum) של  $\rho$ .

### 6.2.4 לסיכום

עודד מחלוקת ע"פ החלטת הרוב. ההחלטות בעלות קוראלציה  $\rho$  וכל משתתף צודק בהסתברות  $p$ . אם  $\frac{1}{2} > p$  ההחלטה הועדה תשתרפ בכל ש- $T$  יגדל בשתי דרכים:  
 היא תצדק הסתברות גבוהה יותר, וגם תהיה יותר עקבית.  
 אם  $0 > \rho$ , הגדלת  $T$  תעוזר לנו רק עד לנוקה מסויימת, ולכן נוונותנו חסם תחתון על השיפור האפשרי מעבר לכך לעודשה שלמה.

## 6.3 ועדות ב- IML

נניח כי ברשותנו  $T$  סט דוגימות אימון מגודל  $m$  שנדרשו באופן בלתי תלוי מ-  $\mathcal{X}$  לפי התפלגות כלשהי  $\mathcal{D}$ .  
 נניח כי ברשותינו אלג' למידה  $A$  אותו נאמן על סטי הדוגימות ונתקבל  $h_{S_1}, \dots, h_{S_T}$ .  
 החיזוי  $(x)$  על  $h_{S_t}$  הוא בלתי תלוי בשאר החיזויים של כלל ההחלטה השונים, אך בעל אותה התפלגות.  
 אין אם השתמש ב-  $h_{S_1}, \dots, h_{S_T}$  כועדה, תוך שימוש בבחירה הרוב, נקבל סיטואציה דומה למקרה שבו שראינו מוקדם.  
 שגיאות הכלכלה תשתרפ ביחיד עם  $T$ , עד לרמה אפסית כאשר  $\infty \rightarrow T$ . (אם לכל כלל בנפרד יש שנייה הכלכלה גדולה ממחצית). כמו שראינו בתרגול, שונות החיזוי קטנה כמו  $\frac{1}{T}$ .

אך בשיטת ה- Batch Learning אין לנו  $T$  סטים. יש לנו רק אחד.  
 הקסם הראשון שנראה הוא איך ליצור  $Y$  סטים של>Data אימון מהסט הבודד  $S$  שלנו,  
 בצורה כזו שתדמיה דוגימות חדשות ובלתי תלויות מגודל  $m$  מהתפלגות  $\mathcal{D}$ .

### 6.3.1 שיטות ועדה - הגדרה

שיטות ועדה הן **מטא-אלגוריתמים**. בשיטת ועדה, אנו לוקחים אלג' למידה  $A$   
 (אשר נקרא לו אלג' למידה "בסיסי" או "חלש"), ונוסף לו  $T$  סטי דוגימות אימון מלאכותיים.  
 בהרצאה זו אנו עובדים עם בעיות סיווג, ולכן אוסף הדוגימות שלנו הוא  $1 = \pm \mathcal{U}$ ,  
 וההחלטה מחלוקת ע"פ  $h(x) = \text{sign} \left( \sum_{t=1}^T h_t(x) \right)$ . כל מה שנראה בהמשך תקף גם עבור בעיות רgression.

## 6.4 שיטת Bootstrap

יצירת סטי>Data "מלאכותיים" מסט דוגימות קיים  $S$  נשמע כמו מהهو בלתי אפשרי.  
 אך למען האמת זה אפשרי. בהינתן סט דוגימות  $\{(x_i, y_i)\}_{i=1}^m = S$ , אנו הולכים ליצור סט דוגימות חדש  $S^{*1}$  באופן הבא:

נדגים  $m$  פעמים **עם החזרה** מהסט  $S$ . הדגימה הראשונה שנדרשו מסומן ע"י  $(\mathbf{x}_1^{*1}, y_1^{*1})$ , השנייה  $(\mathbf{x}_2^{*1}, y_2^{*1})$  וכן הלאה.  
 בעת יש לנו סט דוגימות  $\{(x_i^{*1}, y_i^{*1})\}_{i=1}^m = S^{*1}$ . מבון שיתכנו חזירות בסט זה, גם במידה ו-  $S$  עצמו לא היו חזירות.  
 נחזור על תהליך זה בסה"כ  $B$  פעמים, ונקבל  $B$  סטי>Dogimates אימון שונים מגודל  $m$ :  $S^{*1}, \dots, S^{*B}$ .

שיטת זו נקראת **בוטסטראפ**, וסט הדגימות  $S^{*b}$  נקרא דגימת בוטסטראפ שנוצרה מ- $S$ . נניח כי בעיית הלמידה שלנו היא דגימות זהות ובلتוי תלויות מההתפלגות לא ידועה  $\mathcal{D}$  מעל  $\mathcal{X} \times \mathcal{Y}$ . אנו מוקווים שכל בוטסטראפ מ- $S$  ייכוחו מתנהג כמו סט דגימות חדש שנדגם ובאופן אחד ובלתי מ- $\mathcal{D}$ . זה עלול להישמע מטורף שדגימות הבוטסטראפ יכולות לשמש אותנו כאילו נדגמו מההתפלגות  $\mathcal{D}$ , אך אכן לעיתים קרובות מאוד זהו המקרה.

#### 6.4.1 ההתפלגות האמפירית של $S$

נניח סט דגימות אימון  $S$  כלשהו, ולמען הפשטות נניח כי של הנק'  $x \in S$  ישנוות זו מזו. נגיד את **ההתפלגות האמפירית**  $\hat{\mathcal{D}}_S$  המושראית ע"י  $S$  על  $\mathcal{Y} \times \mathcal{X}$  כהתפלגות הסתברותית על  $\mathcal{Y} \times \mathcal{X}$ : עבור תת קבוצה  $C \subseteq \mathcal{Y} \times \mathcal{X}$ , נגיד:

$$\hat{\mathcal{D}}_S((X, Y) = (x, y)) = \begin{cases} \frac{1}{m} & (x, y) \in S \\ 0 & (x, y) \notin S \end{cases}$$

או באופן שקול לכל  $C \subseteq \mathcal{Y} \times \mathcal{X}$  נגיד  $\hat{\mathcal{D}}_S = \frac{|C \cap S|}{m}$ .

נבחן כי דגימת בוטסטראפ  $S^{*b}$  היא פשוט דגימה אחידה ובلتוי תליה של  $m$  נקודות מההתפלגות האמפירית  $\hat{\mathcal{D}}_S$ . ככל ש- $m$  גדול, ככל רצוי ככל שסט האימון  $S$  גדול, ההתפלגות האמפירית  $\hat{\mathcal{D}}_S$  מתכנסת להתפלגות ל- $\mathcal{D}$ .

הweeney המקורי הבוטסטראפ הוא ש- $\hat{\mathcal{D}}_S$  לא שונה בהרבה מ- $\mathcal{D}$ , וכן  $m$  דגימות זהות ובلتוי תלויות מ- $\hat{\mathcal{D}}_S$  הן שערוך טוב ל- $m$  דגימות זהות ובلتוי תלויות מ- $\mathcal{D}$ .

עליה השאלה כמה נקודות של  $S$  נשארות מוחוץ לכל דגימת בוטסטראפ בדרך כלל? התשובה היא בערך שלישי, והчисוב הושאר כתרגילים.

#### 6.5 Bagging שיטת

זה בעצם ועידה ע"י שימוש בדגימות בוטסטראפ, המאפשרת שיפור דיוק של אלג' למידה קיימ. נתחילה עם אלג' למידה בסיסי כלשהו  $A$  וסט דגימות  $S$ . נבחר  $T$  כלשהו וניצור  $T$  סטי דגימות בוטסטראפ  $S^{*1}, \dots, S^{*T}$  מוגדל  $m$  כל אחד. נאמן את אלג' הלמידה  $A$  **בנפרד** על כל אחד מ- $T$  סטי דגימות הבוטסטראפ. ניצור ועדה  $x$ ,  $h_{S^{*1}}, \dots, h_{S^{*T}}$ , ונאחסן את כל  $T$  המודלים מהאומנים. על מנת לסוג נקודה חדשה  $x \in \mathcal{X}$  נריץ את  $x$  על כל כללי החלטה ונסוווג את  $x$  ע"פ החלטת הרוב:

$$h_{\text{bag}}(x) = \text{sign} \left( \sum_{t=1}^T h_{S^{*t}}(x) \right)$$

### 6.5.1 דגימות חוזרות

על אף הלמידה  $A$ , שלנו לדעת איך להתמודד עם דגימות חוזרות. חלק מآل' הלמידה שראינו לא אהבים דגימות חוזרות, מכיוון שהן גורמות לשגיאות נומריות. (לדוגמא גרסיה ליניארית ולוגיסטי). לחלקם الآخر, כלל לא אכפת. (עצי החלטה, NN-k).

### 6.5.2 הורצת השונות

ראינו כי שיטת ועידה עם הכרעת הרוב מוריידה את השונות, אך עד לנוקודה מסויימת, הנקבעת על פי ערך הקוראלציה בין משתפי הוועדה. לכן, נוכל לצפות כי שיטת ה- Bagging תקטין את השונות בכך ש-  $T$  יגדל, ובכך תפחית את שגיאת ההכללה - אך שוב עד לנוקודה מסויימת. הנקבעת ע"פ ערך הקוראלציה בין כללי ההחלטה שנוצרים.

## 6.6 די-קוראלציה + Random Forest

הסקנו כי ניתן לשפר את שיטת ה- Bagging אם נמצא דרך למער את הקוראלציה. דרך אחת לבצע זאת היא ע"י "איזיקת ידים של אלגוריתמי הלמידה". אם נעשה זאת בצורה אקראית ומעודנת נוכל לקוות שהביצועים יעלו כתוצאה מהדי-קוראליזציה יותר מאשר ירדו ע"י איזיקת ידים של כל אלג' למיניהם. הדוגמא הטובה ביותר לזה היא Random Forests.

### 6.6.1 עצי החלטה

נזכיר במסוגי עצי ההחלטה מעלה  $\mathbb{R}^d = \mathcal{X}$ . בידנו סט אימון בעל  $m$  דגימות. מסוג העיר הרנדומלי מושג ע"י שימוש בשיטת ה- Bagging על מסווגי חצי החלטה, **עם טויסט עיקרי עבור די-קוראליזציה**: האלגוריתם מחזק פרמטר  $d \leq k$ , כאשר מגדלים כל עץ ההחלטה, עברו כל פיצול שעשו האלגוריתם, נגריל  $k$  קורדינטות מתוך  $d$  הקורדינטות הקיימות באופן אחיד ואקראי, ונשר פיצול רק לאורך  $k$  קורדינאות אלה.

### 6.6.2 הגדרה פורמללית

- $R \in \mathbb{N}$ . עומק העץ המקסימלי של כל עץ.
- $m_{min}$ . המספר המינימלי של דגימות בכל עלה של כל עץ.
- $N \in \mathbb{N}$ . מס' דגימות ה- Bagging (= מס' העצים בעיר).
- $N \in \mathbb{N}$ . מס' הקורדינטות לאורכן נרצה לבצע פיצולים.
- משתנה נוסף שאחראי על **גיוזם**, אותו נלמד בשבוע הבא.

---

**Algorithm 4** Random Forests

```

1: procedure RANDOM-FOREST(training set  $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ , maximal tree depth  $R \in \mathbb{N}$ , minimal
   training samples in any leaf  $m_{min}$ , number of trees  $T$ , number of coordinates to choose from in
   each split  $k$ )
2:   for  $t = 1, \dots, T$  do
3:     Draw a Bootstrap sample  $S^{st}$  from  $S$ 
4:     Train a decision tree  $h_{S^{st}}$  on the sample  $S^{st}$  where at each split
5:       if Not reached maximal depth  $R$  or minimal number of samples  $m_{min}$  then
6:         Select uniformly at random  $k$  features from  $[d]$ 
7:         Choose the best feature to split by from the st of  $k$  features
8:         Split based on selected feature and threshold
9:       end if
10:      end for
11:      return  $h_S(\mathbf{x}) = sign\left(\sum_{t=1}^T w_t h_t(\mathbf{x})\right)$ 
12: end procedure
```

---

### 6.6.3 תובנות

- **האם שיטת ה- Bagging יכולה פגוע בנו?**

כן. אם אלג' הלמידה הבסיסי שלנו הוא חלש, ביצוע bagging רק יחליש אותו עוד יותר. ("ועדה של טיפשים גרועה יותר מטיפש אחד").
- **מהן החסרונות של Bagging ?**
  - עליינו לאמן ולאחסן  $T$  מודלים ולא אחד.
  - עבור כל חיזוי עליינו לשאול  $T$  מודלים ולא אחד.
  - איבוד יכולת הפירוש שככל כך מובהקת לעצם החלטה.
- **הסתברות לשיווך?**

רצוי לא להשתמש בחלק היחס של 1+ בעודה כשערון ההסתברות לסיווג במחלקה.
- **IMPLEMENTATION**

כל חבר ועדה מאמין באופן בלתי תלוי ולכון ניתן לבצע IMPLEMENTATION במקביל.

### 6.6.4 סיכום

- מסיווג Random Forest הוא מסיווג מאוד פופולרי.
- מכיוון שהפרמטר  $T$  לא גורם ל- overfit, לרוב נססה להימנע מערכי  $T$  גדולים לשםיעילות.
- על מנת לבחור  $k$ , כלל האצבע הוא בדרך כלל  $\sqrt{d} = k$ , כאשר  $d$  הוא המימד של  $\mathcal{X}$ .

## 6.7 שיטת Boosting

ניקח אלג' למיניה  $A$  חלש, אלג' שהוא טוב יותר מבחירה אקראית כלשהי אך עדין לא בעל דיוק טוב כל כך, ונניח לו בסיס – נשימוש בשיטת ועידה חכמה על מנת לקבל אלג' למיניה בסיסי  $A$  וסט דגימות  $S$ . הקסם העיקרי והמרכזי כאן הוא רעיון שונה לגמרי ליצירת ועידה של כלל החלטה מאלג' למיניה בסיסי  $A$  וסט דגימות  $S$ . בעוד בשיטת ה- Bagging "העמדנו פנים" כאילו יש לנו  $S_1, S_2, \dots, S_T$  סטי דגימות מההתפלגות  $D$ , וכל חבר ועדה יוכל להתאמן על סט אחר, במקרה של Boosting אנו "נעמיד פנים" שיש לנו מספר התפלגות בסיס  $D$  שמהן דאטא האימון שלנו נדגם.

באופן יותר ספציפי, בשיטה זו כל חבר ועדה  $h_t$  הוא תוצאה הרצת האlg'  $A$  על דאטא אימון  $S_t$  שמחקה סט דגימות בגודל  $m$  הנדגם באופן אחיד ובבלתי תלוי מההתפלגות כלשהי  $D^t$ . בנויגוד לשיטת ה- Bagging, כאן כל חבר ועדה מאמין **אחד לאחר השני** וכל אחד מהווה במובן מסוים שיפור של הקודם לו. הרעיון החכם מאחורי שיטתה זו הוא שלאחר שנסיים לאמן את חבר הוועדה  $h_t$ , על סמך ההתפלגות  $D^t$  נעדכן את ההתפלגות בצורה כזו שתגדיל את ההתפלגות על דגימות הדאטא עליהם  $h_t$  טעה. בדומה זו, הכלל  $h_{t+1}$  יעשה מאמץ גדול על מנת לא לטעות בדגימות אלה, וכך הלאה.

### 6.7.1 דגימות ממושקלות

נרצה להבין מה המשמעות של "הרצת האלג'  $\mathcal{A}$ " על DATA אימון  $S_t$  שמחקה סט דגימות בגודל  $m$  הנדגם באופן אחיד ובلتוי מהתפלגות כלשהי  $D^t$ .

1. דרך אחת לפרש זאת, היא לחתת **דגימת בוטסטרהפ ממושקלת** מסט הדגימות  $S$ ,  
כאשר ההסתברות לבחור דגימה  $(x, y) \in S$  היא לפי המשקל  $D^t(x, y)$ .

2. דרך פשוטה יותר היא במידה והאלג'  $\mathcal{A}$  משתמש בכלל ה- ERM, לדוגמה עבור סטנדרטי,  
כלומר מנסה למצער את הסיכון האמפירי הממושקל:

$$L_S(h) = \sum_{i=1}^m \mathbb{1}_{(y_i \neq h(x_i))}$$

אפשר להשתמש ב-  $S$  עצמו, ולשנות את אלג' הלמידה הבסיסי כך שיימיצער את **הסיכון האמפירי הממושקל**:

$$L_{S,D^t}(h) = \sum_{i=1}^m D_i^t \cdot \mathbb{1}_{(y_i \neq h(x_i))}$$

כאשר לכל  $D_i^t = 1$  נכתוב  $(x_i, y_i) \in S$  כך ש-  $D_i^t = D^t(x_i, y_i)$

לרוב נעדיף לבחור בשיטה השנייה על פני הראשונה מכיוון שהיא חסיבה באופן יותר יעיל,  
ולא דורשת שימוש בדgesים חוזרים. מנגד, השיטה הראשונה זמינה תמיד, בעוד שהשנייה לא.

### 6.8 שיטת Adaboost

זהו מטא-אלגוריתם אחד מיני רבים של boosting.  
מתאפיין באתחול של התפלגות אחת בלבד על סט הדגימות  $S$ , כך ש-  
משם נעדכן את ההתפלגות שלנו בצורה אקספוננציאלית בכל איטרציה:

$$D_i^{t+1} \leftarrow \frac{D_i^t \cdot \exp(-w_t y_i h_t(x_i))}{\sum_{j=1}^m (D_j^t \cdot \exp(-w_t y_j h_t(x_j)))}$$

עבור אקספוננט  $w_t > 0$ . נבחן כי במידה ונקודה  $i$  מסווגת נכון, הביטוי  $y_i \cdot h_t(x_i) = 1$  והמשקל יורדת באיטרציה הבעה.  
באופן הפוך, במידה ונקודה  $i$  מסווגת לא נכון, הביטוי  $y_i \cdot h_t(x_i) = -1$  והמשקל עולה באיטרציה הבאה.

אנו נבחר את האקספוננט  $w_t$  בצורה כזו כך שנקבל:

$$\sum_{i=1}^m D_i^{t+1} \cdot \mathbb{1}_{(y_i \neq h_t(x_i))} = \frac{1}{2}$$

כל חבר ועדה  $h_t$  מצביע עם משקלות  $w_t$ , כך שהtaggit שנחזית על ידי הועדה היא:

$$h_{\text{boost}}(x) = \text{sign} \left( \sum_{t=1}^T w_t h_t(x) \right)$$

הרעיון פשוט: במעבר מהאיטרציה ה- $t$  לאיטרציה ה- $t+1$  נרצה **להגדיל** את משקל הדגימות שסווגו לא נכון ע"י  $h_t$ , ולהקטין את המשקל של הדגימות נכון ע"י  $h_t$ . נרצה להפוך את בעיית הסיווג ל"קשה ביותר האפשרית" במובן זה שהסיכון האמפירי הממושקל של  $h_t$  ביחס למשקולות של  $D^{t+1}$  הוא הגרוע ביותר  $-\frac{1}{2}$ . לבסוף, ממצאים כלליים להחלטה בועדה ע"י המשקולות  $w_t$ .

### 6.8.1 טענה - האקספוננט הדורש

נטען כי האקספוננט לו אנו זוקקים הוא בדיק  $\frac{1}{2} \log \left( \frac{1}{\varepsilon_t} - 1 \right)$  הסיכון האמפירי הממושקל של  $h_t$ .

הוכחה

$$\begin{aligned} \sum_{i=1}^m D_i^{t+1} \cdot \mathbb{1}_{(y_i \neq h_t(x_i))} &= \frac{\sum_{i=1}^m D_i^t \cdot e^{-w_t y_i h_t(x_i)} \cdot \mathbb{1}_{(y_i \neq h_t(x_i))}}{\sum_{j=1}^m D_j^t \cdot e^{-w_t y_j h_t(x_j)}} \\ &\stackrel{i}{=} \frac{\varepsilon_t \cdot e^{w_t}}{\sum_{j=1}^m D_j^t \cdot e^{-w_t y_j h_t(x_j)}} \stackrel{ii}{=} \frac{\varepsilon_t \cdot e^{w_t}}{\varepsilon_t \cdot e^{w_t} + e^{-w_t} (1 - \varepsilon_t)} \\ &\stackrel{iii}{=} \frac{\varepsilon_t}{\varepsilon_t + e^{-2w_t} (1 - \varepsilon_t)} \stackrel{iv}{=} \frac{\varepsilon_t}{\varepsilon_t + e^{-\log(\frac{1}{\varepsilon_t} - 1)} (1 - \varepsilon_t)} \\ &= \frac{\varepsilon_t}{\varepsilon_t + \frac{1}{\frac{1}{\varepsilon_t} - 1} \cdot (1 - \varepsilon_t)} = \frac{\varepsilon_t}{\varepsilon_t + \frac{\varepsilon_t}{1 - \varepsilon_t} \cdot (1 - \varepsilon_t)} = \frac{1}{2} \end{aligned}$$

(i) האינדיקטור במונה מאפס כל דוגמה  $y_i = h_t(x_i) = -1$  ולכן נקבל  $e^{w_t}$ . לכן עבור  $i$  שנשארו  $y_i \neq h_t(x_i)$  ונפעיל כmo (ii) נפרק את הסכום לדגימות  $y_i = h_t(x_i)$  ו- (iii) נחלק מונה ומכנה ב-  $e^{w_t}$  לפי הטענה

## 6.8.2 האלגוריתם

**Algorithm 6 Adaptive Boosting**


---

```

1: procedure ADABOOST(training set  $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ , base learner  $\mathcal{A}$ , number of rounds  $T$ )
2:   Set initial distribution to be uniform:  $\mathcal{D}^{(1)} \leftarrow (\frac{1}{m}, \dots, \frac{1}{m})$   $\triangleright$  Initialize parameters
3:   for  $t = 1, \dots, T$  do
4:     Invoke base learner  $h_t = \mathcal{A}(\mathcal{D}^{(t)}, S)$ 
5:     Compute  $\varepsilon_t = \sum \mathcal{D}^{(t)} \mathbb{1}[y_i \neq h_t(\mathbf{x}_i)]$ 
6:     Set  $w_t = \frac{1}{2} \ln \left( \frac{1-\varepsilon_t}{\varepsilon_t} \right) = \frac{1}{2} \ln \left( \frac{1}{\varepsilon_t} - 1 \right)$ .
7:     Update sample weights  $\mathcal{D}_i^{(t+1)} = \mathcal{D}_i^{(t)} \exp(-y_i \cdot w_t h_t(\mathbf{x}_i))$ ,  $i = 1, \dots, m$ 
8:     Normalize weights  $\mathcal{D}_i^{(t+1)} = \frac{\mathcal{D}_i^{(t+1)}}{\sum_j \mathcal{D}_j^{(t+1)}} \quad i = 1, \dots, m$ 
9:   end for
10:  return  $h_S(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^T w_i h_i(\mathbf{x})\right)$ 
11: end procedure

```

---

## 6.9 שיטת Boosting מנק' מבט של מודל PAC

## 6.9.1 הגדרות

אלג' למידה  $\mathcal{A}$  הוא **אלג' למידה  $\gamma$ -חלש** עבור מחלוקת היפותזות  $\mathcal{H}$  אם קיימות פונקציה  $\mathbb{N} \rightarrow \{(0, 1) : \mathcal{H}\}$  כך שלכל  $1 < \delta < 0$ , לכל התפלגות  $\mathcal{D}$  מעל מרחב הדגימות  $\mathcal{X}$ , ולכל פונקציית תיווג  $\{\pm 1\} \rightarrow \mathcal{X}$ , אם המקרה הריאליizable מתקיים עבור  $f, \mathcal{D}, \mathcal{H}$ , אז כאשר נריצ' את האlg'  $\mathcal{A}$  על דאטא אימון בגודל  $(1 - \delta)^{m_{\mathcal{H}}} \geq m_{\mathcal{H}}$  דגימות  $\mathcal{D} \stackrel{\text{i.i.d.}}{\sim}$  המתויגות ע"י  $f$ , האלגוריתם יחזיר היפותזה  $\mathcal{A}(S) = h_S$  כך שהסתברות של לפחות  $\delta - 1$  מתקיים  $\gamma$   $L_{\mathcal{D}, f}(h_S) \leq \frac{1}{2} - \gamma$ .

חלוקת היפותזות  $\mathcal{H}$  היא **למידה  $\gamma$ -חלש** אם קיים אלגוריתם למידה  $\gamma$ -חלש עבורה.

## 6.9.2 המוטיבציה ל-Boosting

נניח כי מחלוקת  $\mathcal{H}$  היא למידה PAC. אזי ERM $_{\mathcal{H}}$  יכול ללמידה עם מס' דגימות כמעט מינימלי  $m_{\mathcal{H}}$ . אך מה עם ERM $_{\mathcal{H}}$  קשה חישובית? נניח כי נוכל למצוות מחלוקת היפותזות **פשוות** (חלוקת בסיס)  $\mathcal{H}_{\text{base}}$ , כך ש- ERM $_{\mathcal{H}_{\text{base}}}$  ניתן לחישוב ביעילות, ומהוות אלג' למידה  $\gamma$ -חלש עבור  $\mathcal{H}$  עבור  $\gamma$  קלשה. משמעות הדבר היא שמצאנו דרך עיליה חישובית ללמידה עם דיוק  $\gamma - \frac{1}{2}$  עבור  $\gamma$  קלשה. האם נוכל לתת بواسטן ERM $_{\mathcal{H}_{\text{base}}}$  בדרכ' חישוב ביעילות, ולקבל אלג' למידה  $\mathcal{A}$  יעיל שקרוב למוצע ERM מעל  $\mathcal{H}$ ?

## 6.9.3 משפט

יהי  $S$  סט דגימות אימון. נניח כי בכל איטרציה של אלג' הלמידה הבסיסי  $\mathcal{A}$  מוחזר כלל החלטה  $h_t$  עבורו המשקל האמפירי מקיים:

$$\sum_{i=1}^m D_i^{t+1} \cdot \mathbb{1}_{(y_i \neq h_t(x_i))} \leq \frac{1}{2} - \gamma$$

אזי המשקל האמפירי הרגיל של כלל החלטה אותו ייצור  $h_{\text{boost}}$ , Adaboost, מקיים:

$$L_S(h_{\text{boost}}) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{(y_i \neq h_{\text{boost}}(x_i))} \leq e^{-2\gamma^2 T}$$

#### 6.9.4 Bias-varaince tradeoff

נרצה לקוות שאין לא מגדילים את  $T$  עד כדי overfitting, על מנת שמצוור הסיכון האמפירי יגרור שגיאת הכללה נמוכה. נניח ואני מרכיבים  $T$  איטרציות של Adaboost על אלג' למידה  $\mathcal{A}_{\text{base}}$  המחויר היפותיה מהמחלקה  $\mathcal{H}_{\text{base}}$ . מה היא מחלוקת היפותיות שיש לנו כתוצאה בפועל, וכמה גדולה היא? בדיק -

$$\mathcal{H}_T = \left\{ x \mapsto \sum_{t=1}^T w_t h_t(x) \mid w_t \in [0, \infty), \sum_t w_t = 1, h_t \in \mathcal{H}_{\text{base}} \right\}$$

כלומר מחלוקת הצירופים הקומפקטיבים של היפותיות מהמחלקה  $\mathcal{H}_{\text{base}}$ .

נבחן כי  $\mathcal{H}_T$  גדלה ככל שהפרמטר  $T$  גדל, אך היא לא גדלה במהירות יחד עם  $T$ . מכך נסיק כי boosting יגדיל את השונות, ככל ש-  $T$  גדל, אך לא בהרבה.

לעומת זאת, ה- boosting מוריד את הטעיה - נובע מכך שהסיכון האמפירי קטן ככל ש-  $T$  גדל.

למעשה הסיכון האמפירי קטן בצורה אקספוננציאלית עם  $T$  ועל כן קטנה בצורה מהירה.

לסיכום, ניתן לבדוק אם boosting מוריד את הטעיה הרבה מהר מעליית השונות,

זהו לפחות הוא משפר לרוב את שגיאות הכללה בצורה דרמטית.

### 6.10 השוואת Bagging & Boosting

Bagging	Boosting	
באופן מקביל	אחד אחרי השני	למידת חבריו ועידה
דגימות אימון	בוטסטראף ממושקל או $S$ המקורי	דגימות אימון
די-קוראלציה	לא נחוץ	
$T$ גדול מידי	עלול לגרום לכך	
שיעור	מוריד הטעיה	
עם עצי החלטה נשתמש ב-	גדם	
מימוש במקביל	לא	
הצבעת הוועדה	ממושקלת	

## 7 הרצאה 7 - רגולרייזציה ובחירה מודל

### 7.1 רגולרייזציה

עקרון המאפשר לקחת מודל ולהפוך אותו למשפחה רציפה של מודלים  $A_\lambda$  עם אותה מחלוקת היפוטזות. נחשוב על  $\mathcal{H}$ , א' כלליים. נניח כי קיים אלג' למידה  $A_0$  הבוחר  $\mathcal{H} \in h_S \in \mathcal{F}_S(h)$  לפי מיעור פונקציה ( $\mathcal{F}_S$  עבור  $\mathcal{H}$ ) משמעות הדבר היא ( $S$  נתן ע'':

$$h_S = \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{F}_S(h)$$

הfonקציה  $\mathcal{F}$  מודדת כמה טוב  $h$  מתאימה את DATA האימון. נקרא ל-  $\mathcal{F}(h)$  Fidelity Term.

ברגרסיה ליניארית, הפונקציה  $\mathcal{F}_S(h)$  מדדה את סכום הריבועים. ברגרסיה לוגיסטי,  $\mathcal{F}_S(h)$  – היה הנראות עבור מודל הרגרסיה הלוגיסטי.

ב-SVM,  $\mathcal{F}_S(h)$  – היה ה- Margin. \* המינוסים מופיעים במקומות בהם רצינו למקם את הפונקציה.

#### 7.1.1 הוספת פרמטר רגולרייזציה

אם מחלוקת היפוטזות  $\mathcal{H}$  גדולה מדי, אנו נחשוש שמיוער  $\mathcal{F}_S$  מעל  $\mathcal{H}$  עלול להוביל לאובר-פייטינג ולא עבור הכללה. דרך אחת לפתור בעיה זו היא ע' הגבלת  $\mathcal{H}$ . דרך אלגנטית יותר היא להשאיר את  $\mathcal{H}$  כמו שהיא, ולשנות את אלג' הלמידה  $A_0$ . נעשה זאת ע' שינוי בעיית האופטימיזציה ש-  $A_0$  משתמש בה לבחירת  $h_S$ . נבחר  $0 \leq \lambda$  ונגידיר אלג' למידה  $A_\lambda : S \mapsto h_S$  ע'':

$$h_S = \operatorname{argmin}_{h \in \mathcal{H}} (\mathcal{F}_S(h) + \lambda \mathcal{R}(h))$$

כאשר  $\mathcal{R}(h)$  נקרא איבר הרגולרייזציה. תפקידו למדוד את "סיבוכיות" היפוטזה  $h$ . ככל ש-  $h$  תהא יותר מסובכת  $\mathcal{R}(h)$  יהיה גדול יותר. כעת בעת המיעור, קיים trade-off:

- מצד אחד, ככל ש-  $h$  יותר מסובכת, היא יכולה לתאר באופן טוב יותר את DATA האימון  $S$ ,
- ו-  $\mathcal{F}_S(h)$  יהיה קטן יותר. מצד שני, ככל ש-  $h$  יותר מסובכת, כך יהיה  $\mathcal{R}(h)$  גדול יותר.

#### 7.1.2 λ הוא פרמטר ה- Trade-off

כמובן שעבור  $\lambda = 0$  אין רגולרייזציה, ונקבל מיוער עבור  $\mathcal{F}_S(h)$ .

כאשר  $\lambda = \infty$ , בעיית המינימיזציה לא מתייחסת ל-  $\mathcal{F}$ , ואנו מקבלים את היפוטזה  $h \in \mathcal{H}$  הפשוטה ביותר.

כל ערך אחר  $(\lambda \in (0, \infty))$  מגידר trade-off כלשהו בין הצורך ב-  $\mathcal{F}_S(h)$  קטן לבין הצורך בהיפוטזה פשוטה ( $\mathcal{R}(h)$  קטן). מכאן שהפרמטר  $\lambda$  יוצר לנו משפחה של אלג' למידה  $\{A_\lambda\}_{\lambda \in [0, \infty)}$ .

פרמטר זה גם שולט ב- B-V trade-off: כאשר  $\lambda = 0$  נקבל הטיה נמוכה ושותנות גבוהה - היפוטזה מסובכת, ועבור  $\lambda = \infty$  נקבל הטיה גבוהה ושותנות נמוכה - היפוטזה פשוטה.

### 7.1.3 המשך הדרך

לאורך רוב הרצאה זו, נתמקד בבעיות רגרסיה כאשר  $\mathcal{X} = \mathbb{R}^d$  ו-  $\mathcal{Y} = \mathbb{R}$ .  
 הסיכון האמפירי במקרה שלנו (בעיות רגרסיה) יחושב ע"י פונק' ההפסד squared loss -  $\ell(h(x), y) = (h(x) - y)^2$ . מכאן קיבל כי:

$$L_S(h) = \sum_{i=1}^m (h(x_i) - y_i)^2$$

## 7.2 עצים מסווגים

ניבור כעת לדבר על עצים רגרסיביים.  
 נניח חלוקה  $\bigcup_{j=1}^N B_j = \mathbb{R}^d$  ל-  $N$  קופסאות מקבילות לצירים. נניח תגיות  $\mathbb{R} \in c_j \in [N]$  לכל  $j \in [N]$  כך שהtagית  $c_j$  שייכת לקופסה  $B_j$ . ע"י הרגרסיה  $h \in \mathcal{H}_{\text{RT}} : \mathbb{R}^d \rightarrow \mathbb{R}$  המוגדרת ע"י:

$$h(\mathbf{x}) = \sum_{j=1}^N c_j \mathbb{1}_{B_j}(\mathbf{x})$$

### 7.2.1 גידול העץ

בהתאם סט דוגמאות אימון  $S$ , נרצה למצוא את העץ  $h \in \mathcal{H}_{\text{RT}}$  המזער את הסיכון האמפירי. גם כאן כמו בעיות סיווג בעיה זו היא NP קשה, ועל כן נשתמש ביריסטיקט CART. העיקרונו כאן דומה מאד למה שלמדנו בעבר בעיות סיווג, פרט לכך שהכרעות הרוב מוחלפות בממוצע התגיות. מכאן שבහינתן עצ חלוקה, הסיכון האמפירי ימוזער ע"י התגיות  $c_j = \text{avg}(y_i \mid y_i \in B_j)$ . היריסטיקה מתחילה לגדל עצ רגרסיה עם  $\mathbb{R}^d = B_0$  ובאופן רקורסיבי מפצלת כל קופסה בנקודה הטובה ביותר, ונונתנת לה תגית ע"י חישוב ממוצע הנקודות שבתוכה אותה קופסה. ממשיך לפצל קופסאות עד אחד משני התאים הבאים:  
 (1) הגיעו לעומק העץ המקסימלי אותו הגדרנו    (2) הגיעו למס' נקודות מינימלי בקופה.

### 7.2.2 גיזום העץ

לאחר גידול העץ, יתכן ונגלה כי בצענו יותר מדי פיצולים. אומנם כל פיצול מקטין לנו את הטעיה, אך גם מגדיל לנו את השונות. על כן השלב האחרון של CART הינו גיזום העץ - הקטנת גודלו של העץ ע"י איחוד של קופסאות אותן יצרנו בתהליך הגידול, על מנת למצוא יחס V-B טוב יותר שיבילל להכללה טובה יותר.

נניח סט אימון  $S$  ועכ' רגרסיה  $T$  המושירה מהחלוקת  $\mathbb{R}^d = \bigcup_{j=1}^N B_j$ . הסיכון האמפירי של  $T$  על  $S$  מוגדר ע"י:

$$L_S(T) = \sum_{j=1}^N \sum_{i: \mathbf{x}_i \in B_j} (y_i - \hat{y}_S(B_j))^2$$

זהו ה-  $\mathcal{F}_S(T)$  Fidelity Term. יהי  $T_0$  עכ' המתkeletal בסיום תהליכי הגידול של CART. כתוב  $T \subseteq T_0$  אם  $T$  הוא תת עכ' של  $T_0$  - כלומר אם מתkeletal מ-  $T_0$  ע"י איחוד קופסאות. פרמטר הרגולרייזציה שלנו במקרה זה יהיה פשוט  $|T|$ , כאשר  $|T| = |\mathcal{R}(T)|$ , כאשר  $\mathcal{R}(T)$  הוא מס' הקופסאות (העלים) של העכ'  $T$ . זה מגד' טוב לסבירות של היפוטזה בחלוקת  $\mathcal{H}_{\text{RT}}$ , שכן יותר עלים גורר עכ' מסובך יותר. כתוב הבעה שלנו היא מהצורה:

$$\min_{T \subseteq T_0} (L_S(T) + \lambda \cdot |T|)$$

כלומר מבין כל תת-העכ'ים של  $T_0$ , נחפש את האופטימלי המאזן את הסיכון האמפירי וסבירות ההיפוטזה.

### 7.3 בחזרה לרגרסיה

נעבור כעת לשיטות רגרסיה מודרניות המתבססות על מחלוקת היפוטזות ליניארית - עם רגולרייזציה. נזכיר במחלוקת היפוטזות של פונק' ליניארית עבור רגרסיה:

$$\mathcal{H}_{\text{lin}} = \left\{ h \mid h(x_1, \dots, x_d) = w_0 + \sum_{i=1}^d x_i w_i, \quad w_0, \dots, w_d \in \mathbb{R} \right\}$$

כאשר דיברנו על רגרסיה ליניארית, הנחנו  $m$  דגימות ו-  $d$  פיצרים, כך ש-  $d \geq m$ . ראיינו כי נעדיף  $m \ll d$  מכיוון שאם  $d \sim m$  השונות של היפוטזה הליניארית שנמצא ע"י ERM יכולה להיות גדולה. במידה מודרנית בימנו, מס' הפיצרים  $d$  יכול להיות גדול מאוד, מכיוון שהניה קל מאוד לאוסף פיצרים. ובעית רגרסיה טיפוסית לרוב בעלות  $m \sim d$  או אפילו  $m \gg d$ . במקרים כאלה, יש לנו פיצרים בעלי קוראלציה. המשקלות הנבחרות ע"י הרגרסיה יקבעו בצורה גרוועה, לדוגמה משקלות חיובית גדולה יכול להתבטל ע"י משקלות שליליות גדולות של פיצ'ר כמעט זהה. מכאן שהרגרסיה שראינו לא יכולה לעבוד עבור  $m > d$ .

#### 7.3.1 רגרסית Best Subset

נזכיר כי  $h \in \mathcal{H}_{\text{lin}}$  מתקשרות לוקטור משקלות ייחודי  $w = (w_1, \dots, w_d)^T \in \mathbb{R}^d$ , כאשר  $w_0 \in \mathbb{R}$  עם אינטראספט. ב מקרה שלנו כעת לא נכנס את האינטראספט  $w_0$  לתוך הוקטור  $w$ . כעת קיבל מטריצה  $X_{d \times m}$  ווקטור  $y$  ונוכל לכתוב:

$$L_S(w) = \|w_0 \mathbf{1} + Xw - y\|^2$$

סיכון האמפירי שלנו, כאשר  $\mathbf{1}$  הוא וקטור אחדות. הפתרון האולטימטיבי לביעות רגרסיה ליניארית עבור ערך  $d$  גדול ידוע כ- best subset selection

minimize $w \in \mathbb{R}^d, w_0 \in \mathbb{R}$	$\ w_0 \mathbf{1} + Xw - y\ ^2$
subject to	$\ w\ _0 \leq t$

עבור "נורמת" האפס המוגדרת ע"י  $\|w\|_0 = \#\{i \mid w_i \neq 0\}$ .

על פניו, נראה מושלם - נגיד  $t$  וنمצא מבין כל תת-הקבוצות של  $t$  פיצ'רים מותוק  $d$  פיצ'רים את הקבוצה עם הסיכון האמפירי הנמוך ביותר. החזרות הרעות הן שבעה זו היא NP קשה, שכן  $\|w\|_0$  אינה קמורה, והחישוב לעיל נראה כמו חיפוש קומבינטורי על פני כל תת-הקבוצות.

נרצה להחליף את התנאי על וקטור המשקולות כך שעדין ימודד את סיבוכיותו אך יהיה ניתן לחישובעיל.

נגיד כעת את הבעיה מחדש מחדש עבור נורמה  $\|\cdot\|$  על  $\mathbb{R}^d$ :

$$\begin{array}{ll} \underset{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}}{\text{minimize}} & \|w_0 \mathbf{1} + X\mathbf{w} - \mathbf{y}\|^2 \\ \text{subject to} & \|\mathbf{w}\| \leq t \end{array}$$

ולמען הפשטות נמירה בעיה שcolaה ללא תנאים ע"י הגדרת פרמטר רגולרייזציה:

$$\underset{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}}{\text{argmin}} (L_S(w_0, \mathbf{w}) + \lambda \|\mathbf{w}\|)$$

### 7.3.2 מודלים חדשים

נראה כעת שלוש שיטות גרסיה שכולן בוחרות היפותזה ליניארית ב-  $\mathcal{H}_{\text{RT}}$  בעזרת  $\mathcal{F}_S(h)$  fidelity term (סכום ריבועים):

$$L_S(w_0, \mathbf{w}) = \sum_{i=1}^m (w_0 + \langle \mathbf{w} | \mathbf{x}_i \rangle - y_i)^2$$

ואיבר הרגולרייזציה  $\mathcal{R}(h)$  ימדוד את סיבוכיות ההיפותזה הליניארית  $\mathbf{w}$ .

#### o אלג' Ridge Regression \ $\ell_2$ -regularized linear regression

$$\underset{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}}{\text{argmin}} \|w_0 \mathbf{1} + X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_2^2 \quad \left| \quad \|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^d v_i^2} \right.$$

#### o אלג' Lasso Regression \ $\ell_1$ -regularized linear regression

$$\underset{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}}{\text{argmin}} \|w_0 \mathbf{1} + X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_1 \quad \left| \quad \|\mathbf{v}\|_1 = \sum_{i=1}^d |v_i| \right.$$

#### o אלג' Best Subset Regression

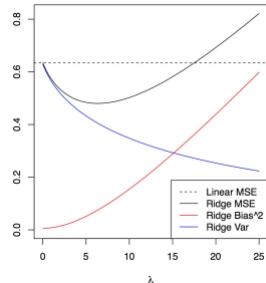
$$\underset{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}}{\text{argmin}} \|w_0 \mathbf{1} + X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|w\|_0 \quad \left| \quad \|w\|_0 = \#\{i \mid w_i \neq 0\} \right.$$

נבחן כי  $\lambda$  גורר גרסיה ליניארית רגילה.  $\lambda = 0 \rightarrow \hat{\mathbf{w}}$  והgresיה מותאמת רק את האינטרספט.

## 7.4 אלג' Ridge Regression

$$\underset{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}}{\operatorname{argmin}} \|w_0 \mathbf{1} + X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_2^2$$

נבחן כי זהה בעיית אופטימיזציה קמורה - QP. נסמן ב-  $\hat{\mathbf{w}}_{\lambda}^{\text{ridge}}$  את המינימום בעיית ה- ridge עבור  $0 = \lambda$  נקבל פתרון בעיית רגרסיה ליניארית רגילה, ועבור  $\infty \rightarrow \lambda$  נקבל  $0 \rightarrow \hat{\mathbf{w}}_{\lambda}^{\text{ridge}}$ . ככל ש-  $\lambda$  גדלה, ההטיה קטנה ורודה.



כיצד נמצא את הפתרון? לא נזדקק לפותר QP הפעם.

עבור אלג' זה יש לנו נוסחה סגורה למציאת המינימום.

אם נפעל בדרך דומה בה פעלו עבור רגרסיה ליניארית נמצא ש-

$$\forall i \in [d] : \frac{\partial}{\partial w_i} (\|w_0 \mathbf{1} + X\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_2^2) = 0$$

MOVED למערכת המשוואות:

$$X^T \mathbf{y} = (X^T X + \lambda I) \mathbf{w}$$

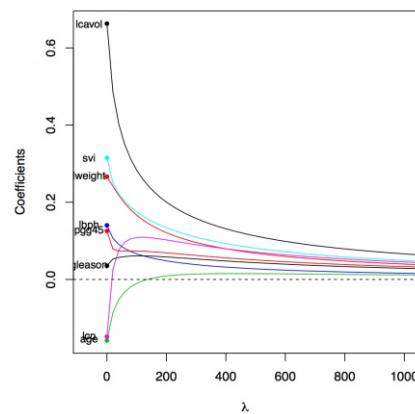
כאשר עבור  $0 = \lambda$  נקבל את המשוואות הנורמליות לרגרסיה ליניארית.

היתרון של רגולרייזציה נראה בבירור כאן - אפילו אם  $X^T X$  אינה הפיכה, ואפילו אם  $m > d$ , לפחות  $0 > \lambda$  מתקיים כי  $I + \lambda X^T X$  הפיכה. (כל ע"ע 0 של  $X^T X$  גדול וכבר אין אפס). שיטת ה- SVD לחישוב המשקלות של הרגרסיה הליניארית מוכללת גם לרגרסיה זו:

$$\hat{\mathbf{w}}_{\lambda}^{\text{ridge}} = V \Sigma_{\lambda} U^T \mathbf{y} \quad \left| \quad [\Sigma_{\lambda}]_{i,i} = \frac{\sigma_i}{\sigma_i^2 + \lambda} \right.$$

### 7.4.1 מסלול הרגולרייזציה

נרצה לעקוב אחריו כל משקלות  $w_i$  לפי שינויים בערכי  $\lambda$ . פלוט זה נקרא Regularization Path וניתן בעזרתו להבחן איך המשקלות מתחילה כמשקלות של רגרסיה ליניארית ולאט לאט יורדות בערךן.



## 7.5 אלג' Lasso Regression

ראינו כי ה- Ridge Regression היא שיטה טובה עבור רגרסיה ליניארית כאשר  $d$  גדול, כאשר  $m > d$  או כאשר  $X^\top X$  אינה הפיכה. היא גם שיטה טובה לקבלת שליטה על ה- B-V trade-off מה שהיא לא עשויה, זה לבחור סט פיצרים ספציפי בוא נשימוש עבור הרגרסיה.

$$\underset{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}}{\operatorname{argmin}} \quad \|w_0 \mathbf{1} + X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 \quad \left| \quad \|\mathbf{v}\|_1 = \sum_{j=1}^d |v_j| \right.$$

נבחן כי זהה בעית אופטימיזציה קמורה, ובפועל הינה אחת השיטות היעילות והאפקטיביות ביותר עבור רגרסיה מודרנית. אלג' זה מתנהג כמו Ridge עבור  $\lambda = 0$ ,  $\lambda \rightarrow \infty$  אך פועל הצורה שונה מאוד בערכים בין לבין. הפתרונות שלlasso לasso, וקטורי המשקלות  $\hat{\mathbf{w}}$  הם מדוילים - יש בהם מעט קורדיינאות שאינן אפס. ככל ש-  $\lambda$  גדלה, לרוב נקבע יותר אפסים ב-  $\hat{\mathbf{w}}$ , וכך הוקטור מדויל יותר ויוטר עד שבגבול  $\lambda \rightarrow \infty$  הוא וקטור אפסים.

ונכל להסתכל על ה- 'active set' - הפיצרים בעלי קורדיינאות שונות מאפס בפתרון שלlasso - זו דרך לדעת אלו פיצרים חשובים עבור הרגרסיה. ככל ש-  $\lambda$  גדלה, נקבע פחות פיצרים אקטיביים. מכאן שלlasso יתרכז **משמעותי** על פני Ridge - הוא ניתן לפירוש ובוחר עבורנו את אוסף הפיצרים הטובים ביותר.

נשאלת השאלה - מדוע אלג' הlasso המשתמש בנורמת  $\ell_1$  נותן פתרונות מדוילים בעוד Ridge המשתמש בנורמת  $\ell_2$  לא? נתנו בהרצאה שתי הסברים לכך - האחד משתמש בצדורי היחידה והשני במקרה פרטיא של מטריצת  $X$  אורתוגונלית. ראו תרגול 10 לפירוט רחב יותר.

## 7.6 אלג' רגרסיה לוגיסטיבית $\ell_1$ -regularized

נזכיר ברגression לוגיסטיבית, בה נתונים האימון שלנו מסודרת במטריצת רגרסיה  $X$  וקטור התוצאות מקיים  $\mathbf{y} \in \{0, 1\}^m$ . אם  $m \sim d$  ורגרסיה הלוגיסטיבית תסבול מאותן בעיות כמו זריסה ליניארית וגילוח. הרגרסיה הלוגיסטיבית מוצאת את וקטור המשקלות ע"י פתרת הבעה:

$$\hat{\mathbf{w}} = \underset{w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d}{\operatorname{argmax}} \sum_{i=1}^m \left( y_i (w_0 + \langle \mathbf{x}_i | \mathbf{w} \rangle) - \log (1 + e^{w_0 + \langle \mathbf{x}_i | \mathbf{w} \rangle}) \right)$$

אלג' למידה זה פשוט מוסף איבר רגולרייזציה ע"י פותר את הבעה:

$$\underset{w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \left( \underbrace{\sum_{i=1}^m \left( \log (1 + e^{w_0 + \langle \mathbf{x}_i | \mathbf{w} \rangle}) - y_i (w_0 + \langle \mathbf{x}_i | \mathbf{w} \rangle) \right)}_{-\mathcal{F}_S(\mathbf{w})} + \lambda \|\mathbf{w}\|_1 \right)$$

זהו גם בעיה קמורה וקיים עבורה פותרים ייחודיים. זהו מסווג נפלא עבור  $\mathbb{R}^d = \mathcal{X}$ , מכיוון שהוא בעל שונות נמוכה, ניתן לשולט ב- B-V trade-off ע"י בחירת  $\lambda$  והוא מאוד ניתן לפירוש.

**הערה** נבחן כי עבורנו מבועית **מקסימיזציה** לבעית **מינימיזציה** ולכן לקחנו את  $(\mathbf{w})$ .

## 7.7 בחירת מודל והערכתו

עד כה רأינו אלג' למידה עבור בעיות גראסיה וסיווג, ובנוסף רأינו מטא-אלגוריתמים. השלב הבא הוא יכולת להחליט:

- umi מהאלג' שלרשותינו כדי להשתמש בעיה נתונה כלשהי?

- כיצד נבחר את הפרמטרים הרלוונטיים לאוטו אלג'?

(מס' שכנים, פרמטר רגולרייזציה של Soft-SVM, עומק עץ מקסימלי ופרמטר גיזום ב-CART..  
גם בעת שימוש ב-Boosting \ Bagging עלינו להחליט על פרמטרים רלוונטיים..)

שלב זה נקרא **בחירה המודל**.

### 7.7.1 הערכת המודל

משימה נוספת היא להעריך את ביצועי (שגיאת הכללה) המודל שבחרנו, לפניה שמשיך איתנו לביקורת דגימות חדשות. זהו שלב חשוב מכמה סיבות:

1. אם שגיאת הכללה שלנו לא טובה, יתכן וננו עובדים עם אלג' למידה שגוי עבור הבעיה ואולי עדיף לחזור לשלב בחירת המודל ולמצוא מועד חדש.

2. בד"כ כאשר משתמשים בלמידה על מנת לפתור בעיות אמיטיות, נרצה לדעת כיצד אלג' הלמידה שלנו מתנהג לפניה שנתחיל להשתמש בו (חצבו לדוגמא על אלג' שמהליט על קבלת הלואות).

### 7.7.2 שגיאת הכללה

מה הוא בדיקת הטעויות אותה אנו מנסים לאמוד? נניח פונקציית הפסד  $\ell(\cdot, \cdot)$  כלשהי. ההגדרה שלנו לשגיאת הכללה תלולה במודל גינרוצ' הדאטא שלנו. רأינו כבר מס' הגדירות לשגיאת הכללה:

- אם אין הנחה כלשהי לגבי איך נוצר הדאטא, ההגדרה היחידה של שגיאת הכללה היא על סט בדיקה כלשהו  $T = \{(x_j, y_j)\}_{j=1}^{|T|}$  של מודל PAC, נקבע:

$$L(h) = \sum_{j=1}^{|T|} \ell(h(x_j), y_j)$$

- ידועה התפלגות על  $\mathcal{X}$ , ולא ידועה פונק' התיאוג  $\mathcal{D} \rightarrow \mathcal{X}$ . אם נניח את ההנחות של מודל PAC, נקבל:

$$L_{\mathcal{D}, f}(h) = \mathbb{E}_{x \sim \mathcal{D}} (\ell(h(x), f(x)))$$

- ידועה התפלגות על  $\mathcal{X} \times \mathcal{Y}$ . אם נניח את ההנחות של מודל Agnostic PAC נקבל:

$$L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}} (\ell(h(x), y))$$

### 7.7.3 פירוק הטיה ושונות

בביעות גרסיה, כאשר אנו משתמשים בשיטת מעור סכום הריבועים, שגיאת הכלכלה ניתנת לפירוק של סכום השונות עם הטיה בריבוע. נניח סט אימון  $S$  וסט בדיקה  $T = (\bar{x}_j, \bar{y}_j)$ . ניצור את וקטור התגיות של סט הבדיקה  $(\bar{y}_{|T|}, \dots, \bar{y}_1) = \bar{y}$ . (**תוצאות אמיתיות**). נשמש באלג' הלמידה המאומן על  $S$  על מנת לקבל את  $h_S$ ,  $\tilde{y}_j = h_S(\bar{x}_j)$ , ונכתוב  $\tilde{y} = \bar{y} - \tilde{y}_S$ . עבור החיזוי שלנו על סט הבדיקה  $T$ . אם נשמש ב-Squared Error loss נקבל שגיאת הכלכלה מהצורה  $\|\tilde{y}\|^2$ .

מכך שסט הדגימות  $S$  הינו אקראי, נקבל כי  $h_S$  אקראית גם היא ועל כן  $\tilde{y}$  הוא וקטור אקראי.

$$\begin{aligned} \mathbb{E}(\|\bar{y} - \tilde{y}\|^2) &= \mathbb{E}(\|\bar{y} - \mathbb{E}(\tilde{y}) + \mathbb{E}(\tilde{y}) - \tilde{y}\|^2) \\ &= \mathbb{E}(\|\bar{y} - \mathbb{E}(\tilde{y})\|^2) + \mathbb{E}(\|\mathbb{E}(\tilde{y}) - \tilde{y}\|^2) + \overbrace{2\mathbb{E}(\langle \bar{y} - \mathbb{E}(\tilde{y}) | \mathbb{E}(\tilde{y}) - \tilde{y} \rangle)}^{\text{מתאפס}} \\ &= \|\bar{y} - \mathbb{E}(\tilde{y})\|^2 + \text{var}(\tilde{y}) = \text{bias}(\tilde{y})^2 + \text{var}(\tilde{y}) \end{aligned}$$

ובכך פירקנו את שגיאת ה经济学家 להטייה בריבוע + השונות.

**הטייה מודדת את המרחק בין "ממוחע" \ החיזוי ה"טיפוסי" לבין האמת.**

**השונות מודדת כמה וריאלי החיזוי.**

יוטר דאטא  $\leftarrow$  שונות נמוכה. מודל מסובך יותר  $\leftarrow$  שונות גבוהה. רעש גבוה  $\leftarrow$  שונות גבוהה.

### 7.7.4 מטרות בחירת המודל

כאשר יש לנו שליטות על סיבוכיות המודל, האתגר העיקרי של בחירת מודל הוא למצוא את ה-sweet-spot שבו הטייה והשונות לא גבוהים מדי - איפה שנקלט את שגיאת ה经济学家 הנמוכה ביותר. כיצד נעשו זאת?

#### 7.7.5 שיטה ראשונה | שימוש בדאטא אימון

הגישה הנאיבית ביותר תהיה להשתמש בדאטא האימון  $S$  עבור הכל. התהליך יראה בערך ככזה. בהינתן משפחה של אלג' למידה  $\{\mathcal{A}_\alpha\}$ , בצע:

- **שלב אימון.** האמן כל אלג' על  $S$  ונקבל  $h_\alpha = \mathcal{A}_\alpha(S)$ .
- **שלב בחירת המודל.** נבחר את  $\alpha^* = \underset{\alpha}{\operatorname{argmin}} L_S(h_\alpha)$ .
- **שלב העורכת המודל.** נאמוד את שגיאת ה经济学家 של המודל הנבחר  $h_{\alpha^*}$  ע"י חישוב הסיכון האמפירי על  $S$ , קלומר נאמוד את  $L_S(h_{\alpha^*})$  ע"י  $L_D(h_{\alpha^*})$ .

שיטה זו תכשל. למודל שנבחר לא תהיה ה经济学家 טובה, והאמידה  $(h_{\alpha^*})_S$  תהיה מאוד אופטימית עבור שגיאת ה经济学家 האמיתית. למה? כי אלג' הלמידה שלנו מנסה לאמץ ולספוג כמה שיותר מדאטא האימון במסגרת האפשרות שלו ביותר  $\mathcal{H} \in \mathcal{A}$  ככל שלאלג' הלמידה שונות גבוהות יותר, כך הוא יספוג יותר תוכנות ספציפיות של דאטא האימון, ויכיליל פחות טוב.

### 7.7.6 שיטה שנייה | דאטא אינסופי

במקרה זה נרצה שלושה סטים שונים של דאטא, אחד לכל שלב. נניח כי ברשותינו סט אימון  $S$ , סט אימונות  $V$  וסט הערכה  $T$ . נפעל באופן הבא:

- **שלב אימון.** האמן כל אלג' על  $S$  ונקבל  $\mathcal{A}_\alpha(S) = h_\alpha$ .
- **שלב בחירת המודל.** נבחר את  $\alpha^* = \arg\min_{\alpha} L_V(h_\alpha)$  - ההפסד הממוצע על סט האימונות.
- **שלב הערכת המודל.** נאמוד את שגיאת ההכללה של המודל הנבחר  $h_{\alpha^*}$  ע"י חישוב הסיכון האמפירי על  $T$ , ככלומר נאמוד את  $L_T(h_{\alpha^*})$  ע"י  $L_D(h_{\alpha^*})$ .

שיטה זו תעבור, אך בפועל ב- batch learning אין לנו גישה אינסופית לדאטא, אלא רק אוסף אחד  $S$ .

### 7.7.7 שיטה שלישית | Cross Validation

נתבונן בשיטת ה- k-fold cross validation:

1. נחלק את הדאטא ל-  $k$  חלקים שווים וזרים, הנקראים folds.
2. עבור  $i = 1, \dots, k$  נבצע:
  - נאמן את המודל על כל הדאטא חוץ מה폴ד ה-  $i$ .
  - נחשב את ההפסד על הפולד ה-  $i$ , Cainilo היה סט בדיקה.
3. נחשב את התוחלת ואת סטיית התקן של  $k$  ההפסדים.

שיטת CV היא השיטה הנפוצה ביותר עבור בחירת פרמטרים. על מנת לעשות זאת, נאמן כל מועמד  $\mathcal{A}_\alpha$  בסה"כ  $k$  פעמים - כאשר כל פעם נשאיר פולד אחד בחוץ. נבחר את אלג' הלמידה  $\mathcal{A}_\alpha$  שמשמעותו השגיאה שלו היה הנמוך ביותר. **חשוב להבין** כי לאחר שנבחר מודל בעזרת שיטת ה-CV ונגידיר אותו במלואו, נאנושוב על כל סט האימון, שכן אנו רוצים לאמן את אלג' הלמידה שלנו על כמה שיותר נקודות.

שיטת CV טוביה גם עבור הערכת המודל. לאחר שהשכנו אלג' למידה סופי ומוגדר במלואו, נוכל להריץ CV ולחשב את תוחלת השגיאה ואת סטיית התקן. בכך נאמוד את שגיאת ההכללה וגם נספק הערכה כלשהי על הדיקוק של אומד זה.

#### כיצד נבחר את $k$ ?

נבחן כי עבור  $k = 1$  אין CV. עבור  $k = 2$  קיבל סט אימון של רק חצי מהدادטא, ושגיאת ה-CV תהיה גדולה.  $k$  קטן מדי - עלולים לאמן על דאטא סט קטן מדי ושגיאת ה-CV עלולה להיות מותה כלפי מעלה.  $k$  גדול מדי - כל סטי האימון כמעט זהים, ולכן שגיאת ה-CV כוללת ממוצע של משתנים בעלי קוראלציה גבוהה, דבר שעלול להוביל לשונות גבוהה.

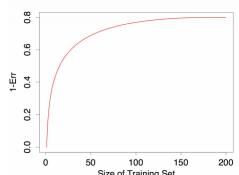
### 7.7.8 בוטסטראפ

שיטת הבוטסטראפ יכולה לבוא לידי שימוש גם עבור אמידת שגיאת ההכללה -  
עובד בחירת והערכת מודל בהתבסס על הדטא היחיד  $S$  שברשותינו.  
על מנת לאמוד את שגיאת הכללה של אלג' למידה  $A$ , נדגים  $B$  דוגמאות בוטסטראפ,  
נסמן את הדגימה ה- $b$ -ע"י  $S^{(b)}$ . נגידר את הסט ה- $b$ -ע"י  $S \setminus S^{(b)}$  -  
כלומר כל הנקודות ב- $S$  שלא נבחרו בסט הבוטסטראפ ה- $b$ .  
נאמן את  $A$  על  $S^{(b)}$  ונבחן אותו על  $T^{(b)}$ . נחשב תוחלת וסטיית תקן של שגיאות ההכללה.

### 7.7.9 2 שגיאות נפוצות

- שגיאת over-estimation. בשיטת ה-CV, אנו מאמנים את  $A$  על סט אימון קטן מ- $S$ :  
נאמנו על  $\frac{m(k-1)}{s}$  דוגמאות עבור  $m = |S|$  ו-  $k$  הוא מס' הפלדים.

אם  $k$  לא גדול, משמעות הדבר היא שהכל פעם שנאמן במהלך CV נפחית את מס' הדוגמאות האפשרי.  
ニיצרך מודל PAC שקיים מס' מינימלי של דוגמאות הנחוצות למדידת היפוטזה מחלוקת היפוטזות.  
אם  $|S|$  מספיק, ו-  $\frac{m(k-1)}{s}$  לא מספיק?



במקרה זה שיטת ה-CV תבצע over-estimation על שגיאת ההכללה.

נסיק מכאן שכדי ידעת איך מס' הדוגמאות משפייע על אלג' הלמידה שלנו - לכל אלג' עוקמה המתרארת את שגיאת ההכללה שלו כפונקציה של גודל סט האימון  $m$ . אם נוכל לאמוד עוקמה זו בדרך כלשהי, נוכל להיות חכמים יותר בעת בחירת מס' הפלדים  $k$  בהינתן סט אימון מוגדל  $m$ .

- שגיאת under-estimation. זהה בעיה נפוצה הרבה יותר. נניח ויש לנו  $m$  דוגמאות אימון ואנו לא יכולים להשיג עוד. אנו מתחילה לבדוק מודלים, כל אחד עם הפרמטרים הייחודיים לו. שלב זה נקרא snooping model.

נניח ואנו מטפלים בסט האימון - מורידים פיצרים, יוצרים פיצרים חדשים, מסננים דוגמאות בעיתיות וכו'.

שלב זה נקרא snooping data. לבסוף נמצא מודל מוגדר ו"סט דוגמאות נקי" ש"עובד טוב". CUT נרצה להשתמש ב-CV או בוטסטראפ על מנת לאמוד את שגיאת ההכללה של אלג' הלמידה שבחרנו. מה הבעיה?  
- ע"י התאמת אובייסיבית של פרמטרים ונסיון של מס' רב של אלג', אנו גורמים ל- overfit על סט האימון.  
אם נשתמש בסט אימום, לאחר שנשתמש בו מס' רב של פעמים נבעץ overfit גם עליו.  
- כאשר דטא חדש שלא נראה לפני גיע, הוא לא יקבל את הטיפול שהעבכנו את דטא האימון שלנו.  
שיטת CV עלולה להפיק over-estimates על דטא חדש.

- המנו מעיבוד מידע ידני. קודדו את כל שלב ה-preprocessing, ובכל שלב של בוטסטראפ או CV הריצו את שלב עיבוד המידע המקורי - כמו שהוא ירצו כאשר אנו חוזים דטא חדש.

- הגבילו את שלב ה-model,data snooping למתוך קבוצה קטנה של  $S$  שלא נקרא "מודכמת עם אופטימיות".  
חשוב מאד לשמר דטא לשלב מאוחר יותר בעיצוב אלג' הלמידה.

## 8 הרצאה 8 - למידה Unsupervised

### 8.1 בעיה שונה לממרי

נubby לטיפול בעיות שונות מהרגיל - בעיות בהן אין תגית  $y$  בכלל.  
הדאטה אותו קיבל ונעבד יהיה מהצורה  $\{x_i\}_{i=1}^{\infty}$  עבור  $x_i \in \mathcal{X}$ .  
בעיות למידה כאלה נקראות **unsupervised learning problems**.

### 8.2 הורדת מימד - PCA

לעתים דאטה ב-  $\mathbb{R}^d$  עבור  $d$  גבוהה רק נראה כאילו הוא מממד גבוה.

לדוגמא, תמונות של ספרות הכתובות בכתב יד. למרות שלאו תמונות  $28 \times 28$  פיקסלים המיוצגות ב-  $\mathbb{R}^{28 \times 28}$ , ישנו רק מס' בודד של צורות בהן אנשים כתובים בספרות. ניתן לראות כי ברוב התמונות יש חלק קבוע שנשאר ריק, שמייצג פיקסלים לא אינפורטטיבים עבור חיזוי התמונה. בנוסף למרות שתי תמונות של ספרה כלשהי נבדלות זו מזו, הן כנראה מייצגות את אותו הדבר. מכאן שאולי ניתן להוריד את מימד הדגימות למימד נמוך יותר עם ייצוג קומפקטי יותר.

יהי  $x_1, \dots, x_m \in \mathbb{R}^d$  הדאטה שלנו. בהורדת מימד אנו מוחפשים מיפוי  $W : \mathbb{R}^d \rightarrow \mathbb{R}^k$  עם  $k \ll d$

כך שכל דבר שנרצה לעשות עם  $x_1, \dots, x_m$  יוכל לעשות עם  $(W(x_1), \dots, W(x_m))$ .

אם  $W$  היא העתקה ליניארית - הפעולה תקרא **הורדת מימד ליניארית**. אחרת, הפעולה תקרא **הורדת מימד לא ליניארית**.

#### 8.2.1 סיבות

1. **למידה.** ראיינו כי קיימים אלג' למידה על  $\mathbb{R}^d$  שעובדים טוב יותר כאשר המימד  $d$  קטן בהשוואה לגודל סט האימון  $m$ , וכי חלק אף נכשלים עבור  $d$  גדול מדי.

2. **ווייזואлизציה.** אם נצליח להוריד את המימד  $-4 \leq k \leq 4$  נוכל לעשות זאת לדאטה על דף,  
כאשר המימד הרביעי יוצג ע"י צבע.

3. **חישוביות.** כפי שראינו סיבוכיות הזמן והמקום של הרבה אלג' למידה גדול ביחס עם  $d$ .  
ע"י הורדת המימד חלק מתהליק לעיבוד מידע מוקדים, נוכל להשתמש בפחות משאבי חישוביים.

#### 8.2.2 הורדת מימד ליניארית

יהי  $x_1, \dots, x_m \in \mathbb{R}^d$  הדאטה שלנו כך ש-  $d$  גדול, תחת הנחה כי הדאטה לא מכסה את כל המרחב  $\mathbb{R}^d$   
וכמעט כי בתוך תת-מרחב  $k$  מימדי של  $\mathbb{R}^d$ .

נרצה להטיל את הדאטה שלנו על תת המרחב זהה, ע"י העתקה  $W : \mathbb{R}^d \rightarrow \mathbb{R}^k$  כלשהי.

בעיות:

1. אנו לא יודעים מה הוא תת המרחב זהה. יש לנו רק את הדטא, ועלינו למצוא בסיס לאותו תת מרחב.
2. גם אם היינו יודעים מי הוא תת המרחב, הטלה אינה מספיקה. הנקודה המוטלת עדיין תהיה ב-  $\mathbb{R}^d$ . מה שאנו בוכלים רוצים זה למצוא בסיס אורתונורמלי לתת המרחב זהה, אז למצוא  $k$  קורדיינאות שמייצגות את הנקודה המוטלת ע"י הבסיס לתת המרחב. התוצאה תהיה  $k$  מספרים ממשיים, ואכן תייג הורדת מימד אמיתית.

### 8.2.3 האלגוריתם

אלגוריתם PCA הוא הנפוץ ביותר להורדת מימד. נבחר  $k$  כלשהו ונחפש העתקה  $U : \mathbb{R}^k \rightarrow \mathbb{R}^d$ . עם  $W$  נחפש גם את ההעתקה ההפכית  $W : \mathbb{R}^d \rightarrow \mathbb{R}^k$ . נמודד את הטעות הנגרמת כתוצאה מהתהליך ע"י סכום הריבועים:

$$\sum_{i=1}^m \|x_i - UWx_i\|^2$$

אotto נשאף למזער. מכאן שביעית ה-PCA בהינתן נקודות  $x_1, \dots, x_m$  היא למצוא:

$$W^*, U^* = \underset{W \in \mathbb{R}^{k \times d}, U \in \mathbb{R}^{d \times k}}{\operatorname{argmin}} \sum_{i=1}^m \|x_i - UWx_i\|^2$$

**למה**

יהי  $(W, U)$  פתרון לבעה לעיל. איז העמודות של  $U$  הן אורתונורמליות, ו-  $W = U^\top$ .

**הוכחה**

יהיו  $W, U$  מטריצות, ותהי העתקה  $x \mapsto UW$ . המטריצה  $UW$  היא מטריצה  $d \times d$  ממימד  $k$  ולכן בתמונה שלה הוא תת מרחב של  $\mathbb{R}^d$ , שמיומו הוא לכל היותר  $k$ . נסמן תת מרחב זה ע"י  $S = \operatorname{Im}(UW)$ . מכיוון ש-  $x \in S$ , ומתקיים כי ההטלה שמזערת את  $\|x - UWx\|$  הינה העתקה  $x \mapsto UWx$  על תת המרחב  $S$  (לפי תכונה של הטלה אורתוגונלית).

בנוסף, הנק' הכוי קרובה ל-  $x$  ב-  $S$ , כלומר ההטלה האורתוגונלית של  $x$  על  $S$  נתונה ע"י  $VV^\top x$ , עבור  $V$  שעמודותיו הן בסיס אורתונורמלי של  $S$ :

$$\forall u \in S : \|x - u\|_2 \geq \|x - VV^\top x\|_2$$

מכאן שפתרון לבעה לעיל היא  $U$  עם עמודות אורתונורמליות, ו-  $W = U^\top$ .

בהתמך על הלמה לעיל נסיק כי  $W = U^\top$  וכי עמודות המטריצה  $U$  הם אורתונורמלים. מתקיים:

$$\begin{aligned} \|x - UU^\top x\|^2 &\stackrel{i}{=} \|x\|^2 - 2x^\top UU^\top x + x^\top UU^\top UU^\top x \\ &\stackrel{ii}{=} \|x\|^2 - 2x^\top UU^\top x + x^\top UU^\top x \\ &= \|x\|^2 - x^\top UU^\top x \\ &= \|x\|^2 - (U^\top x)^\top U^\top x \\ &\stackrel{iii}{=} \|x\|^2 - \text{trace}(U^\top x (U^\top x)^\top) \\ &= \|x\|^2 - \text{trace}(U^\top x x^\top U) \end{aligned}$$

(i) ליניאריות של  $U^\top U = I$  (ii)  $\langle x - UU^\top x | x - UU^\top x \rangle$  מתקיים כי

$$\forall v, u \in \mathbb{R}^k : v^\top u = \text{trace}(uv^\top) \quad (iii)$$

ניתן להוכיח כי בעיה זו שකלה לבעה היא מהצורה:

$$U^* = \underset{U \in \mathbb{R}^{d \times k}: U^\top U = I}{\operatorname{argmax}} \text{trace}\left(U^\top \sum_{i=1}^m x_i x_i^\top U\right)$$

נבחן כי השורה האחורונה בפיתוח הראשונה מכילה את  $x$  שאינו תלוי ב-  $U$  ועל כן ניתן להוריד אותו.

#### 8.2.4 משפט

תהי  $A = \sum_{i=1}^m x_i x_i^\top$  ויהיו  $u_1, \dots, u_n$  ה-  $k$  וקטורים העצמיים הראשוניים של  $A$ . אזי הפתרון לבניית PCA הוא: המטריצה  $U$  תהיה מטריצה  $d \times k$  שעמודותיה הם  $u_1, \dots, u_k$ .

#### הוכחה

אם נסמן  $A = \sum_{i=1}^m x_i x_i^\top$  קיבל:

$$\underset{U \in \mathbb{R}^{d \times k}: U^\top U = I}{\operatorname{argmax}} \text{trace}(U^\top A U)$$

כאשר המטריצה  $A$  היא סימטרית מוגדרת חיובית. יהי  $A = VDV^\top$  הפירוק הספקטורי של  $A$ . עבורו המטריצה  $D$  היא בעלת אלכסון המכיל את הע"ע של  $A$  בסדר יורד, ועמודות  $V$  הווקטורים העצמיים המתאימים לאוותם ע"ע. לפי טענה שלא נכנס לפרטי הוכחתה, לכל מטריצה  $d \times k$  אורתונורמלית  $U$  מתקיים:

$$\text{trace}(U^\top A U) \leq \sum_{i=1}^k D_{i,i}$$

כאשר אגף ימין הוא סכום  $k$  הע"ע המוביילים. לפי טענה זו, אם ניקח  $\tilde{U} = U$ , מטריצה  $d \times k$  שעמודותיה הן  $k$  הוקטורים העצמיים המוביילים של  $A$ , אז  $U$  היא אורתונורמלית ומתקיים:

$$\text{trace} \left( \tilde{U}^\top A \tilde{U} \right) = \sum_{i=1}^k D_{i,i}$$

כאשר קיבלנו שיוויון עם החסם העליון, ועל כן מקסמנו את הביטוי, כנדרש.

### 8.2.5 מרחב אפיני

בפועל אין סיבה שתת המרחב הליינרי עבר בראשית הצירים. במקרים אחרים, נרצה לאפשר להעתקה  $W$  להיות העתקה אפינית ולא רק לייניארית. נכליל את המקרה שראינו ונתבונן ב-  $W : \mathbb{R}^d \rightarrow \mathbb{R}^k$  מהצורה:

$$W(\mathbf{x}) = \tilde{W}(\mathbf{x} - \mu)$$

כאשר  $\mu \in \mathbb{R}^d$  ו-  $\tilde{W} : \mathbb{R}^d \rightarrow \mathbb{R}^k$  העתקה לייניארית. שינוי זה מאפשר לנו "להזיז" את הדאטא לפני הפעלת העתקה  $\tilde{W}$ . כאשר מוסיפים לביעית האופטימיזציה את  $\mu$ , אופטימיזציה לערך  $\mu$  מעל בעיית PCA נתונה על ידי:

$$\mu = \frac{1}{m} \sum_{i=1}^d \mathbf{x}_i \stackrel{\text{סימוי}}{=} \bar{\mathbf{x}}$$

נבחן כי זה הוא הממוצע האמפירי של הדאטא שלנו. לאחר ביצוע השינוי, המטריצה  $A$  שלנו מוגדרת כעת ע"י:

$$A = \frac{1}{m} \sum_{i=1}^m (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

ונבחן כי זהו האומד למטריצת השוניות המשותפות אותה ראיינו בעבר. ניתן לתהות כיצד על הקשר בין PCA לבין שונות ושונות משותפת - את הקשר נראה בתרגול.

### 8.2.6 סיכום ביינימ

במידה ונרצה להוריד מימד  $d$  למימד  $k$  ע"י שימוש בהורדת מימד לייניארית, תוקן מדידת איקות הורדת המימד ע"י סכום הריבועים, הבעה:

$$\underset{W \in \mathbb{R}^{k \times d}, U \in \mathbb{R}^{d \times k}}{\text{argmin}} \sum_{i=1}^m \|\mathbf{x}_i - UW\mathbf{x}_i\|^2$$

תפתר ע"י המטריצה  $\tilde{U}$  אשר עמודותיה מרכיבת מ-  $k$  הוקטורים העצמיים של מטריצת השוניות המשותפות:

$$A = \frac{1}{m} \sum_{i=1}^m (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

ומטריצת  $W$  מוגדרת ע"י  $\tilde{U}^\top$ .

עבור DATA  $x_m, x_1, \dots, x_d$  ו-  $u_1, \dots, u_d$  הוקטורים העצמיים של  $S$  המתאימים לע"ע  $\lambda_1 \geq \dots \geq \lambda_d$  כך ש-

- המספר  $i$  נקרא **הערך-**  $i$  **העיקרי** של  $x_m, x_1, \dots, x_d$ .

הוקטור  $u_i$  נקרא **הוקטור-**  $i$  **העיקרי** של  $x_m, x_1, \dots, x_d$ .  
וקטורים אלה אורטוגונרמליים ב-  $\mathbb{R}^d$  ובחרו בקפידה כך ש-  $k$  הוקטורים הראשונים מספקים את הקירוב הליניארי הטוב ביותר מミמד  $k$  לדאטא. במובן מסוים הם "נקודות טיפוסיות", השונות זו מזו בוצרה מקסימלית.

מכיוון ש-  $A$  היא מטריצה ריבועית  $d \times d$  מוגדרת חיובית, הוקטורים העצמיים שלה  $v_1, \dots, v_d$  ("הרכיבים העיקריים") מהווים בסיס אורטוגונרמלי ל-  $\mathbb{R}^d$ . עבור  $k$  קבוע, הורדתPCA ל-  $k$  מימדים מפה כל נקודה  $x_i$  ל-  $U^\top x_i$  ל-  $v_k, \dots, v_1$ .  
עבור  $U$  מטריצה שעמודותיה הם  $v_1, \dots, v_d$ . נבחן כי  $U^\top x_i$  הוא רק וקטור של  $k$  קורדינאות של  $x_i$  לפני  $v_k, \dots, v_1$ .  
כלומר קורדינאות ההטלה של  $x_i$  על תת המרחב האופטימלי, הנitin ע"י  $(v_k, \dots, v_1)$ .

מכיוון ש-  $v_d, \dots, v_1$  בסיס אורטוגונרמלי, נקבל:

- כל נקודה  $x \in \mathbb{R}^d$  יכולה להכתב כצירוף ליניארי של  $v_d, \dots, v_1$ .

ו-  $v_i | x$  הוא פירוק ייחיד של  $x \in \mathbb{R}^d$  לפי הרכיבים העיקריים.  
ונכל להשתמש בו על מנת למפות את  $x$  ולהוריד את מימדו עם אותה העתקה ליניארית,  
ע"י מיפוי ל-  $(v_k | x), \dots, (v_1 | x)^\top \in \mathbb{R}^k$ .

חשוב להבין את ההבדל בין ההטלה על תת המרחב הנפרש ע"י  $k$  הוקטורים העיקריים לבין הקורדינאות של הטלה זו לפני הוקטורים העיקריים. ההטלה תיזג ע"י  $v_1, \dots, v_k$   $x = \sum_{i=1}^k \langle x | v_i \rangle v_i \in \mathbb{R}^d$ , והקורדינאות ייזגו ע"י הוקטור  $(v_k | x), \dots, (v_1 | x)^\top \in \mathbb{R}^k$ .

#### 8.2.7 שימור מרבית השונות (תרגול 11)

אם  $X \in \mathbb{R}^{m \times d}$  מטריצת דיזיין ו-  $S$  מטריצת השינויות המשותפות, ההטלה של  $X$  על תת מרחב ליניארי ממימד  $k$  המשמרת את מרבית השונות של  $X$  נתונה ע"י המטריצה  $U \in \mathbb{R}^{d \times k}$  שעמודותיה הם  $k$  ה"ע" ראשונים של  $S$ .

##### הוכחה

בלי הגבלת הכלליות, יהיו  $v \in \mathbb{R}^d$  וקטור ייחידה בו השתמשנו להטיל את הדאטא עליו ( $k = 1$ ).  
התוחלת על הדאטא המוטל היא:

$$\mathbb{E}_x (v^\top x) = \frac{1}{m} \sum v^\top x_i = v^\top \frac{1}{m} \sum x_i = v^\top \bar{x}$$

(i) מדובר בתוחלת של וקטור, סכום קורדינאות חלקי מס' קורדינאות

מכאן שונות הדאטה המוטל היא:

$$\begin{aligned} Var_x(v^\top x) &\stackrel{i}{=} \mathbb{E}_x \left( (v^\top x - \mathbb{E}_x(v^\top x))^2 \right) \stackrel{ii}{=} \frac{1}{m} \sum (v^\top x - v^\top \bar{x})^2 \\ &= \frac{1}{m} \sum (v^\top (x - \bar{x}))^2 \stackrel{iii}{=} \frac{1}{m} \sum (v^\top (x - \bar{x})) (v^\top (x - \bar{x}))^\top \\ &= \frac{1}{m} \sum (v^\top (x - \bar{x}) (x - \bar{x})^\top v) = v^\top \left( \frac{1}{m} \sum (x - \bar{x}) (x - \bar{x})^\top \right) v = v^\top S v \end{aligned}$$

(i) הדרה (ii) הצבה לפי החישוב לעיל (iii)  $v^\top (x - \bar{x}) \in \mathbb{R}$  ועל כן שווה לטרנספוז שלו

מכאן שפתרון שורצוה למינוס את השונות של הדאטה המוטל ביחס ל- $v$  הוא:

$$\hat{v} = \underset{\|v\|=1}{\operatorname{argmax}} v^\top S v$$

על מנת לפתור את בעיית האופטימיזציה הנ"ל נשתמש בשיטת כופלי לגראנג' עם האילוץ  $v$ .  
מכאן שפונק' הגראנג'יאן נראה כך:

$$\mathcal{L} \stackrel{i}{=} v^\top S v + \lambda g(v) = v^\top S v + \lambda (1 - v^\top v)$$

(i) השתמשנו כאן ב- (a)  $g(v) = 1 - v^\top v$  ולא במינוס לפי הדרה, כי לא יכולה "לבלו" את המינוס.

אם נבצע נגזרת חלקית לפי  $v$  ונשווה לאפס נקבל:

$$\frac{\partial}{\partial v} \mathcal{L} = 2Sv - 2\lambda v = 0 \iff Sv = \lambda v$$

מכאן ש-  $v$  הוא וקטור עצמי של המטריצה  $S$  והוא בעל ערך עצמי  $\lambda$ . אם נכפול את שני האגפים משמאלו ב-  $v^\top$  נקבל:

$$v^\top S v = v^\top \lambda v \iff v^\top S v = \lambda v^\top v \iff v^\top S v = \lambda \|v\|^2 \iff v^\top S v = \lambda$$

ובכך קיבלנו שהוקטור העצמי הוא הוקטור שנutan ששממר את מרבית השונות, והע"ע שלו הוא בדיקת השונות ששמורה.

מכאן שהוקטור ששממר את מרבית השונות הוא הערך הגדול ביותר  $\lambda_1$ , המתקיים  $u_1^\top v = \lambda_1 u_1$ .

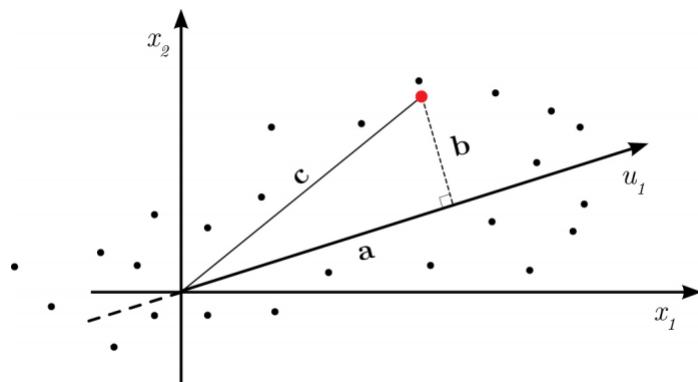
כעת נרצה למצוא את הוקטור הבא, השני בגודל שימור השונות שלו. מכיוון שאנו מחפשים הטלה אורתוגונלית, נוסיף את האילוץ  $u_1^\top v = 0$ .

$$\hat{v} = \underset{\|v\|=1, u_1^\top v = 0}{\operatorname{argmax}} v^\top S v$$

כמו מוקדם, כאשר נפתרור את בעיית האופטימיזציה קיבל שהוקטור  $v$  אותו אנו מחפשים הוא בדיקת  $u_2^\top v = 0$ ,  $u_2$  עם שונות של  $\lambda_2$ . ע"י הוכחה באינדוקציה אפשר לקבל לכל  $d \leq k$ , המיד  $d$ -ה ממד המשמר את מרבית השונות כאשר מטילים את  $X$  עליו הוא נתה המרחב הנפרש ע"י  $k$  הוראונות של  $S$ .

### 8.2.8 הקשר בין תת המרחב בקרוב ביותר לשימוש מרבית השונות (תרגול 11)

עד כה רأינו שני פירושים ל-PCA: הראשון כתת המרחב הקרוב ביותר, והשני כמשמעות מרבית השונות. על מנת להבין איך השניים קשורים זה לזה נتابון בדעתא שמופיעה בגרף הבא:



כאשר הנקודה האדומה מסומנת ע"י  $x_i$ . ע"י כך שנintel אורתוגונליות את  $x_i$  על  $u_1$ , ניצור משולש ישר זווית כאשר:

- הצלע המסומנת  $a$  היא גודל ההטלה של  $x_i$  על  $u_1$ :  $a = \|x_i^\top u_1\|$ .
- זהו הפרמטר אותו אנו **מקסימים** בשיטת מקסום השונות.
- הצלע המסומנת  $b$  היא המרחק בין הנקודה המקורית  $x_i$  להטלה האורתוגונלית שלו על  $u_1$ :  $b = \|x_i^\top - x_i^\top u_1\|$ .
- זהו המרחק אותו אנו **מימזאים** בשיטת תת המרחב הקרוב ביותר.
- הצלע המסומנת  $c$  היא הגודל של  $x_i$ :  $c = \|x_i\|$ .

מכיוון שהמשולש שנוצר הינו ישר זווית, ממשפט פיתגורס אני מקבלים כי  $a^2 + b^2 = c^2$ .  
מכך שאם נמצא פתרון PCA שמיוצר את  $b$ , נקבל פתרון שמקסם את  $a$ , ולהיפך.

### 8.2.9 חישוב מהיר

כאשר  $m \gg d$ , אנו עובדים עם דאטא סט מממד גובה מאד, וזהי סיבה טובה לנסוט לביצוע הורדת מימד. הבעה היא שליכISON של מטריצת השינויות המשותפת מממד  $d \times d$  עליה  $O(d^3)$ .

כאשר  $d$  ענק,  $O(d^3)$  יכול לעלות לנו הרבה. במקרים כאלה נוכל להשתמש בטריק ולחשב את ה"ע" של מטריצת  $m \times m$  במקומות. זה יעלה לנו  $O(m^3)$ . קיים גם אלג' שיחשב את PCA בזמן של  $O(m^2 \cdot d)$ .

**אלגוריתם 1 PCA**

Input: ○ A matrix of  $m$  samples,  $X \in \mathbb{R}^{m \times d}$

○ Number of components,  $n$

1. **if** ( $m > d$ ):  
 1.1.  $A = X^\top X$   
 1.2. Let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  be the eigenvectors of  $A$  with largest eigenvalues
2. **else**  
 2.1.  $B = XX^\top$   
 2.2. Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be the eigenvectors of  $B$  with largest eigenvalues  
 2.3. **for**  $i = 1, \dots, n$  **set**  $\mathbf{u}_i = \frac{1}{\|X^\top \mathbf{v}_i\|} X^\top \mathbf{v}_i$
3. **return**  $\mathbf{u}_1, \dots, \mathbf{u}_n$

**8.2.10 בחירת  $k$** 

בחירת הערך  $k$  היא קריטית - אם נבחר  $k$  גבוהה מדי אונחנו "בזבזנים" בכך שלא זוקקים למיד מכך גובה. אם  $k$  קטן מדי, אנו זורקים חלקים קריטיים מהדadata והורדת המימד לא תופסת את הפיצרים הנחוצים מהדadata.

נתחיל בלהבוח כי ניתן לגנות דרך PCA האם הדadata שלנו יושב בדיק על תת מרחב  $k$  מימי של  $\mathbb{R}^d$ . נניח דата  $x_1, \dots, x_m$  המוכל בתת מרחב  $\mathbb{R}^d$  בעל  $k = \text{dim}(V) \leq d$ . תהי  $S$  מטריצת השונוויות המשותפות מסדר  $d \times d$  מותקיים  $\text{rank}(S) \leq k$  ומשמעות הדבר היא שאם הדadata אכן יושב בדיק על תת מרחב  $k$  מימי, קיבל  $k$  ערכים עיקריים שונים מפאס, והשאר יהיו אפס. אך מה קורה עם הדadata שלנו נמצא קרוב לתת המרחב? קיבל  $k$  ערכים עיקריים גדולים, והשאר יהיו מאוד קטנים אך שונים מפאס. מכאן שהדרך הטובה ביותר לבחור את  $k$  מהדadata הוא כמה ערכים עיקריים "גדולים" יש לנו. זה הרעיון הכללי מאחורי הבחירה כאשר השיטה העיקרית בה משתמשים כיום הוא Scree Plot - דקירת הערכים העיקריים בדרך יורדת על גראף והסקת ניחוש מושכל בהסתמך על הגראף המתתקבל.

**8.3 קלאסטרינג Clustering**

זהו בעיית unsupervised learning, בה בהינתן DATA  $x_1, \dots, x_m$  נרצה לחלק את  $m$  הנקודות  $-k$  סטים שונים. לעיתים יהיה לנו מושג מה  $k$  צריך להיות, ולעתים נדרש להבין לבד בהסתמך על הדadata.

**8.3.1 סיבות לשימוש**

1. שיטה זו מחלקת את הדadata לסטים שווים של נקודות "דומות".

שלב מקדים, אולי יהיה מעונייננו לחלק את הנקודות על מנת לבדוק תכונות אלה של סטים.

2. אולי נרצה לוודא חלוקה כלשהי שימושה נתן לנו, ע"י הרצת האלג' עם הנקודות על מנת לראות אם התקבל אותה החלוקה.

נבחן כי במקרה זה אין לנו תוצאות ולכן אין אמת מוחלטת. מכאן ששיטה זו לא מוגדרת היטב ואינה בעלת תשובה יחידה.

**K-Means 8.3.2 גישה**

נניח כי יש לנו שיטת מדידת מרחק על מרחב הדגימות שלנו. לדוגמה אם  $\mathcal{X} = \mathbb{R}^d$  נוכל לחשב על הנורמה האוקלידית. מבחרת הגדרה, קלאסטרינג של דאטא סט  $\{x_1, \dots, x_m\}$  ל-  $k$  קלאסטרים הוא פשוט חלוקה  $\bigcup_{j=1}^k C_j$  בהינתן חלוקה צו מעל הדאטא שלנו, נוכל להגדיר **פונקציית עלות** עבור החלוקה.

עבור פונקציית מרחק  $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  כלהלן, נגדיר:

$$G_d(C_1, \dots, C_k, \mu_1, \dots, \mu_k) = \sum_{j=1}^k \sum_{x \in C_j} d(x, \mu_j)$$

עבור  $\mu_1, \dots, \mu_k$  ה"נציגים" של הקלאסטרים  $C_1, \dots, C_k$ . בעזרת פונקציה זו, נרצה למצוא חלוקה שסימעת את:

$$\{C_1, \dots, C_k\}^* = \operatorname{argmin}_{\{C_j, \mu_j\}_{j=1}^k} G_d(C_1, \dots, C_k, \mu_1, \dots, \mu_k) = \operatorname{argmin}_{\{C_j, \mu_j\}_{j=1}^k} \sum_{j=1}^k \sum_{x \in C_j} d(x, \mu_j)$$

כאשר במקרה של אלגוריתם ה-K-Means הנציגים נבחרים להיות:

$$\mu_j(C_j) = \operatorname{argmin}_{\mu \in \mathcal{X}} \sum_{x \in C_j} d(x, \mu)$$

והערך המתתקבל נקרא **הסנטרואיד** של הסט  $C_j$ .

איך נמצא את המינימום של  $G_d$ ? נקבע  $k$  ונקבל פונק' של החלוקות  $C_1, \dots, C_k$ . מזורה כרוכ במעבר על כל החלוקות האפשריות של  $m$  איברים ל-  $k$  סטים. הבעיה קומבינטורית ו- NP קשה, ועל כן עוברים ליריסטיות.

**8.3.3 האלגוריתם**

אלג' זה משתמש באלגוריתם של לoid - גישה יוריסטית לסייע  $G$ .  
נניח לשם פשוטות כי  $\mathcal{X} = \mathbb{R}^d$  ופונק' מרחק  $d$  שהיא הנורמה האוקלידית.

**אלגוריתם 2 K-Means**

**קלט:** סט  $x_1, \dots, x_m$  ומס' סטים  $k$ .

1. בחר סנטרואידיים ההתחלתיים  $\mu_1, \dots, \mu_k$ .
2. חזר על התהליך עד להתכנסות:
  - (א) השם לתוך  $C_j$  את כל הנקודות  $x$  שהכى קרובות ל-  $\mu_j$  מאשר לשאר הסנטרואידיים.

(ב) עדכן  $\mu$  להיות הסנטרואיד של  $C_j$  ע"י:

$$\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$$

3. החזר  $C_1, \dots, C_k$ .

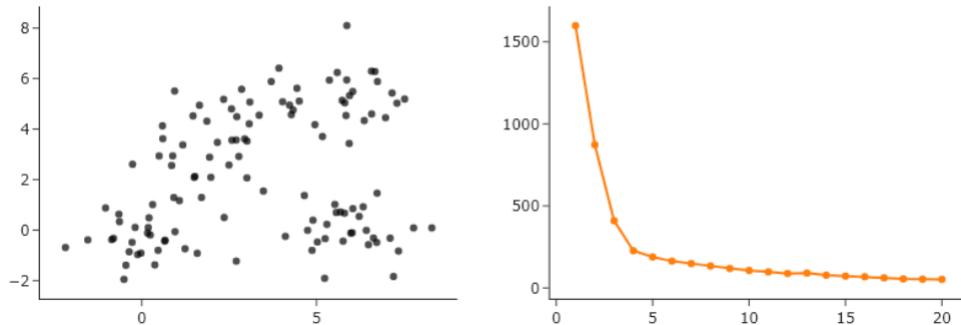
**תכונות:**

1. הוריבאליות של כל סט יורדת מאייטרציה לאיתרציה.
2. האלגוריתם תמיד מתכנס - אך יכול להתכנס מאוד לאט.
3. התוצאה הסופית מושפעת בצורה דרסטית מההשומות ההתחלתיות.  
ייתכן ונקבל תוצאה סאב-אופטימלית: כלומר יתכן והtoutזאה  $C_1, C_2, \dots, C_k$  לא בהכרח תהיה החלוקה שמצוירת את  $G_d$ . זהה תוצאה שנובעת מכך שפונק' המטרה אינה קמורה, ועל כן קיימות לה מס' נקודות מקומיות, אשר כל אחת משנה מינימום שונה.

**בחירה k:**

מסתבר ש- $k$ -הו פרמטר שלשלוט על ה- bias-variance. ככל שהוא גדול יותר (נחלק ליותר סטים) נקבל ערך נמוך יותר עבור  $G$ . זהה אנלוגיה ל- overfitting בלמידה מפוקחת, ככל שאנו מגדילים את סיבוכיות המודל. שיטת הבחירה דומה לשיטה בה השתמשנו עבור PCA - נציג על גרף את התוצאות עבור ערכי  $k$  שונים, ובחר את הערך הראשון שהשיפור אחריו הוא יחסית קטן.

Sum Of Squared Distances As Function Of k



בתמונה - ניתן לראות כי הערך שנבחר הוא  $k = 4$ . במקרה זה הדטא אacen נוצר מאربעה התפלגויות גאוסיאניות שונות, והשיטה אacen הצלחה למצאו את ערך החלוקה הנכון.

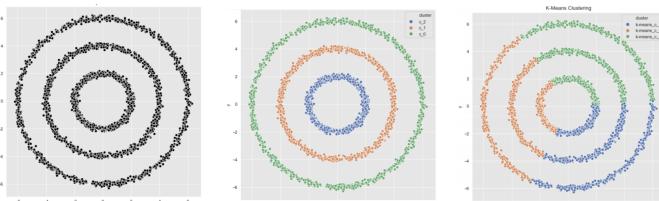
**8.4 קלאסטרינג ספקטרלי**

מהשיטה הקודמת עולה מס' בעיות:

- האם נוכל להתעלם ממרחקים גדולים בעת קליסטו?
- האם נוכל לקלסטר בהסתמך על מרחקים בין זוגות?

נזהה איך נוכל לקלסטר דטא מהתורה הבאה:

חלוקת אותה נרצה לקבל היא האמצעית. שימוש בשיטת ה- K-Means תניב את החלוקה הימנית.



### 8.4.1 השיטה

נרצה להשתמש באיחוד של מס' רעיונות טוביים:

1. נסתכל רק על המרחק בין זוגות של איברים, תוך הצלולות מרחקים גדולים ושמירת מרחקים קטנים.
2. נבנה גרפ סימטרי ממושקל, אשר קודקודיו הם הדגימות  $m$ . צלעותיו הם האפיניות בין הזוגות, כאשר אפיניות גבוהה שköלה למרחק קטן.
3. נתמיע את הדadata בתוך  $\mathbb{R}^k$  בעזרת  $k$  הוקטוריהם העצמיים המובילים של הגרף.

### 8.4.2 הגרף

נדיר את הגרף עי מטריצת סמיכוויות המוגדרת עי:

$$A_{i,j} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\varepsilon}\right)$$

עבור  $0 < \varepsilon$  פרמטר כיוול. נבחן כי אם הדגימות לא נמצאות ב- $\mathbb{R}^d$  (תמונות, קבצי טקסט וכו'), נוכל להגדיר דרך מדידה אחרת לחישוב האפיניות  $A_{i,j}$ .

בכך קיבלנו דרך לבצע unsupervised learning לא רק למרחב דגימות  $\mathbb{R}^d = \mathcal{X}$ .

המטריצה  $A$  היא מטריצת משקלות עבור גרפ סימטרי ממושקל בעל  $m$  קודקודים.

נדיר כעת מטריצה  $D = D^{-1}A = \sum_{j=1}^m A_{i,j}$  עבורה מטריצה אלכסונית  $D_{i,i}$ .

המטריצה  $L$  הינה  $m \times m$ , אי שלילית אשר שורותיה נסכמוות ל-1. היא נקראת גם גרפ לאפליסיין מנורמל.

אם  $A_{i,j} > 0$  רק עבור דגימות  $j, i$  שקרובות אחת לשניה, אז הגרף המתואר עי  $A$  יהיה בעל מס' רכיבי קשרות.

אם נניח כי יש בגרף בדיק  $k$  רכיבי קשרות, נקבל כי  $A$  היא מטריצה  $k$  בלוקים.

מכיוון ש-  $L$  היא רק גרסה מנורמלת של  $A$ , תכונה זו תהיה נcona גם עבורה.

### 8.4.3 הקסם מאחוריו הוקטוריהם העצמיים

מכיוון שהשורות של  $L$  נסכמוות ל-1, למטריצה  $L$  לפחות וע"ז עצמי שמתאים לע"ז 1.

נניח כי קיימים 3 רכיבי קשרות, ויהיו  $v_1, v_2, v_3$  הו"ז המתאימים של  $L$ .

נתבונן בדגם ? . אם נמפה דגימה ? לוקטור ב- $\mathbb{R}^3$  שמכיל את הקורדיינטה ה-  $i$  של  $v_1, v_2, v_3$ .

וקטור זה יהיה מהצורה  $(a, 0, 0)$  או  $(0, b, 0)$  או  $(0, 0, c)$  כתלות בסט של הדגימה ה-  $i$ .

כעת נוכל להפעיל K-Means על  $\mathbb{R}^3$  ולקבל את החלוקה שנרצה.

### 8.4.4 עלות החישוב

על מנת להשתמש בклиיסטור ספקטורי על  $m$  דגימות, נctrיך לאחסן וללכSEN מטריצה  $m \times m$ .

בפועל זאת משימה לא נוראה - הגרף  $L$  הוא לרוב מטריצה מדוולת אשר ניתן לשמר במבנה נתונים

מיוחדים. את  $k$  הערכים העצמיים נוכל למצוא עי power methods. משם נרץ K-Means על מימד יחסית נמוך.

## 9 הרצאה 9 - Kernels Methods - 9

### 9.1 רעיון כללי

נזכיר בחלוקת הhipothesis היליניארית עבור רגרסיה:

$$\mathcal{H}_{lin} = \left\{ (x_1, \dots, x_d) \mapsto w_0 + \sum_{i=1}^d w_i x_i \mid w_0, w_1, \dots, w_d \in \mathbb{R} \right\}$$

עבור סיווג:

$$\mathcal{H}_{lin} = \left\{ (x_1, \dots, x_d) \mapsto \text{sign} \left( w_0 + \sum_{i=1}^d w_i x_i \right) \mid w_0, w_1, \dots, w_d \in \mathbb{R} \right\}$$

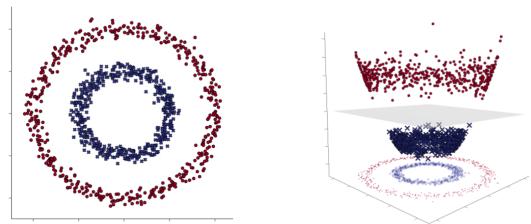
נבחר מיפוי  $\psi(\mathbf{x})_1, \dots, \psi(\mathbf{x})_k \in \mathbb{R}^k$  עבור  $d > k$ . לכן עבור  $\mathbf{x} \in \mathbb{R}^d$  נקבל  $\psi(\mathbf{x}) \in \mathbb{R}^k$  עם קורדיינאות  $\psi(\mathbf{x})_1, \dots, \psi(\mathbf{x})_k \in \mathbb{R}^k$ .

נתבונן בחלוקת הבאות:

$$\mathcal{H}_\psi = \left\{ \mathbf{x} \mapsto w_0 + \mathbf{w}^\top \psi(\mathbf{x}) \mid w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d \right\} \quad \left| \quad \mathcal{H}_\psi = \left\{ \mathbf{x} \mapsto \text{sign}(w_0 + \mathbf{w}^\top \psi(\mathbf{x})) \mid w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d \right\} \right.$$

מודלים אלו ליניארים ב- $\psi(\mathbf{x}) \in \mathbb{R}^k$  ולא ב- $\mathbf{x} \in \mathbb{R}^d$ .

#### 9.1.1 מוטיבציות



(a) Two class dataset that is not linearly separable

(b) Dataset mapped to  $\mathbb{R}^3$  using the mapping  $(x_1, x_2) \rightarrow (x_1^2, x_2^2, x_1^2 + x_2^2)^\top$

1. העשרה שלחלוקת הhipothesis היליניארית.

2. יעיל חישובית: מציאת  $h_S \in \mathcal{H}$  ללא הסתכילות בפיצ'רים שבמימדים הגבוהים.

3. אפשר עבודה ללא פיצ'רים, רק מרחקים בזוגות.

## 9.2 דוגמא ראשונה - Polynomial Fitting

נתבונן בבעיית רגרסיה  $\mathcal{Y} = \mathbb{R}$  ב- $\mathcal{X} = \mathbb{R}$ . יש לנו סט אימון  $(x_1, y_1), \dots, (x_m, y_m)$ . נניח כי התווגים האמיטיים מוגעים מהפונקציה  $y = x^2 + x + b$  (לדוגמא). המודל היליניארי  $h: \mathbb{R} \rightarrow \mathbb{R}$  הוא?

$$\psi(x) = (1, x, x^2)$$

המודל היליניארי ב- $\mathbb{R}^3$  שיאומן על  $(\psi(x_1), y_1), \dots, (\psi(x_m), y_m)$  יהיה מצוין. השיטה לה קראנו "התאמת פולינומית" - שימוש ברגרסיה ליניארית על מנת ללמידה פונקציה מאוד לא ליניארית, משתמשת ב- **גרעין פולינומי** מדרגה  $n$ . לדוגמה עבור  $\mathcal{X} = \mathbb{R}$ , הגרעין הפולינומי משתמש בפונקציה:

$$\psi(x) = (1, x, \dots, x^n)$$

### 9.2.1 $\mathbb{R}^2$ הפללה ל-

אם נוכל ללמידה פונקציה לא ליניארית  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^k$  ע"י מודל ליניארי שימושי בפיצ'רים במימד גובה יותר  $\mathbb{R}^k$ ? נזכיר בקורסכם של פולינומים רב מושתנים:

$$p(x_1, x_2) = w_{(0,0)} + w_{(1,0)}x_1 + w_{(0,1)}x_2 + w_{(2,0)}x_1^2 + w_{(0,2)}x_2^2 + w_{(1,1)}x_1x_2$$

אם נגדיר  $\mathbf{w} \in \mathbb{R}^6$ .  $\mathbf{w}$  נקבע שוב  $\psi(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$  עבור  $\psi(\mathbf{x}) = \langle \mathbf{w}, \psi(\mathbf{x}) \rangle$

### 9.2.2 $\mathbb{R}^d$ הפללה ל-

פולינום רב מושתנים מדרגה  $n$   $\mathbb{R}^d \rightarrow \mathbb{R}$  הוא מהצורה:

$$p(\mathbf{x}) = \sum_{\mathbf{a} \in \mathbb{N}^d: \sum a_i \leq n} w_{\mathbf{a}} \prod_{i=1}^d x_i^{a_i}$$

עבור  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^k$ . כמוקדם, נוכל שוב כתוב  $p(\mathbf{x}) = \langle \mathbf{w}, \psi(\mathbf{x}) \rangle$  עבור  $\mathbf{x} = (x_1, \dots, x_d)^T$ . במקורה זה,  $\mathbf{a} \in \mathbb{N}^d$ ,  $k = \#\{\mathbf{a} \in \mathbb{N}^d \mid \sum a_i \leq n\}$ , וכל קורדינאטה של  $\psi(\mathbf{x}) \in \mathbb{R}^k$  מאונדקסת ע"י כלשהו עם  $\psi(\mathbf{x})_{\mathbf{a}} = \prod_{i=1}^d x_i^{a_i}$ .

### Kernel Trick - 9.3

עבור סט אימון  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$ , והמחלקות  $\mathcal{H}_{\psi}$  אותן ראיינו עבור רגרסיה וסיווג, נתבונן באיל' למידה הבוחר  $h_S \in \mathcal{H}_{\psi}$  לפי בעיית האופטימיזציה הבאה:

$$(1) \quad \mathbf{w}_S = \underset{\mathbf{w}}{\operatorname{argmin}} f(\langle \mathbf{w} \mid \psi(\mathbf{x}_1) \rangle, \dots, \langle \mathbf{w} \mid \psi(\mathbf{x}_m) \rangle) + \lambda \|\mathbf{w}\|^2$$

עבור  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  ו-  $\mathbf{w} \in \mathbb{R}^k$

cutet נתבונן בעיית אופטימיזציה נוספת:

$$(2) \quad \alpha_S = \underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} (f(G\alpha) + \lambda \alpha^T G \alpha)$$

עבור  $f, \lambda$  ו-  $G_{i,j} = \langle \psi(\mathbf{x}_i), \psi(\mathbf{x}_j) \rangle$  זהות לאלה מהבעיה הראשונה.

### Kernel Trick - 9.3.1 Kernel Trick - 9.3.1

המיצער  $\alpha_S$  עבור בעיית האופטימיזציה השנייה הקשור למיצער  $\mathbf{w}_S$  עבור הבעיה הראשונה, ע"י כך שמותקיים:

$$\mathbf{w}_S = \sum_{i=1}^m (\alpha_S)_i \psi(\mathbf{x}_i)$$

### 9.3.2 הרחבה מושג ה- Kernel

נשאלת השאלה מה הוא בעצם Kernel ?  
לפי הגדרה, פונקציית Kernel  $K(\cdot, \cdot)$  המתאימה לפונק'  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^k$  היא מהצורה:

$$K(\mathbf{x}, \mathbf{x}') = \langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle$$

או במלילים אחרות, ה- Kernel הוא מכפלה פנימית במינימד גובה  $\mathbb{R}^k$ .

לפי משפט ה- Kernel Trick, במקומות מסוימים ניתן לפשט את בעיית האופטימיזציה הראשונה, נוכל להסתפק בפתרו את השניה.  
נבחן כי בעיית האופטימיזציה השניה לא רואה את מינימד  $\mathbb{R}^k$  או את הפונקציה  $\psi$ .  
היא תלויה רק במטריצה גראם שכניםיסותיה מקיימות ( $G_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$  עבור  $\mathbf{x}_m, \mathbf{x}_{m-1}, \dots, \mathbf{x}_1$  מסט האימון).  
מכאן שעל מנת לאמן את המודול כל שעליינו לדעת זה את  $K(\mathbf{x}_i, \mathbf{x}_j)$  ולא  $\psi(\mathbf{x}_i)$ .

### 9.3.3 אימון וחיזוי בעזרת הטריך

בהתנן סט אימון, אם נוכל לחשב את  $K(\mathbf{x}, \mathbf{x}')$  באופן יעיל לכל  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ , נוכל לחשב את המטריצה  $G$  ולמצוא

$$\alpha_S = \underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} (f(G\alpha) + \lambda \alpha^\top G \alpha)$$

נזכיר כי במקרה של רגרסיה מחלוקת ההיפותזות שלנו היא  $\langle \mathbf{w}, \psi(\mathbf{x}) \rangle \mid \mathbf{w} \in \mathbb{R}$  וועל כן נבחר בהיפוטזה:

$$\mathbf{w}_S = \sum_{i=1}^m (\alpha_S)_i \psi(\mathbf{x}_i)$$

כעת, בהינתן דוגמה חדשה  $\mathbf{x} \in \mathbb{R}^d$ , נחשב:

$$\begin{aligned} \langle \mathbf{w}_S \mid \psi(\mathbf{x}) \rangle &= \left\langle \sum_{i=1}^m (\alpha_S)_i \psi(\mathbf{x}_i) \mid \psi(\mathbf{x}) \right\rangle \\ &= \sum_{i=1}^m (\alpha_S)_i \langle \psi(\mathbf{x}_i) \mid \psi(\mathbf{x}) \rangle = \boxed{\sum_{i=1}^m (\alpha_S)_i K(\mathbf{x}_i, \mathbf{x})} \end{aligned}$$

### 9.3.4 בחזרה למוטיבציות

ניתן לראות כי  $\psi$  אכן מרחיב ומעシリ את מחלוקת ההיפותזות שלנו.

אם נבחר בנוסף  $\psi$  כזה כך ש-  $\langle \psi(\mathbf{x}'), \psi(\mathbf{x}) \rangle = K(\mathbf{x}, \mathbf{x}')$  חישיב ביעילות,

לא נצורך לחשב את  $\psi(\mathbf{x})$  או אפילו להסתכל על המרחיב  $\mathbb{R}^k$  בזמן האימון והחיזוי.

במקרה בו אין פיצרים, אך יש אפינים  $A_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$  - מרחקים בין זוגות, יתכן ונוכל להעמיד פנים כי  $(A_{i,j})$  פיצרים מימייד גובה יותר.  
ובעצם להשתמש בשיטות גרעין למרות שאין לנו  $\mathbb{R}^d$  (פיצרים מוקריים) ו-  $\mathbb{R}^k$  (פיצרים מימייד גובה יותר).

## 9.3.5 המשפט המיצג

לכל כל למידה מהצורה:

$$\mathbf{w}_S = \underset{\mathbf{w}}{\operatorname{argmin}} f(\langle \mathbf{w} | \psi(\mathbf{x}_1) \rangle, \dots, \langle \mathbf{w} | \psi(\mathbf{x}_m) \rangle) + \lambda \|\mathbf{w}\|^2$$

עבור  $\mathbf{w}^* = \sum_{i=1}^m \alpha_i \psi(\mathbf{x}_i)$  ש-  $\alpha \in \mathbb{R}^m$ ,  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  ו-  $\mathbf{w} \in \mathbb{R}^k$

## הוכחה

$$\text{נסמן } \|\tilde{\mathbf{w}}\|^2 = f(\langle \tilde{\mathbf{w}} | \psi(\mathbf{x}_1) \rangle, \dots, \langle \tilde{\mathbf{w}} | \psi(\mathbf{x}_m) \rangle) + \lambda \|\tilde{\mathbf{w}}\|^2$$

יהי  $\mathbf{w}$  הטלת האורתוגונליות של  $\text{span}(\psi(\mathbf{x}_1), \dots, \psi(\mathbf{x}_m))$  על המרחב הנפרש ע"י  $\mathbf{w}^* = \underset{\tilde{\mathbf{w}} \in \mathbb{R}^k}{\operatorname{argmin}} G(\tilde{\mathbf{w}})$  ונסמן  $\mathbf{w} - \mathbf{w}^* = \mathbf{u}$ . אז מתקיים משפט פיתגורס כי  $\|\mathbf{w}^*\|^2 = \|\mathbf{w}\|^2 + \|\mathbf{u}\|^2$  ובנוסף מתכונת הטלת האורתוגונליות:

$$\forall i \in [m] : \langle \mathbf{w} | \psi(\mathbf{x}_i) \rangle = \langle \mathbf{w}^* | \psi(\mathbf{x}_i) \rangle$$

מכאן שמתקיים:

$$G(\mathbf{w}) = (\langle \mathbf{w}^* | \psi(\mathbf{x}_1) \rangle, \dots, \langle \mathbf{w}^* | \psi(\mathbf{x}_m) \rangle) + \lambda (\|\mathbf{w}^*\|^2 - \|\mathbf{u}\|^2)$$

(i) מסתמך על 2 התכונות שלעיל.

מאותIMALיות הפתרון  $\mathbf{w}^*$  אנו יודעים כי  $G(\mathbf{w}^*) \leq G(\mathbf{w})$  אבל  $G(\mathbf{w}^*) \leq G(\mathbf{w})$  מכיוון שחייב להתקיים כי  $\|\mathbf{u}\| = 0$  ולכן  $G(\mathbf{w}^*) = G(\mathbf{w})$ .

## 9.3.6 הוכחת הטריך

לפי המשפט המיצג, פתרונו אופטימלי ניתן לכתיבה כ-  $\mathbf{w} = \sum_i \alpha_i \psi(\mathbf{x}_i)$ . נסמן ב-  $G$  את מטריצת גראם המקימה  $G_{i,j} = \langle \psi(\mathbf{x}_i) | \psi(\mathbf{x}_j) \rangle$  כיו:

$$\langle \mathbf{w} | \psi(\mathbf{x}_i) \rangle = \left\langle \sum_j \alpha_j \psi(\mathbf{x}_j) | \psi(\mathbf{x}_i) \right\rangle = \sum_j \alpha_j \langle \psi(\mathbf{x}_j) | \psi(\mathbf{x}_i) \rangle = (G\alpha)_i$$

$$\|\mathbf{w}\|^2 = \left\langle \sum_j \alpha_j \psi(\mathbf{x}_j) | \sum_j \alpha_j \psi(\mathbf{x}_j) \right\rangle = \sum_{i,j=1}^m \alpha_i \alpha_j \langle \psi(\mathbf{x}_i) | \psi(\mathbf{x}_j) \rangle = \alpha^\top G \alpha$$

ועל כן נוכל לקבל פתרון אופטימלי ע"י מזעור  $\alpha$  מעל:

$$\underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} (f(G\alpha) + \lambda \alpha^\top G \alpha)$$

## 9.4 אלגוריתמי Kernel

### 9.4.1 Kernel Hard SVM

ראינו בהרצאה על מסובגים כי במקרה ההומוגני ( $b = 0$ ) של SVM-Hard מתקיים:

$$\operatorname{argmin}_{\mathbf{w}} \left\{ \|\mathbf{w}\|^2 \quad s.t. \quad \forall i : y_i \langle \mathbf{w} | \mathbf{x}_i \rangle \geq 1 \right\}$$

על מנת "לגרען" את SVM-Hard, עליינו (1) להחליף את  $\mathbf{x}$  ב-  $\psi(\mathbf{x})$  לבטא הכל באמצעות מטריצת גראם  $G$  על פיה המשפט המיצג, גרסאות Kernel Hard-SVM היא מהצורה:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} \quad s.t. \quad \forall i : y_i (G \boldsymbol{\alpha})_i \geq 1$$

### 9.4.2 Kernel Soft SVM

ראינו בהרצאה על מסובגים כי במקרה ההומוגני ( $b = 0$ ) של SVM-Soft מתקיים:

$\operatorname{argmin}_{\mathbf{w}, \xi} \lambda \ \mathbf{w}\ ^2 + \frac{1}{m} \sum_{i=1}^m \xi_i$
subject to $\forall i \in [m] : y_i \langle \mathbf{x}_i   \mathbf{w} \rangle \geq 1 - \xi_i \quad \wedge \quad \xi_i \geq 0$

ראינו כי בעיה שוקלה ללא אילוצים היא מהצורה:

$$\operatorname{argmin}_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^m \ell^{\text{hinge}}(y_i \cdot \langle \mathbf{w} | \mathbf{x}_i \rangle)$$

כאשר הפונקציה  $\ell^{\text{hinge}}$  מוגדרת ע"י  $\ell^{\text{hinge}}(a) = \max\{0, 1 - a\}$  עבור  $a \in \mathbb{R}$ . על פיה המשפט המיצג, גרסאות Kernel Soft-SVM היא מהצורה:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \lambda \boldsymbol{\alpha}^\top G \boldsymbol{\alpha} + \frac{1}{m} \sum_{i=1}^m \ell^{\text{hinge}}(y_i \cdot (G \boldsymbol{\alpha})_i)$$

### 9.4.3 Ridge Regression

ראינו כי Ridge Regression במקרה בו  $w_0 = 0$  היא מהצורה:

$$\operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^m |y_i - \langle \mathbf{w} | \mathbf{x}_i \rangle|^2 + \lambda \|\mathbf{w}\|_2^2$$

על מנת לגרען, נחליף את  $\mathbf{x}$  ב-  $\psi(\mathbf{x})$ :

$$\operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^m |y_i - \langle \mathbf{w} | \psi(\mathbf{x}_i) \rangle|^2 + \lambda \|\mathbf{w}\|_2^2$$

ומה המשפט המיצג נקבל כי הבעיה שකולה ל-

$$\underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} \|G\alpha - y\|^2 + \lambda \alpha^\top G \alpha$$

מתקיים:

$$\|G\alpha - y\|^2 + \lambda \alpha^\top G \alpha = \alpha^\top G^2 \alpha - 2y^\top G \alpha + \|y\|^2 + \lambda \alpha^\top G \alpha$$

ניקח גרדיאנט לפי  $\alpha$  ושווה אותו לאפס:

$$\nabla_\alpha E(\alpha) = 2G^2 \alpha - 2Gy + 2\lambda G \alpha = 0$$

נחלק ב-2, נעביר אגפים ונוציה גורמים משותפים:

$$G(G + \lambda I)\alpha = Gy \Rightarrow \alpha = (G + \lambda I)^{-1}y$$

מכאן שהחיזוי נקודה חדשה  $x$  נתונה ע"י:

$$\hat{y}(x) = k^\top \alpha = k^\top (G + \lambda I)^{-1}y$$

עבור  $.k_i(x) = K(x_i, x)$

#### 9.4.4 רגרסיב Kernel Logistic

ראינו כי רגרסיה לוגיסטיבית במקרה בו  $w_0 = 0$  היא מהצורה:

$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left( \sum_{i=1}^m \log(1 + e^{\langle x_i | w \rangle}) - \sum_{i=1}^m y_i \cdot \langle x_i | w \rangle \right)$$

ונכל להוסיף פרמטר רגולרייזציה  $\ell_2$  ולקבל:

$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left( \sum_{i=1}^m \log(1 + e^{\langle x_i | w \rangle}) - \sum_{i=1}^m y_i \cdot \langle x_i | w \rangle + \lambda \|w\|^2 \right)$$

ומה המשפט המיצג נקבל כי הבעיה שකולה ל-

$$\underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} \left( \sum_{i=1}^m \log(1 + e^{(G\alpha)_i}) - \sum_{i=1}^m y_i \cdot (G\alpha)_i + \lambda \alpha^\top G \alpha \right)$$

ניקח גרדיאנט לפי  $\alpha$  ושווה אותו לאפס:

$$(\nabla_\alpha E(\alpha))_j = \sum_{i=1}^m G_{i,j} \cdot \frac{e^{(G\alpha)_i}}{1 + e^{(G\alpha)_i}} - (Gy)_j + 2\lambda (G\alpha)_j = 0$$

$$\sum_{i=1}^m G_{i,j} \left( \frac{1}{1 - e^{-(G\alpha)_i}} + 2\lambda \alpha_i \right) = \sum_{i=1}^m G_{i,j} y_i$$

ומכאן שהוא פתרון למשוואה:

$$\boxed{\frac{1}{1 - e^{-(G\alpha)_i}} + 2\lambda \alpha_i = y_i}$$

## 9.4.5 PCA גרעין

תהי מטריצה  $\mathbf{X} \in \mathbb{R}^{m \times d}$  מmorczat (מקיימת  $\sum_i (\mathbf{x}_i)_j = 0$  לכל  $j \in [d]$ ).  
 ראיינו כי באלגוריתם PCA אנו פותרים את בעיית הע"ע של המטריצה  $C$  המתאים לערך עצמי  $0 \neq \lambda$  ש- $\lambda\mathbf{v} = C\mathbf{v}$ .  
 נרצה להביע ביטוי זה ע"י מכפלות פנימיות. יהיו  $\mathbf{v}$  וקטור עצמי של  $C$  המתאים לערך עצמי  $0 \neq \lambda$  כך ש- $\lambda\mathbf{v} = C\mathbf{v}$   
 משמעות הדבר היא ש- $\mathbf{v} \in \text{Im}(C)$  ולכן שיקד ל- $(\mathbf{x}_1, \dots, \mathbf{x}_m)$ .  
 לפיכך, פתרת בעיית הע"ע של  $C$  שקופה למציאות וקטור  $\mathbf{v}$  עבורו:

$$\clubsuit \quad \forall i \in [m] : \lambda \langle \mathbf{x}_i, \mathbf{v} \rangle = \langle \mathbf{x}_i, \lambda\mathbf{v} \rangle = \langle \mathbf{x}_i | C\mathbf{v} \rangle$$

בנוסף, ע"פ הגדרת  $C$  מתקיים כי  $\langle \mathbf{x}_i | \mathbf{v} \rangle = \sum_i \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}$  ולכן מתקיים:

$$\mathbf{v} = \sum_i \frac{1}{\lambda} \mathbf{x}_i \langle \mathbf{x}_i | \mathbf{v} \rangle$$

$$\mathbf{v} = \sum_{i=1}^m \alpha_i \mathbf{x}_i = \mathbf{X}^\top \boldsymbol{\alpha} \quad \text{נקבל כי } \alpha_i = \frac{1}{\lambda} \langle \mathbf{x}_i | \mathbf{v} \rangle$$

במוקם לפטור את ♣, נבחן כי עבור  $i \in [m]$  קלשו מתקיים:

$$\begin{aligned} \lambda \langle \mathbf{x}_i, \mathbf{v} \rangle &= \langle \mathbf{x}_i, \lambda\mathbf{v} \rangle = \langle \mathbf{x}_i, C\mathbf{v} \rangle = \left\langle \mathbf{x}_i, C \left( \sum_{j=1}^m \alpha_j \mathbf{x}_j \right) \right\rangle \\ &= \sum_{j=1}^m \alpha_j \langle \mathbf{x}_i, C\mathbf{x}_j \rangle = \sum_{j=1}^m \alpha_j \left\langle \mathbf{x}_i, \sum_{\ell} \mathbf{x}_{\ell} \mathbf{x}_{\ell}^\top \mathbf{x}_j \right\rangle \\ &= \sum_{j=1}^m \alpha_j \left\langle \mathbf{x}_i, \sum_{\ell} \mathbf{x}_{\ell} \langle \mathbf{x}_{\ell} | \mathbf{x}_j \rangle \right\rangle = \sum_{j=1}^m \alpha_j \sum_{\ell} \langle \mathbf{x}_{\ell} | \mathbf{x}_j \rangle \langle \mathbf{x}_i, \mathbf{x}_{\ell} \rangle \end{aligned}$$

ולכן קיבל כי:

$$\lambda \sum_j \alpha_j \langle \mathbf{x}_i | \mathbf{x}_j \rangle = \sum_{j=1}^m \alpha_j \sum_{\ell} \langle \mathbf{x}_{\ell} | \mathbf{x}_j \rangle \langle \mathbf{x}_i, \mathbf{x}_{\ell} \rangle$$

חיבור  $m$  המשוואות וכתיבה ע"י נוטציה מטריציונית תוביל למשוואת  $\lambda G\boldsymbol{\alpha} = G^2\boldsymbol{\alpha}$   
 כאשר  $G$  מטריצת גראם המוגדרת ע"י  $: \mathbf{x}_1, \dots, \mathbf{x}_m$   
 הכניסה ה- $i, j$  מקיימת  $G_{i,j} = \langle \mathbf{x}_i | \mathbf{x}_j \rangle = \mathbf{x}_i^\top \mathbf{x}_j$  ולהילופין  
 עבור הקטורים העצמיים של  $G$ , שלא מתאימים לע"ע אפס, הפתרון ל- $\lambda G\boldsymbol{\alpha} = G^2\boldsymbol{\alpha}$  שקול לפטור  $\lambda\boldsymbol{\alpha} = G\boldsymbol{\alpha}$

ברגע שנמצא  $\alpha^m, \alpha^1, \dots, \alpha^0$  וקטוריים עצמיים של  $G$  נוכל:

$$\mathbf{1.} \text{ לקבלת וקטוריים עצמיים של } C \text{ ע"י } \sum_{i=1}^m \alpha_i^\ell \mathbf{x}_i.$$

**2. להטיל את נקודות הדאטה לתת-מרחב ממימד נמוך יותר ע"י:**

$$\tilde{\mathbf{x}}_\ell = \langle \mathbf{v}^\ell | \mathbf{x} \rangle = \sum_i \alpha_i^\ell \langle \mathbf{x}_i | \mathbf{x} \rangle$$

ע"י כך שהראיינו כי ניתן לחשב את הע"ע וההטלה בעזרת הדאטה ומכפלות פנימיות בלבד, הראיינו בעצם כי ניתן להפעיל את השיטת הגרעין גם על אלגוריתם ה-PCA. לכן עבור מיפוי  $\psi$  עבורו  $\psi(\mathbf{x}_i)^\top \psi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$

$$\mathbf{1.} \text{ נפתרו את בעיית הע"ע } G\alpha = \lambda \text{ עבור } \langle \psi(\mathbf{x}_i) | \psi(\mathbf{x}_j) \rangle.$$

$$\mathbf{2.} \text{ נטיל את נקודות הדאטה ע"י נס庭 } \tilde{\mathbf{x}}_\ell = \langle \mathbf{v}^\ell | \psi(\mathbf{x}) \rangle = \sum_i \alpha_i^\ell \langle \psi(\mathbf{x}_i) | \psi(\mathbf{x}_j) \rangle = \sum_i \alpha_i^\ell k(\mathbf{x}_i, \mathbf{x}).$$

## 9.5 מספר Kernels מפורטים

### 9.5.1 הגרעין הפולינומי

נקבע פולינום מדרגה  $n$ . בתחילת ההרצתה הצענו  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^k$  שקורדינאות הפלט שלה מאונדקשות ע"י  $\mathbf{a} \in \mathbb{N}^d$  המקיימים  $\sum_i a_i \leq n$  ומוגדרות ע"י:  $\psi(\mathbf{x})_a = \prod_{i=1}^d x_i^{a_i}$ . אם נחשב את  $k$  כפונקציה של  $n$ ,  $d$  קיבל כי  $k = \binom{n+d}{n} = \frac{(n+d)!}{n! \cdot d!} = O(d^n)$ .

אנו עלולים לחושש כתע כי חישוב  $G$  ייקח לנו המון זמן, אך מסתבר שאם נגידיר את  $\psi$  בעזרת מס' קבועים נוספים (שאינם מעניינים אותנו מכיוון שאנו לא משתמשים אפילו על  $\psi$ ), אנו מקבל כי:  $K(\mathbf{x}, \mathbf{x}') = (1 + \langle \mathbf{x} | \mathbf{x}' \rangle)^n$ .

### 9.5.2 הגרעין האוציאיניי מעל $\mathbb{R}$

נתבונן בבעיה בה  $\mathcal{X} = \mathbb{R}$  ונבחן את המיפוי  $\psi$  הבא:

$$\psi(x) = \left( 1, e^{-\frac{x^2}{2}} x, \frac{1}{\sqrt{2}} e^{-\frac{x^2}{2}} x^2, \dots \right) \quad \forall n \in \mathbb{N} : \psi(x)_n = \frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n$$

**הגדרה - מרחבי היוצרת.** מרחב הילברט  $H$  הוא מרחב וקטורי מעל  $\mathbb{R}$  בעל מכפלה פנימית  $\langle \cdot | \cdot \rangle$  עם התכונה לפיה כל סדרת קושי בnormה  $\sqrt{\langle \mathbf{x} | \mathbf{x} \rangle} = \| \mathbf{x} \|$  מתכנסת ב- $H$ .

**דוגמה:** המרחב  $\ell_2$  מורכב מכל הרצפים האינסופיים  $(x_1, x_2, x_3, \dots)$  כך ש-  $x_i \in \mathbb{R}$  והסדרה  $\sum_{i=1}^{\infty} x_i^2$  מתכנסת.

בצירוף המכפלה הפנימית  $\langle \mathbf{x} | \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i$  זה מרחב הילברט ממימד אינסופי.

○ כאשר לכל סדרת קושי יש גבול, אנו קוראים למרחב **שלם**. תנאי זה מבטיח שלכל וקטור יש הטלה אורתוגונלית ייחודית על כל תת מרחב סגור של  $H$ .

מתקיים (תווך התעלמות מקבועים כמו עצרת ומקדמי מולטינום) כי:

$$\begin{aligned} \langle \psi(x) \mid \psi(x') \rangle &= \sum_{i=0}^{\infty} \left( \frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n \right) \left( \frac{1}{\sqrt{n!}} e^{-\frac{(x')^2}{2}} (x')^n \right) \\ &= e^{-\left(\frac{x^2+(x')^2}{2}\right)} \sum_{i=0}^{\infty} \left( \frac{(xx')^n}{n!} \right) = e^{-\frac{(x-x')^2}{2}} \end{aligned}$$

ועל  $\mathbb{R}^d = \mathcal{X}$  נוכל להגיד כי  $\psi$  כאשר הקורדיינאות של  $(x)$   $\psi$  מאונדקשות ע"י הסט האינסופי  $\{a \in \mathbb{N}^d \mid \psi(a) \neq 0\}$

$$\psi(x)_a = \frac{1}{\sqrt{n!}} e^{-\frac{\|x\|^2}{2}} \prod_{i=1}^d x_i^{a_i}$$

וחישוב דומה למקרה בו  $d = 1$  מראה כי:

$$K(x, x') = \langle \psi(x), \psi(x') \rangle = e^{-\frac{\|x-x'\|^2}{2}}$$

## 9.6 אפיון לפונקציות גרעין

נניח ויש לנו פונקציית גרעין  $K(\cdot, \cdot)$  כלשהי מעל  $\mathbb{R}^d$ . איך נדע האם זהה פונקציה "כשרה"?  
כלומר איך נדע האם קיימת פונקציה  $\psi$  כלשהי מיילר  $\mathbb{R}^d$  עבורה  $\langle \psi(x), \psi(x') \rangle = K(x, x')$  לכל  $x, x' \in \mathbb{R}^d$ .

### 9.6.1 הלמה של מרסר

פונקציית סימטרית  $\mathbb{R} \rightarrow \mathcal{X} \times \mathcal{X} : K$  ממשת מכפלת פנימית במרחב הילברט כלשהו אם היא מטריצה חי-מוגדרת חיובית. כלומר לכל  $x_1, \dots, x_m$ , מטריצת גרם  $G_{i,j} = K(x_i, x_j)$  היא מטריצה חי-מוגדרת חיובית.

### 9.6.2 יצירת גרעינים מגרעינים קיימים

نוכל לבנות סוגים שונים של גרעינים מגרעינים פשוטים יותר. נראה כמה כלליים שיעזרו לנו.  
נניח  $(x', x)$  ו-  $K_1(x, x')$  ו-  $K_2(x, x')$  גרעינים חוקיים. אז כל הבאים גרעינים חוקיים:

1. לכל פונק'  $f$   $K(x, x') = f(x) K_1(x, x') f(x') - f(x) K_1(x, x') f(x')$

2. החצירות הלייניארי הא-שלילי של גרעינים חוקיים הוא גרעין חוקי. עבור  $a_1, a_2 \geq 0$  נקבל:

$$K(x, x') = a_1 K_1(x, x') + a_2 K_2(x, x')$$

3. המכפלת של גרעינים חוקיים היא גרעין חוקי.  $K(x, x') = K_1(x, x') \cdot K_2(x, x')$

## 10 הרצאה 10 - אופטימיזציה קמורה ו-

### 10.1 הקדמה

ראינו במהלך הקורס מס' עקרונות למידה בהם בחירת ההיפוטזה היא ע"י פתרת בעית אופטימיזציה קמורה. חישוב ה- Squared Loss, בעיות ה- Hard/Soft SVM, רגרסיה לוגיסטיבית, רידג' ולאסו.

### 10.2 הגדרות

#### 10.2.1 קבוצה קמורה

יהי  $V$  מרחב וקטורי. קבוצה  $C \subseteq V$  תקרא **קמורה** אם לכל שני וקטורים  $u, v \in C$  וסקלר  $\alpha \in [0, 1]$  מתקיים  $\alpha u + (1 - \alpha)v \in C$ . ביטוי זה נקרא **צירוף קמור**. מבחינה גיאומטרית, הקבוצה  $C$  היא קמורה אם ורק אם הישר המחבר כל שתי נקודות  $v \in C, u \in C$  מוכל ב- $C$ .

#### 10.2.2 דוגמאות לקבוצות קמורות

- כדור היחידה:  $\{x \mid \|x\| \leq r\} \subseteq \mathbb{R}^d$
- על מישור:  $\{x \mid w^\top x = b\} \subseteq \mathbb{R}^d, w \in \mathbb{R}^d, b \in \mathbb{R}$
- חצאי על מישור:  $\{x \mid w^\top x \leq b\} \subseteq \mathbb{R}^d, w \in \mathbb{R}^d, b \in \mathbb{R}$
- תת-מרחב אפיני:  $\{x \mid Ax = b\} \subseteq \mathbb{R}^d, A \in \mathbb{R}^{d \times n}, b \in \mathbb{R}^d$
- קבוצת כל המטריצות המשויות הסמטריות וה- PSD מוגדל  $n \times n$ :  $\mathbb{S}_+^n \subseteq \mathbb{R}^{n^2}$

#### 10.2.3 על מישור מפheid

אם  $C, D \subseteq \mathbb{R}^d, b \in \mathbb{R}$  שתי קבוצות זרות, קמורות ולא ריקות ב-  $\mathbb{R}^d$ , אז קיים על מישור המורכב מ-  $a \in \mathbb{C}$  ש-  $C \subseteq \{x \mid a^\top x \leq b\}, D \subseteq \{x \mid a^\top x \geq b\}$

כלומר על המישור מפheid בין שתי הקבוצות.

#### 10.2.4 פעולות משמרות קמירויות

- אם  $C, D \subseteq \mathbb{R}^d$  קבוצות קמורות, אז הקבוצות הבאות גם קמורות:
- החיתוך  $C \cap D$  ◦
  - הزاיה והכפלה בסקלר:  $\{ax + b \mid a \in \mathbb{R}, b \in \mathbb{R}^d\}$  ◦
  - העתקה אפינית:  $f(C), f^{-1}(C)$  ◦

### 10.2.5 פונקציות קמורות

תהי  $C \subseteq \mathbb{R}^d$  קבוצה קמורה.

הfonקציה  $f : C \rightarrow \mathbb{R}$  נקראת **פונקציה קמורה** אם לכל  $\mathbf{u}, \mathbf{v} \in C$  וסקלר  $0 \leq \alpha \leq 1$  מתקיים:

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v})$$

הfonקציה  $f$  תקרא **קמורה ממש** אם אי-השוויון יתקיים לכל  $\mathbf{u} \neq \mathbf{v} \in \mathbb{R}^d$  ו-  $0 < \alpha < 1$

- פונקציה  $f$  היא קמורה אם ורק אם  $\text{epi}(f)$  אוסף כל הנקודות שמעל הפונקציה.

$$\text{epigraph}(f) = \{(\mathbf{x}, \beta) \mid f(\mathbf{x}) \leq \beta\}$$

### 10.2.6 פונקציות קמורות חשובות

נכתב  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  עבור הקורידינטות של  $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ . הפונקציות של  $x$  הבאות הן קמורות:

◦  $x \mapsto -\log(x)$ , ( $a \leq 0$  או  $a \geq 1$ ) עבור  $x \mapsto x^a$ , ( $a \geq 1$  עבור  $x \mapsto e^{ax}$ ).

◦ פונקציה אפינית  $\mathbf{w}^\top \mathbf{x} + b$ .

◦ תבנית ריבועית  $\alpha \in \mathbb{R}$  ו-  $\mathbf{w} \in \mathbb{R}^d$ ,  $A \in \mathbb{S}_+^n$  עבור  $\mathbf{x} \mapsto \mathbf{x}^\top A \mathbf{x} + \mathbf{w}^\top \mathbf{x} + \alpha$ .

◦ סכום הריבועים  $\|\mathbf{y} - Ax\|_2^2$  לכל מטריצה  $A$ .

◦ כל נורמות  $\ell_p$  עבור  $\mathbf{x} \mapsto \left( \sum_{i=1}^d x_i^p \right)^{\frac{1}{p}}$  :  $p \geq 1$ .

◦ נורמת האינסוף  $\mathbf{x} \mapsto \max_i \{|x_i|\}$ .

◦ פונקציית מקס  $\mathbf{x} \mapsto \max_i \{x_i\}$ .

◦ פונקציית אינדיקטור  $C \subseteq \mathbb{R}^d$  עבור  $\mathbf{x} \mapsto \mathbb{1}_C(\mathbf{x})$  קמורה.

### 10.2.7 אפיון קמירות

תהי  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ . נזכיר כי הגרדיינט  $\nabla f(\mathbf{x})$  וההסיאן  $\nabla^2 f(\mathbf{x})$  של  $f$  ב-  $\mathbf{x} \in \mathbb{R}^d$  הם הווקטור המטריצה הבאים:

$$\nabla f(\mathbf{x})_i =: \frac{\partial f}{\partial x_i} \quad \left| \quad \nabla^2 f(\mathbf{x})_{i,j} =: \frac{\partial^2 f}{\partial x_i \partial x_j} \right.$$

**תנאי מסדר ראשון:** נניח כי  $f$  דיפרנציאבילית. אז  $f$  קמורה אם ורק אם  $\text{dom}(f) \subseteq \mathbb{R}^d$  קבוצה קמורה וגם:

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in \text{dom}(f) : f(\mathbf{x}_2) \geq f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^\top (\mathbf{x}_2 - \mathbf{x}_1)$$

**תנאי מסדר שני:** נניח כי  $f$  דיפרנציאבילית פעמיים. אז  $f$  קמורה אם ורק אם  $\text{dom}(f) \subseteq \mathbb{R}^d$  קבוצה קמורה וגם:

$$\forall \mathbf{x} \in \text{dom}(f) : \nabla^2 f(\mathbf{x}) \in \mathbb{S}_+^n$$

### 10.2.8 פעולות נשמרות קמירות (על פונקציות)

- צירוף ליניארי אי-שלילי ומקסימום: אם  $a_1, \dots, a_m \geq 0$  (עבור  $f_1, \dots, f_m$  פונק' קמורותizi) אז  $\sum_i a_i f_i$  פונק' קמורה. אם  $\max_i f_i$  פונק' קמורות.
- מזעור חלק: נניח  $\mathbb{R}^d \rightarrow \mathbb{R}$ :  $g$ , המוגדרת ע"י  $\mathbb{R}^{d+k} \rightarrow \mathbb{R}$  עבור  $x_1 \in \mathbb{R}^d, x_2 \in \mathbb{R}^k$  ( $x_1, x_2 \mapsto g(x_1, x_2)$ ) אז  $x_1$  היא קמורה. אז המזעור החלקי  $\mathbb{R}^d \rightarrow \mathbb{R}$ :  $h(x_1) = \min_{x_2 \in C} g(x_1, x_2)$  (עבור  $C \subseteq \mathbb{R}^k$  קבועה קמורה) היא פונק' קמורה.
- אם  $f$  פונק' קמורהizi אז  $\mapsto x$  היא פונק' קמורה.
- אם  $\mathbb{R}^d \rightarrow \mathbb{R}$ :  $g$  פונק' קמורה ו-  $\mathbb{R} \rightarrow \mathbb{R}$ :  $h$  פונק' קמורה ומונוטונית לא יורדת אז ההרכבה  $\mapsto x \mapsto h(g(x))$  פונק' קמורה.
- אם  $\mathbb{R}^k \rightarrow \mathbb{R}$ :  $h$  פונק' קמורה ומונוטונית לא יורדת בכל אחת מהקורדיינאות שלה, ו-  $\mathbb{R}^d \rightarrow \mathbb{R}$ :  $g_i$  פונק' קמורה. לכל  $i \in [k]$ , אז  $\mathbb{R}^d \rightarrow \mathbb{R}$ :  $f(x) = h(g_1(x), \dots, g_k(x))$  היא פונק' קמורה.

### דוגמאות 10.2.9

1. נתבונן בפונק'  $\mathbb{R}^d \rightarrow \mathbb{R}$  המוגדרת ע"י:

$$f(\mathbf{x}) = \log \left( \sum_{i=1}^k e^{\mathbf{w}_i^\top \mathbf{x} + b_i} \right)$$

עבור  $\mathbf{w}_i \in \mathbb{R}^d$  ו-  $b_i \in \mathbb{R}$  קבועים לכל  $i \in [k]$ . נקראת log-sum-exp או soft-max.

ניתן להראות באמצעות רשיימת הפונק' הקמורות ופעולות נשמרות קמירות כי הפונק' זו קמורה:

(תרגול 13) מכיוון שהרכבה אפינית נשמרת קמירות, נבחן כי הפונק' עליה אנו מרכיבים את העתקה האפינית

$$\log \left( \sum_{i=1}^m e^{z_i(\mathbf{w})} \right)$$

נוכל להכניס את הקורדיינאות  $z_i = (z_1, \dots, z_m) \in \mathbb{R}^m$  לקטור  $\mathbf{z}$ , ולהראות כי היא קמורה. תכונה נחמדה של LogSumExp היא שהגרדיינט שלה היא פונק' ה-Softmax:

$$\nabla f(\mathbf{z}) = S(\mathbf{z}) = \begin{pmatrix} \frac{e^{z_1}}{\sum\limits_{j=1}^m e^{z_j}} \\ \vdots \\ \frac{e^{z_m}}{\sum\limits_{j=1}^m e^{z_j}} \end{pmatrix}$$

ההסיאן של  $f$  היא המטריצה  $H \in \mathbb{R}^{m \times m}$  והיא מקיימת:

$$H = \nabla S = \text{diag}(S) - SS^\top \iff H_{i,j} = \partial_i S_j = \delta_{i,j} S_i - S_i S_j$$

אם נוכל להראות כי  $H$  היא מטריצה PSD נקבל כי  $f$  קמורה. בambilים אחרים נרצה להראות כי  $0$

נסמן:  $C = \sum_{j=1}^m e^{z_j}$  ו-  $\eta_i = e^{z_i}$

$$\begin{aligned} \mathbf{y}^\top H \mathbf{y} \geq 0 &\iff \mathbf{y}^\top (\text{diag}(S) - SS^\top) \mathbf{y} \geq 0 \\ &\iff \mathbf{y}^\top \text{diag}(S) \mathbf{y} \geq \mathbf{y}^\top SS^\top \mathbf{y} \\ &\iff \frac{1}{C} \cdot \sum_{i=1}^m \eta_i y_i^2 \geq (S^\top \mathbf{y})^2 = \left( \frac{1}{C} \cdot \sum_{i=1}^m \eta_i y_i \right)^2 // \cdot C^2 \\ &\iff C \cdot \sum_{i=1}^m \eta_i y_i^2 \geq \left( \sum_{i=1}^m \eta_i y_i \right)^2 \\ &\iff \left( \sum_{i=1}^m \eta_i \right) \left( \sum_{i=1}^m \eta_i y_i^2 \right) \geq \left( \sum_{i=1}^m \eta_i y_i \right)^2 \end{aligned}$$

כאשר המשוואה الأخيرة היא נכונה ע"פ אי-שוויון קושי-שوارץ:

$$\left( \sum_i a_i^2 \right) \left( \sum_i b_i^2 \right) \geq \left( \sum_i a_i b_i \right)^2$$

ונקבל כי אי-השוויון מתקיים. על כן הפונקציה LogSumExp קמורה, כנדרש.

נציב:  $b = \begin{pmatrix} y_1 \sqrt{\eta_1} \\ \vdots \\ y_m \sqrt{\eta_m} \end{pmatrix}$  ו-  $a = \begin{pmatrix} \sqrt{\eta_1} \\ \vdots \\ \sqrt{\eta_m} \end{pmatrix}$

2. אם נזכיר בפונק' המטריה של Hard-SVM:

$$f(\mathbf{w}) = \min_{i \in [m]} |\langle \mathbf{w} | \mathbf{x}_i \rangle|$$

עבור  $\mathbf{x}$  קבועים, נוכל להוכיח כי  $f$  היא פונק' קעורה (כלומר  $-f$  היא פונק' קמורה).

3. גם פונק' המטריה ברגression לוגיסטייה ממקסמת את הנראות:

$$f(\mathbf{w}) = \sum_{i=1}^m [y_i \langle \mathbf{x}_i | \mathbf{w} \rangle - \log(1 + e^{\langle \mathbf{x}_i | \mathbf{w} \rangle})]$$

עבור  $\mathbf{x}$  קבועים היא קעורה - ראיינו בהרצאה את הצורה הבאה לעביה:

$$L(\mathbf{w} | \mathbf{y}) = \prod_{i=1}^m p_i(\mathbf{w})^{y_i} \cdot (1 - p_i(\mathbf{w}))^{1-y_i}$$

כאשר ההסתברויות מוגדרות ע"י (פונק' הסיגמוד)  $p_i(\mathbf{w}) = \sigma(x_i^\top \mathbf{w}) = \frac{1}{1 + e^{-x_i^\top \mathbf{w}}}$

נסמן  $w$  ונקבל כי:

$$\sigma(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

מכאן שה- LNN נראה כך:

$$-f(\mathbf{w}) = -\sum_{i=1}^m [y_i \log(\sigma(\mathbf{x}_i^\top \mathbf{w})) + (1-y_i) \log(1-\sigma(\mathbf{x}_i^\top \mathbf{w}))]$$

ומספריק להראות כי  $f$  – קמורה כפונקציה של  $z$ , מכיוון שהן פונק' אפייניות של  $\mathbf{w}$ . נגיד רשותי פונק' סקלריות:

$$(1) : f_1(z) = -\log(\sigma(z)) \quad (2) : f_2(z) = -\log(1-\sigma(z))$$

מכיוון ש- $f$  – היא קומבינציה ליניארית "ש" של הפונקציות הנ"ל, אם נוכיח כי שתיהן קמורות נקבל כי  $f$  – קמורה בסכום פונק' קמורות.

$$f_1(z) = -\log\left(\frac{1}{1+e^{-z}}\right) = -(\log 1 - \log(1+e^{-z})) = \log(1+e^{-z})$$

$$\frac{d}{dz}f_1(z) = -\frac{e^{-z}}{1+e^{-z}} = -1 + \frac{1}{1+e^z} = -1 + \sigma(z)$$

הנגזרת הראשונה של  $f_1(z)$  היא מונוטונית עולה ( $\sigma$  מונ' עולה) ועל כן הנגזרת שלה חיובית תמיד, ועל כן הנגזרת השנייה של  $f_1(z)$  חיובית תמיד ו-  $f_1$  קמורה.

$$f_2(z) = -\log\left(1 - \frac{1}{1+e^{-z}}\right) = -\log\left(\frac{e^{-z}}{1+e^{-z}}\right) = -(\log(e^{-z}) - \log(1+e^{-z})) = z + \log(1+e^{-z}) = z + f_1(z)$$

$$\frac{d}{dz}f_2(z) = 1 + f'_1(z)$$

ושוב, מכיוון שהנגזרת הראשונה של  $f_1$  מונוטונית עולה נקבל כי הנגזרת השנייה של  $f_2$  גם היא חיובית תמיד, ועל כן  $f_2$  קמורה גם היא. מכאן שה-  $f$  –  $(\mathbf{w})$  קמורה, ועל כן  $(\mathbf{w})$   $f$  קעורה.

#### 10.2.10 תכונה חשובה ראשונה | מין' לוקאלי הוא גם גלובי

אם  $f$  קמורה, אז כל נקודת מינימום לוקאלי של  $f$  הוא גם מינימום גלובי.

##### הוכחה

$$\text{יהי } B(\mathbf{u}, r) = \{\mathbf{v} \mid \|\mathbf{v} - \mathbf{u}\| \leq r\}.$$

( $\mathbf{u}$ )  $f(\mathbf{v}) \geq f(\mathbf{u})$  מינימום לוקאלי של  $f$  ב-  $\mathbf{u}$  אם קיימים  $\mathbf{v} \in B(\mathbf{u}, r)$  מתקיים  $\mathbf{v} \neq \mathbf{u}$  ו-  $f(\mathbf{v}) > f(\mathbf{u})$ . לפיה הגדרת הקמירות, לכל  $\mathbf{v}$  (לא בהכרח ב- $B(\mathbf{u}, r)$ ) קיימים  $0 < \alpha < 1$  כך ש-  $\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u}) \in B(\mathbf{u}, r)$  ומכאן:

$$f(\mathbf{u}) \leq f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u}))$$

בנוסף מהגדרת הקמירות נקבל כי:

$$f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u})) = f(\alpha\mathbf{v} + (1-\alpha)\mathbf{u}) \leq (1-\alpha)f(\mathbf{u}) + \alpha f(\mathbf{v})$$

וע"י איחוד שתי המשוואות נקבל כי  $f(\mathbf{u}) \leq f(\mathbf{v})$ . זה מתקיים לכל  $\mathbf{v}$ , ועל כן  $f(\mathbf{u})$  הינה מינימום גלובי של  $f$ .

**10.2.11 תכונה חשובה שנייה | יחס בין פונק' למישור משיק**

זכור בקירוב טילור מסדר ראשון של  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ . אם  $f$  היא דיפרנציאבילית ב-  $w$ , אז עבור  $u$  קרוב ל-  $w$ , ניתנת לשערוך ע"י פונק' ליניארית:

$$f(u) \approx f(w) + \langle \nabla f(w) \mid u - w \rangle$$

אם  $f$  קמורה ודיפרנציאבילית, אז  $f$  תמיד יותר גדולה מהקירוב הלייניארי שלה. באופן פורמלי:

$$\forall u : f(u) \geq f(w) + \langle \nabla f(w) \mid u - w \rangle$$

**10.3 תת-גרדיינטים Sub-gradients**

כאשר פונקציה  $f$  היא קמורה אך לא דיפרנציאבילית בנקודה  $x$ , אין לנו גרדיינט  $\nabla f(x)$ . אך יש לנו משהו אחר.

**10.3.1 תת-גרדיינט**

ש הוא **תת-גרדיינט** של  $f$  בנקודה  $w$  אם:  $\langle v \mid u - w \rangle$ .

**10.3.2 תת-הדיפרנציאל**

תת-הדיפרנציאל הוא קבוצת כל תת-הגרדיינטים ( $w$ ) של  $f$  בנקודה  $w$ .

**10.3.3 למה**

פונקציה  $f$  היא קמורה אם ורק אם לכל  $w$  מתקיים כי  $\partial f(w) \neq \emptyset$ .

**10.3.4 דוגמאות**

הפונקציה  $f : \mathbb{R} \rightarrow \mathbb{R}$  המוגדרת ע"י  $f(x) = |x|$  לכל  $x \neq 0$  מתקיים כי  $\partial f(x) = [-1, 1]$ . עבור  $x = 0$  מתקיים כי  $\partial f(x) = \{\text{sign}(x)\}$ .

הפונקציה  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  המוגדרת ע"י  $f(x) = \|x\|_1$  לכל  $x \neq 0$  מתקיים כי  $\partial f(x) = \{\text{sign}(x_1), \dots, \text{sign}(x_d)\}$ . עבור  $x = 0$  מתקיים כי  $\partial f(x) = \{(\text{sign}(x_1), \dots, \text{sign}(x_d))\}$ .

הפונקציה  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  המוגדרת ע"י  $f(x) = \|x\|_2$  לכל  $x \neq 0$  מתקיים כי  $\partial f(x) = \left\{ \frac{x}{\|x\|_2} \right\}$ . עבור  $x = 0$  מתקיים כי  $\partial f(x) = \{0\}$ .

### 10.3.5 תכונות של תת הדיפרנציאלי

- היא קבוצה קמורה וסגורה (גם עבור פונקציות לא קמורות).

$$\text{אם } f \text{ דיפרנציאבילית בנקודה } x \text{ אז } \partial f(x) = \{\nabla f(x)\}.$$

- בכיוון השני - אם  $x$  היא נקודת בת הדיפרנציאלי הוא ייחדו  $\nabla f(x) = \{g\}$ .

### 10.3.6 חישובו תת-גרדיינטים

$$\text{אם } \alpha > 0 \text{ אז } \partial(\alpha f) = \alpha \cdot \partial f.$$

$$\text{חיבור: } \partial(f_1 + f_2) = \partial(f_1) + \partial(f_2).$$

$$\text{הרכבה אפינית: אם } g(x) = f(Ax + b) \text{ אז } \partial g(x) = A^\top \partial f(Ax + b).$$

## 10.4 אופטימיזציה קמורה

בעיה אופטימיזציה קמורה ב-  $\mathbb{R}^d$  היא בעיה מהצורה:

$\begin{array}{ll} \text{minimize}_{\mathbf{x} \in D} & f(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq b_i \quad i = 1, \dots, n \end{array}$
--

כאשר  $D = \text{dom}(f) \cap \left( \bigcap_{i=1}^d \text{dom}(f_i) \right)$  כולם פונקציות קמורות. תחום הבעיה  $D$  לרוב מוגדר ע"י  $Ax = b$  יתכן בנוספ' אילוצים ליניארים מהצורה

- נקודת  $x \in D$  חמיימת  $g_i(x) \leq b_i$  נקראת **נקודת פיזיבilitiy**.
- הערך האופטימלי של  $f$  מעל כל הנקודות הפיזיבליות נקרא **הערך האופטימלי**, ומסומן  $f^*$ .
- אם  $x$  היא נקודת פיזיבilitiy וגם  $f(x) = f^*$  נקרא **נקודת האופטימלית, הפתרון לבעה או המוצע**.
- קובצת כל הפתרונות של בעיה אופטימיזציה קמורה היא קבוצה קמורה.
- אם נסמן ע"י  $C \subseteq \mathbb{R}^d$  את אוסף כל הנקודות הפיזיבליות, נוכל לכתוב את בעיה האופטימיזציה ע"י  $\text{minimize } f(\mathbf{x}) \text{ subject to } \mathbf{x} \in C$ .

### 10.4.1 אלגוריתמים עבור אופטימיזציה קמורה

פותר אופטימיזציה קמורה הוא אלגוריתם שמקבל כקלט בעיה אופטימיזציה קמורה ומחזיר את הפתרון האופטימלי. פותרי אופטימיזיות קמורויות הוא נושא ענק, וגישה עבורה מתחולקות לשיטות מסדר ראשון, המשמשות בגרדיינט או תת-גרדיינט של פונקציית המטריה, ושיטות מסדר שני שימושות גם במטריצת ההessian של פונק' המטריה.

## 10.5 בעיות למידה קמורות

נרצה לקשור כעט בין בעיות אופטימיזציה קמורות לבין למידה חישובית.  
השאלה הראשונה אותה נשאל היא - האם בעית למידה שנייה להמיר לעית אופטימיזציה קמורה היא למידה PAC?

נזכיר איך במקרה של חצאי מישור ושל רגרסיה הצלחנו להגיע לאופטימיזציה קמורה.  
נניח  $\mathcal{H}$  מחלוקת היפוטזות כך שכל היפוטזה  $\mathcal{U} \rightarrow \mathcal{X} : h \in \mathbb{R}^d$ . לדוגמה מחלוקת ההיפוטזות:

$$HS_d = \left\{ \mathbf{x} \mapsto \text{sign}(\langle \mathbf{w} | \mathbf{x} \rangle) \mid \mathbf{w} \in \mathbb{R}^d \right\}$$

לשם נוחות, נזהה כל היפוטזה כוקטור ונכתב  $\mathbb{R}^d \subseteq \mathcal{H}$ .

### 10.5.1 הגדרה | בעית למידה קמורה

תהי  $\mathcal{U} \times \mathcal{X} = Z$ . בעית למידה  $(\mathcal{H}, Z, \ell)$  נקראת קמורה אם מחלוקת היפוטזות  $\mathcal{H}$  היא קבוצה קמורה  
ולכל  $Z \in \mathcal{Z}$  פונק' ההפסד  $(\cdot, z)$  היא פונק' קמורה.  
( $f(\mathbf{w}) = \ell(\mathbf{w}, z)$  המוגדרת ע"י  $f : \mathcal{H} \rightarrow \mathbb{R}$  היא פונק'  $(\cdot, \cdot)$  הינה פונק' פונקצייתית).

**אבחנה מרכזית:** בעית למידה קמורה, שימוש בעקרון ERM מורייד בעית אופטימיזציה קמורה:

$$\min_{\mathbf{w} \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \ell(\mathbf{w}, z_i)$$

### 10.5.2 טענה

לא כל בעית למידה קמורה היא למידה PAC. אך אם נוסיף את שני התנאים הבאים, נקבל למידות PAC:

- המחלוקת  $\mathcal{H}$  חסומה.

- פונקציית ההפסד  $\ell$  היא פונק' ליפשיצית, כאשר פונקציה  $f : C \rightarrow \mathbb{R}$  היא  $\rho$ -ליפשיצית אם לכל  $\mathbf{w}_1, \mathbf{w}_2 \in C$  מתקיים:

$$|f(\mathbf{w}_1) - f(\mathbf{w}_2)| \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|$$

### 10.5.3 הגדרה | בעית למידה קמורה ליפשיצית וחסומה

בעית למידה  $(\mathcal{H}, Z, \ell)$  היא בעית למידה קמורה, ליפשיצית וחסומה עם פרמטרים  $B, \rho$  אם מתקיים:

1. מחלוקת היפוטזות  $\mathcal{H}$  היא קבוצה קמורה כך שכל  $\mathcal{H} \in \mathcal{H}$  מתקיים כי  $\|\mathbf{w}\| \leq B$ .
2. לכל  $Z \in \mathcal{Z}$  פונק' ההפסד  $(\cdot, z)$  היא קמורה ו-  $\rho$ -ליפשיצית.

**10.5.4 טענה מרכזית**

כל בעית למידה קמורה, לפטישית וחסומה היא למידה PAC, עם סיבוכיות מודל שתלויה ב-  $\rho, \delta, B, \varepsilon$ .

**10.6 אלג' Gradient Descent****10.6.1 תזכורת - תכונות הגרדיינט**

- אם  $f$  דיפרנציאבילית בנקודה  $x$ , אז  $(\nabla f(x))^\top$  מצביע בכיוון העליה התלויה ביוטר, ו-  $-\nabla f(x)$  מצביע בכיוון הירידה התלויה ביוטר.
- עבור  $x \in \mathbb{R}^d$ , נסמן ב-  $L_x(f) = \{x' \mid f(x) = f(x')\}$  את ה-level set של  $f$ . קבוצת כל הנקודות  $x'$  בהן  $f$  בעל אותו ערך.
- נניח כי  $f$  דיפרנציאבילית בנקודה  $x$ , וכי  $w$  הישר המשיק ל-  $L_x(f)$  ב-  $x$ . אז  $0 = \langle \nabla f(x) \mid w \rangle$ . (אנכיים).

**10.6.2 תנאי אופטימליות מסדר ראשון**

עבור בעית אופטימיזציה קמורה מהצורה  $\text{minimize } f(x) \text{ subject to } x \in C$  כאשר  $f$  דיפרנציאבילית, נקודה פיזיבלית  $x$  היא אופטימלית אם ורק אם מתקיים:

$$\forall y \in C : \langle \nabla f(x) \mid y - x \rangle \geq 0$$

או במלים, לכל  $h$ , אם  $h + x$  (הוקטור המצביע מ-  $x$  בכיוון של  $h$ ) הוא פיזיבלי, אז  $h$  לא מכיל רכיב בכיוון של  $-\nabla f(x)$ .  
מקרה מיוחד: אם  $C = \mathbb{R}^d$  (אין אילוצים), אז תנאי זה גורר כי  $x$  הוא אופטימלי אם ורק אם  $(\nabla f(x))^\top h = 0$  לכל  $h \in \mathbb{R}^d$ .

נבחן כי אם  $x$  אינו אופטימלי אז קיים  $h$  כך ש-  $h + x$  פיזיבלי ו-  $0 < \langle \nabla f(x) \mid h \rangle$ .  $h$  נקרא כיוון ירידה מ-  $x$ .  
אם ניקח  $h$  קטן מספיק, נקבל  $\langle \nabla f(x + h) \mid h \rangle < \langle \nabla f(x) \mid h \rangle$ . אך נשאלת השאלה - באיזה כיוון נלך?  
אם יש איזשהו כיוון ירידה, אז  $\nabla f(x)$  הוא כיוון ירידה. השאלה היא מהו גודל הצעד?  
אם הצעד יהיה קטן מדי - אנחנו לא מתקדמים מספיק. אם הצעד גדול מדי - אנו עלולים לעבור את האופטימום!

### 10.6.3 האלגוריתם

ראשית נעבד עם אופטימיזציה קמורה ללא אילוצים. נתחל עם וקטור  $\mathbf{x}^{(1)} \in \mathbb{R}^d$  כלשהו. באיטרציה ה- $t$ , נעדכן:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \nabla f(\mathbf{x}^{(t)})$$

ונעזור בזמן  $T$  כלשהו (תנאי עצירה סביר: כאשר  $\|\nabla f(\mathbf{x}^{(t)})\|$  קטן מערך סף מסוימים). ניתן להוציא לפلت את  $\mathbf{x}^{(T)}$ , את הווקטור הממוצע  $\bar{\mathbf{x}} = \frac{1}{T} \sum_{t=1}^T \mathbf{x}^{(t)}$  או את הווקטור בעל הביצועים הטובים ביותר עבור  $\mathbf{x}^{(t^*)}$ . ערכו  $\eta_t$  נקראים גודל הצעד.

$$\mathbf{x}^{(t^*)} = \underset{1 \leq t \leq T}{\operatorname{argmin}} f(\mathbf{x}^{(t)})$$

### 10.6.4 כיצד להבין את השיטה

נזכיר בקירוב טיילור מסדר שני של  $f$  סביב נקודת  $\mathbf{x}$  כלשהו:

$$f(\mathbf{y}) \approx f(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x} \mid \nabla f(\mathbf{x}) \rangle + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x}) (\mathbf{y} - \mathbf{x})$$

שיטת GD היא שיטה מסדר ראשון ואנו לא מניחים כי  $f$  דיפרנציאbilית פערמיים או שאנו יודעים את ההסיאן  $\nabla^2 f(\mathbf{x})$ . לכן אנחנו נחליף אותו ב-  $I \frac{1}{\eta}$  עבור  $I$  מטריצת הזהות:

$$f(\mathbf{y}) \approx f(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x} \mid \nabla f(\mathbf{x}) \rangle + \frac{1}{2\eta} \|\mathbf{y} - \mathbf{x}\|^2$$

### 10.6.5 כיצד נבחר את גודל הצעד?

בעזרת Backtracking Line Search לבחירת  $\eta_t$  אַדְפְּטִיבִי ב-GD.

נסמן:  $\Delta \mathbf{x} = -\nabla f(\mathbf{x})$ . צעד הירידה שלנו יוזם מ-  $\mathbf{x}$  ל-  $\mathbf{x} + \eta \Delta \mathbf{x}$ .

נרצה לבחור  $\eta$  גדולה מספיק על מנת לקבל ירידה ממשמעותית בערכו של  $f$ , אך לא גדולה מדי.

נראה שיטה נחמדה למציאת  $\eta$  מוגدل נכון: נקבע פרמטר  $\alpha < \frac{1}{2}$  כלשהו.

- עבור  $0 > \eta$  קרוב לאפס, מתקיים:  $f(\mathbf{x}) + \alpha \eta \langle \nabla f(\mathbf{x}) \mid \Delta \mathbf{x} \rangle > f(\mathbf{x} + \eta \Delta \mathbf{x})$
- עבור  $0 > \eta$  גדול מספיק, מתקיים:  $f(\mathbf{x}) + \alpha \eta \langle \nabla f(\mathbf{x}) \mid \Delta \mathbf{x} \rangle < f(\mathbf{x} + \eta \Delta \mathbf{x})$
- מכאן שחייב להיות ערך  $0 > \eta_0$  עבורו מתקיים:  $f(\mathbf{x}) + \alpha \eta_0 \langle \nabla f(\mathbf{x}) \mid \Delta \mathbf{x} \rangle = f(\mathbf{x} + \eta_0 \Delta \mathbf{x})$

כיצד נשערך את  $\eta_0$  באופן מהיר? נקבע פרמטר נוסף,  $\beta < 1$ . נריץ את הלולאה הפשוטה הבאה: נתחל מ-  $\eta = 1$ , ובכל איטרציה נשנה  $\eta \mapsto \beta \eta$  כל עוד

$$f(\mathbf{x}) + \alpha \eta \langle \nabla f(\mathbf{x}) \mid \Delta \mathbf{x} \rangle < f(\mathbf{x} + \eta \Delta \mathbf{x})$$

הפרמטר הראשון שיופיע את התנאי הוא הערך הנבחר.

### 10.6.6 האם האלגוריתם בהכרח מתכנס לנקודה אופטימלית?

נתבונן בניתוח התכונות הפשטוט ביותר בעבר אלגוריתם GD: נתבונן בניתוח התכונות הפשטוט ביותר בעבר אלגוריתם GD: נניח כי  $f$  היא קמורה וdifrenzialibilitiy עם  $\nabla f(\mathbf{x}) = \mathbb{R}^d$ . נניח כי  $\nabla f(\mathbf{x})$  היא לפישיצית עם קבוע  $L$ :

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d : \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2$$

יהי  $f^*$  הערך האופטימלי של  $f$ , ויהי  $\mathbf{x}^*$  הנקודה האופטימלית. אז האל' GD עם גודל צעד קבוע  $\eta$  מקיים:

$$f(\mathbf{x}^{(t)}) - f^* \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\eta \cdot t}$$

ועל כן האלגוריתם אכן מתכנס לנקודה אופטימלית כאשר  $t \rightarrow \infty$ .

### 10.6.7 בעיות אופטימיזציה עם אילוצים

האם נוכל להשתמש באל' GD על מנת לפתור בעיות אופטימיזציה קמורה עם אילוצים? התשובה היא כן. נתבונן שוב בצורה הכללית של בעיית אופטימיזציה קמורה:

$$\text{minimize } f(\mathbf{w}) \text{ subject to } \mathbf{w} \in C$$

יהי  $P_C$  אופרטור ההטלה על הקבוצה הפיזיבלית  $C$  (כאשר נזכיר כי אם הבעיה קמורה אז הקבוצה  $C$  קמורה). האופרטור  $P_C$  מוגדר ע"י (משמעותו לב שזוויות בעצם בעיה קמורה):

$$P_C(\mathbf{x}) = \operatorname{argmin}_{\mathbf{w} \in C} \|\mathbf{x} - \mathbf{w}\|_2$$

שיטת GD מורכבת מהאייטרציה הבאה:

$$\mathbf{x}^{(t+1)} = P_C(\mathbf{x}^{(t)} - \eta_{t+1} \nabla f(\mathbf{x}^{(t)}))$$

כלומר לאחר כל צעד אנו מטילים חזרה על הקבוצה הפיזיבלית  $C$ .

אל' זה הוא יעיל אם  $C$  היא כזו ש-  $P_C(\mathbf{x})$  חסיבה באופן מהיר.

## 10.7 אל' Sub-gradient Descent

מה נעשה כאשר  $f$  אינה דיפרנציאבילית בנקודה  $\mathbf{x}$ ?

נזכיר כי כאשר  $f$  קמורה אז תמיד  $\emptyset \neq \partial f(\mathbf{x}) \subseteq \text{dom}(f)$  לכל  $\mathbf{x} \in \text{dom}(f)$ .

נניח ונוכל למצוא את תת-גרדיינט של  $f$  בנקודה  $\mathbf{x}$ . האם נוכל להשתמש בו עבור ?

### 10.7.1 תנאי אופטימליות למת-גרדיינט

לכל פונקציה  $f$  (קמורה או לא),  $f(\mathbf{x}^*) = \min f(\mathbf{x})$  אם ורק אם  $0 \in \partial f(\mathbf{x}^*)$ . מה? נזכיר בהגדירה של תת-גרדיינט:  $\forall \mathbf{u} : f(\mathbf{u}) \geq f(\mathbf{x}) + \langle \mathbf{v} | \mathbf{u} - \mathbf{x} \rangle$  אם  $\mathbf{v} \in \partial f(\mathbf{x})$ . נבחן כי אם  $f$  דיפרנציאבילית בנקודה  $\mathbf{x}^*$ , אז מצד אחד  $0 \in \partial f(\mathbf{x}^*)$  ומצד שני  $0 \in \{\nabla f(\mathbf{x}^*)\}$

### 10.7.2 האלגוריתם

אלג' - Sub-gradient Descent הוא אלג' מסדר ראשון של בעיית אופטימיזציה קמורה, המתאים לבעיות קמורות עם פונק' מטרה לא דיפרנציאבילית. הרעיון הוא פשוט - נחליף את האיטרציה של GD:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \nabla f(\mathbf{x}^{(t)})$$

באייטרציה הבאה:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \mathbf{g}^{(t)}$$

כאשר  $\mathbf{g}^{(t)} \in \partial f(\mathbf{x}^{(t)})$  הוא תת גרדיאנטו כלשהו של  $f$  ב-  $\mathbf{x}^{(t)}$ .

נרצה גם כן לעזר בזמן  $T$  כלשהו, ולכן נשאל מה הוא תנאי עצירה סביר? \*שאלה ל Kohra\*: התת גרדיאנט של  $f$  בנקודה  $\mathbf{x}$  לא בהכרח בכיוון ירידה, ועל כן אין הגיון בלחשior את הוקטור האחרון  $\mathbf{x}^{(t)}$ . נחזיר או את הוקטור הממוצע, או את הוקטור בעל הביצועים הטובים ביותר.

- ניתן להראות כי אלג' הפרספטורן הוא בעצם אלג' sub-gradient descent על פונקציית המרגינן.

### 10.7.3 גודל הצעד

מכיוון שתת הגרדיאנט של  $f$  בנקודה  $\mathbf{x}$  אינו בהכרח בכיוון ירידה, אין לנו שיטה אדפטיבית לבחירת גודל הצעד, ועל כן הוא יקבע מראש. לרוב נבחר סדרה  $\{\eta_t\}_{t=1}^{\infty}$  המתכנסת לאפס אך לא מהר מדי. לדוגמה סדרה כזו שמקיימת:

$$\sum_{t=1}^{\infty} \eta_t^2 < \infty \quad \text{but} \quad \sum_{t=1}^{\infty} \eta_t = \infty$$

### 10.7.4 האם האלג' בהכרח מתכנס לנקודת אופטימלית?

מכיוון שהכיון בזמן  $t$  אינו בהכרח בכיוון ירידה, לא ברור האם האלג' מתכנס לפתרון אופטימיili. משפט

נניח כי  $f$  פונק' קמורה עם  $\text{dom}(f) = \mathbb{R}^d$ . נניח כי  $f$  ליפשיצית עם קבוע  $L$ :

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d : \|f(\mathbf{x}) - f(\mathbf{y})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2$$

יהי  $f^*$  הערך האופטימי של  $f$  ו-  $\mathbf{x}_{best}^{(t)}$  הוקטור בעל הביצועים הטובים ביותר מבין  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}$ . אזי תת הגרדיאנט בעל גודל צעד שהולך וקטן הוא:

$$\lim_{t \rightarrow \infty} f(\mathbf{x}_{best}^{(t)}) = f^*$$

## 11 הרצאה 11 - SGD ולמידה عمוקה (רשתות נוירוניים)

### 11.1 Stochastic Gradient Descent

נסמן ב-  $G$  וקטור מקרי המקבל ערכים מ-  $\mathbb{R}^d$ , כך שהתחולת ( $G$ ) הוא וקטור קבוע ב-  $\mathbb{R}^d$ .

#### 11.1.1 הגדרה

האיטרציה  $\mathbf{g}^{(t)} \stackrel{\text{indep.}}{\sim} G^{(t)}$  נקראת איטרצית SGD עבור  $f$  אם לכל  $t$  מתקיים  $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \mathbf{g}^{(t)}$ .  $\mathbf{g}^{(t)} \in \partial f(\mathbf{x}^{(t)})$  והוא וקטור מקרי עם  $\mathbb{E}(G^{(t)}) \in \partial f(\mathbf{x}^{(t)})$ . נבחן כי אם  $f$  גזירה ב-  $\mathbf{x}^{(t)}$  אז מתקבל  $\mathbb{E}(G^{(t)}) = \nabla f(\mathbf{x}^{(t)})$ . בambilים אחרות, SGD אבסטרקטי, כל צעד הוא רנדומי, אך בממוצע נופל ב-  $\partial f(\mathbf{x}^{(t)})$  בעצור בזמן  $T$  כלשהו ונוחיר את הוקטור המומוצע של האיטרציה.

#### 11.1.2 דוגמא

נניח בעית אופטימיזציה ללא אילוצים בה נרצה למצוא מינימום לפונק' מטרה קמורה כלשהי מהצורה  $f(\mathbf{w}) = \sum_{i=1}^m f_i(\mathbf{w})$ . נניח ש-  $\sum_{i=1}^m f_i(\mathbf{w}) = \sum_{i=1}^m \partial f_i(\mathbf{w})$  מכיון ש- איטרצית סאב-גרדיינט עבור  $f$  תהיה מהצורה:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \sum_{i=1}^m \mathbf{g}_i^{(t)}$$

עבור  $i \in [m]$   $\mathbf{g}_i^{(t)} \in \partial f_i(\mathbf{x}^{(t)})$

**שיטת  $i$  אקראי:** נבצע את האיטרציה  $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \mathbf{g}_{k(t)}^{(t)}$  עבור  $k(t) \stackrel{\text{i.i.d}}{\sim} \text{Unif}([m])$ . זהה איטרצית SGD עבור  $f$ . **שיטת Random mini-batch:** נוכל להכליל את השיטה לעיל ולבצע את האיטרציה:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\eta_{t+1}}{|K(t)|} \sum_{j \in K(t)} \mathbf{g}_j^{(t)}$$

עבור  $K(t) \subseteq [m]$  תת-קובוצה שנדגמה באופן אחד, או דגימת בוטסטרהף מגודל כלשהו. זהה גם איטרצית SGD עבור  $f$ .

#### 11.1.3 בעיות למידה חישוביות

נניח בעית למידה קמורה עם מחלוקת היפותיות  $\mathcal{H} \subseteq \mathbb{R}^d$  ופונק' הפסד  $\ell(\mathbf{w}, (\mathbf{x}, y))$ . יהי  $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$  סט אימון. נסמן  $(\mathbf{x}, y) = z$ , ו-  $L_S(\mathbf{w})$  הוא הסיכון האמפירי של היפותיה  $w$ . נניח כי היפותיה  $\mathcal{H} \ni w$  אותה אנו מחזירים היא הנקודה האופטימלית של הבעה הקמורה:  $\min_{\mathbf{w} \in \mathcal{H}} L_S(\mathbf{w})$ . כאשר נוכל להחליף את  $L_S$  בכל פונק'  $L(\mathbf{w}) = \sum_{i=1}^m f(\mathbf{w}, (\mathbf{x}_i, y_i))$  ועל כן ההכללה היא מעבר לכללי (ERM) יהיו  $\mathbf{z}_1, \dots, \mathbf{z}_T$  דוגימות המתפלגות באופן אחד ובלתי תלוי מ-  $S$ . אז:  $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_{t+1} \mathbf{g}^{(t)}$  עבור  $\mathbf{g}^{(t)} \in \partial \ell(\mathbf{w}^{(t)}, \mathbf{z}_t)$

### 11.1.4 השוואת GD מול SGD

בכל איטרציה של GD אנו מסתכלים על כל הדטא ומחשבים את הגרדיאנט המלא (או התת-גרדיאנט)  $\nabla L_S(\mathbf{w}^{(t)})$  בנקודה  $\mathbf{w}^{(t)}$ . בכל איטרציה של שיטת i random של SGD אנו מסתכלים על דגימה בודדת אקראית, ומשתמשים ב-  $\nabla \ell(\mathbf{w}^{(t)}, (\mathbf{x}_i, y_i))$ . זה משערך לא מוטה מכיוון ש-

$$\mathbb{E} [\nabla \ell(\mathbf{w}^{(t)}, (\mathbf{x}_i, y_i))] = \nabla \mathbb{E} [\ell(\mathbf{w}^{(t)}, (\mathbf{x}_i, y_i))] = \nabla L_S(\mathbf{w}^{(t)})$$

כאשר ההתולות היא ביחס לערכי  $i$  הנדגמים באופן אחיד מ-  $[m]$ , אך שונות השערוך גבוהה. מיצוע יכול לעזור בהורדת השונות - עבור  $B \subseteq [m]$ mini-batch נשתמש ב-

$$\frac{1}{|B|} \sum_{i \in B} \nabla \ell(\mathbf{w}^{(t)}, (\mathbf{x}_i, y_i))$$

כמשערך עבור  $\nabla L_S(\mathbf{w}^{(t)})$ .

- **חוקים ציקליים.** במקומות לבחור נקודה או אוסף בצורה אקראית, שיטה נפוצה היא לעבור על הדטא בצורה מחזורית.
- **אפוק epoch** הוא מספר צעדי ה- SGD הנחוצים על מנת להשתמש בכל הדטא פעם אחת. אם משתמשים ב- SGD מחזורי עם דגימה אחת בכל פעם, כל אפוק הוא  $m$  צעדי SGD. אם משתמשים ב- SGD מחזורי עם mini-batch גדול  $b$ , אז כל אפוק הוא  $\frac{m}{b}$  צעדי SGD.

### 11.1.5 כבר לא Batch Learning

נניח ואימנו את המודל שלנו על סט אימון  $S_1$  והתחלנו להשתמש במודל המאומן. לאחר זמן מה, מגיע עוד DATA -  $S_2$ . עוד לא יצא לנו להיתקל כזה בקורס - ועבור כל מודל שלמדו עד כה הדרך היחידה להשתמש בDATA החדש הייתה לאמן את המודל מחדש. לעומת זאת עם SGD, נוכל להמשיך להריץ את SGD עוד איטרציות עם DATA החדש. אפשר לחשב גם על מודל.streaming בו המודל כל הזמן מאומן ע"י SGD עם DATA חדש שמנגע.

### 11.1.6 ה- SGD אופטימיזציה עבור שגיאת הכללה

נניח בעיית למידה קמורה עם פונק' הפסד  $\ell$ .  
נשתמש בהנחה ה- PAC Agnostic  $\mathcal{D} \stackrel{\text{i.i.d}}{\sim} z = (\mathbf{x}, y)$  לפיה  $\mathcal{D}$  כלשהי מעל  $\mathcal{X} \times \mathcal{Y}$ .  
המטרה שלנו היא לפתור את הבעיה:

$$\min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) \text{ where } L_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{z \sim \mathcal{D}} [\ell(\mathbf{w}, z)]$$

מכיוון שאין לנו דרך ישירה לאפותם את  $L_{\mathcal{D}}$ , פיתחנו תיאוריה שלמה סביב עקרון ה- ERM, לפיה איפטמנו את  $(\mathbf{w})$  כ"נצח" עבור איפטום  $L_{\mathcal{D}}$  ומצאו תנאים עבור הכללה טובה.

נראה כעת אלג' אופטימיזציה קמורה שמצויר את  $L_{\mathcal{D}}$  ישירות. זה עקרון **למידה חדש** - דרך חדשה לבחור  $\mathbf{h} \in \mathcal{H}$ .  
אם היינו יודעים מה הוא  $L_{\mathcal{D}}$ , היינו משתמשים באיטרצית GD הבאה:  $0 = \mathbf{w}^{(1)} - \eta \nabla L_{\mathcal{D}}(\mathbf{w})$  ו-  $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla L_{\mathcal{D}}(\mathbf{w}^{(t)})$   
אך אנו לא יודעים. למרות זאת, לפי שינוי סדר גירה ותוחלת נקבל: (גם עבור תת-גרדיאנטים)

$$\nabla L_{\mathcal{D}}(\mathbf{w}) = \nabla \mathbb{E}_{z \sim \mathcal{D}} [\ell(\mathbf{w}, z)] = \mathbb{E}_{z \sim \mathcal{D}} [\nabla \ell(\mathbf{w}, z)]$$

ובמילים אחרות,  $\nabla \ell(\mathbf{w}^{(t)}, z_i)$  הוא אומד בלתי מוטה לגרדיאנט  $\nabla L_{\mathcal{D}}(\mathbf{w})$ . מכאן שהאיטרציה:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_{t+1} \nabla \ell(\mathbf{w}^{(t)}, z_i)$$

עבור  $z_i$  הנדגם באקראי מ-  $S$  היא איטרציה SGD עבור שגיאת הכללה  $L_{\mathcal{D}}(\mathbf{w}) \rightarrow \mathbf{w}$ , למרות שאנו לא יודעים את  $L_{\mathcal{D}}$

### 11.1.7 משפט

תהי בעית אופטימיזציה קמורה, לפשיצית וחסומה עם פרמטרים  $B, \rho$ . אזי לכל  $0 < \varepsilon$ , אם נريץ את שיטת SGD עבור מזעור  $(\mathbf{w})$  עם מס' איטרציות המקיימים:

$$T \geq \frac{B^2 \rho^2}{\varepsilon^2}$$

$$\mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \varepsilon \quad \text{וגודל צעד } \eta = \sqrt{\frac{B^2}{\rho^2 T'}}$$

### 11.1.8 אלגוריתמים

#### SGD for Solving Soft-SVM

**goal:** Solve Equation (15.12)

**parameter:**  $T$

**initialize:**  $\theta^{(1)} = \mathbf{0}$

**for**  $t = 1, \dots, T$

    Let  $\mathbf{w}^{(t)} = \frac{1}{\lambda t} \theta^{(t)}$

    Choose  $i$  uniformly at random from  $[m]$

    If  $(y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle < 1)$

        Set  $\theta^{(t+1)} = \theta^{(t)} + y_i \mathbf{x}_i$

    Else

        Set  $\theta^{(t+1)} = \theta^{(t)}$

**output:**  $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}^{(t)}$

#### SGD for Solving Soft-SVM with Kernels

**Goal:** Solve Equation (16.5)

**parameter:**  $T$

**Initialize:**  $\beta^{(1)} = \mathbf{0}$

**for**  $t = 1, \dots, T$

    Let  $\alpha^{(t)} = \frac{1}{\lambda t} \beta^{(t)}$

    Choose  $i$  uniformly at random from  $[m]$

    For all  $j \neq i$  set  $\beta_j^{(t+1)} = \beta_j^{(t)}$

    If  $(y_i \sum_{j=1}^m \alpha_j^{(t)} K(\mathbf{x}_j, \mathbf{x}_i) < 1)$

        Set  $\beta_i^{(t+1)} = \beta_i^{(t)} + y_i$

    Else

        Set  $\beta_i^{(t+1)} = \beta_i^{(t)}$

**Output:**  $\bar{\mathbf{w}} = \sum_{j=1}^m \bar{\alpha}_j \psi(\mathbf{x}_j)$  where  $\bar{\alpha} = \frac{1}{T} \sum_{t=1}^T \alpha^{(t)}$

## 11.2 רשת נוירונים קטנה

נזכיר במודל של רגרסיה לוגיסטיבית. מחלוקת ההיפוטזות הוגדרה ע"י:

$$\mathcal{H} = \{x \mapsto \phi(\langle w | x \rangle) \mid w \in \mathbb{R}^{d+1}\}$$

עבור הפונקציה  $\phi(x) = \frac{e^x}{1+e^x}$ , המפה איברים מ-  $(-\infty, \infty)$  ל-  $(0, 1)$ . [ ניתן כמובן לקחת כל פונק' דומה אחרת ]

עבור וקטור משקלות  $w$  נתון, ודוגמה  $x$ , נפרש את  $1 \leq \phi(\langle w | x \rangle) \leq 0$  כנראות שעבורה התגית של  $x$  היא 1.

עקרון הלמידה בו השתמשנו היה **מקסום הנראות**, בהינתן סט אימון  $\{(x_i, y_i)\}_{i=1}^m$  נבחר וקטור  $w \in \mathcal{H}$  לפי:

$$L(w | y) = \sum_{i=1}^m [\log(\phi(\langle w | x_i \rangle)) \cdot \mathbb{1}_{y_i=1} + \log(1 - \phi(\langle w | x_i \rangle)) \cdot \mathbb{1}_{y_i=0}]$$

### 11.2.1 הרעיון - הרכבת פונקציות

נניח ואנו מרכיבים מודל רגרסיה לוגיסטיבית ומקבלים משקלות  $w_1^{(1)}$  ויכולים לחזות:

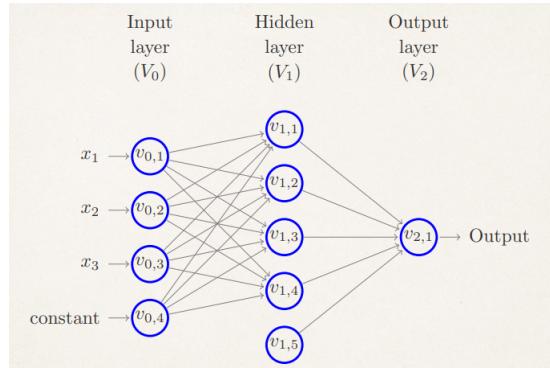
$$x \mapsto \phi(\langle x | w_1^{(1)} \rangle)$$

כעת אנו יכולים לאמן מודל שני ולקבל משקלות  $w_2^{(1)}$  ויכולים לחזות:

$$x \mapsto \phi(\langle x | w_2^{(1)} \rangle)$$

ולהמשך ככה עד שנאמן  $k$  מודלים. כעת בידנו  $dk$  פרמטרים ומחלקת ההיפותזות שלנו מכילה פונק'  $f$  כאשר כל פונקציה  $f_j$  היא פلت של מודל ליניארי או מודל רגרסיה לוגיסטיבית אחר.

### 11.2.2 שכבה סמויה



- השכבה השמאלית ביותר היא "שכבה הקלט" - קורדינאות ה- $x$  שלנו.
- השכבה האמצעית היא "השכבה הסמויה" - היא כוללת פונק' סקלרית לא ליניארית לה נקרא activation. כל נוירון בשכבה זו בעצם מייצג  $\sigma(\langle W_i^\top x | w_i^{(1)} \rangle) = \sigma(W_i^\top x)$ .
- השכבה הימנית היא "שכבה הפלט" - המיצגת ע"י  $\phi(\langle \sigma(W_1^\top x) | w_2 \rangle)$ .
- ה-constant בשכבה השמאלית ביותר משמש כ-intercept.

### 11.2.3 הגדרה

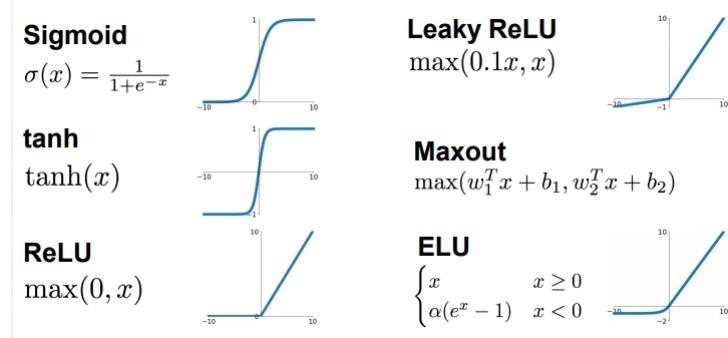
עבור דוגמאות ב-  $\mathbb{R}^d$ , נגדיר את מחלקת ההיפותזות של 'feedforward NN' בעלת שכבה סמויה אחת (המכילה  $k$  נוירונים) ופונק' אקטיבציה  $\sigma$  עבור רגרסיה או סיווג ע"י:

$$\mathcal{H} = \phi(\langle \sigma(W_1^\top x) | w_2 \rangle)$$

עבור  $\mathbb{R} \rightarrow \mathbb{R}$ :  $\sigma$  פונק' אקטיבציה כאשר  $\sigma, \phi$  מופעלות על וקטור איבר  $A$ ,  $W_1, x \in \mathbb{R}^d$ ,  $w_2 \in \mathbb{R}^k$  היא מטריצה  $d \times k$ .

עבור בעיית רגרסיה, משתמש ב-  $x = \phi(x)$  פונק' זהות. עבור בעית סיווג ביןארי משתמש ב-

## 11.2.4 פונק' אקטיבציה אפשריות



## 11.2.5 שכבת הפלט

נבחין כי רשת הנוירונים מקבלת וקטור  $x$  כקלט, ונחזירה כפלט את  $f(x)$ .  
ההחלטה האם לעשות רשת עבור בעית סיווג ביןארית, בעית רגרסיה, או בעיה רבת משתנים היא בשכבה האחורונה.  
עבור רגרסיה או בעית סיווג ביןארי, שכבת הפלט תכיל ניורון אחד בלבד.  
עבור רגרסיה מרובה ( $k$  חיצויים) או סיווג רב מושתנים ( $k$  מחלקות), שכבת הפלט תכיל  $k$  נוירונים.

- עבור בעית רגרסיה, ניורון הפלט ייקח משקלות  $w_2$  וקלט  $\langle o_2 | w_2 \rangle$ .
- עבור רגרסיה מרובה, כל דגימה מהדatta היא מהצורה  $(x, y_1, \dots, y_d)$  עם  $x \in \mathbb{R}^d$  ו  $y_i \in \mathbb{R}$ .  
אנו ניקח  $k$  נוירונים שייקחו משקלות  $W_2$  וקלט  $\langle W_2^\top x | o_2 \rangle$  ויחזרו  $\sigma(W_2^\top x)$  וקלט  $\langle W_1^\top x | o_2 \rangle$  ויחזר:

$$\text{logit}_{w_2}(o_2) = \frac{e^{\langle o_2 | w_2 \rangle}}{1 + e^{\langle o_2 | w_2 \rangle}}$$

כאשר לבחירת הפונק' 2 סיבות עיקריות: הראשונה היא לבדוק כמו ברגרסיה ליניארית אנו משתמשים בעיקרון מקסום הנראות. הסיבה השנייה היא כי הפונק'  $\log(\text{logit}_w(x)) = \frac{\text{sign}(\langle w | x \rangle + 1)}{2}$  שעலום אינה שטוחה.  
עובדת זו שימושית עבור GD.

- עבור סיווג רב מושתנים, במקרה להכניס מודל ליניארי יחיד  $\langle w | x \rangle$  לתוך פונק' הלוגיט, נkeh מס' מודלים:

$$x \mapsto (\langle x | w_1 \rangle, \dots, \langle x | w_k \rangle)$$

ונפעל בצורה זהה. הנראות שחלוקת הנכונה עבור דגימה  $x \in \mathbb{R}^d$  תהיה:

$$\frac{e^{\langle x | w_i \rangle}}{1 + \sum_{j=1}^{k-1} e^{\langle x | w_j \rangle}}$$

והנראות שהמחלקה  $k$  היא הנכונה תהיה:

$$\frac{1}{1 + \sum_{j=1}^{k-1} e^{\langle \mathbf{x} | \mathbf{w}_j \rangle}}$$

במקרה זה שוב נשתמש עבוקון מקסום הנראות על מנת לבחור  $\mathbf{w}_{k-1} \in \mathbb{R}^d$ .

מכאן שכבת הפלט תכיל  $k$  (או  $1 - k$ ) נוירונים. נסמן ב-  $z_j$  את המכפלה הפנימית המוגדרת ע"י נוירון הפלט ה-  $j$ .

הפונק' אוטה הוא ממש היא  $\frac{e^{z_j}}{\sum_i e^{z_i}}$ , ופונק' הוקטור המוגדרת ע"י:

$$(z_1, \dots, z_k) \mapsto \left( \frac{e^{z_1}}{\sum_i e^{z_i}}, \dots, \frac{e^{z_k}}{\sum_i e^{z_i}} \right)$$

נקראת פונק' ה- Softmax.

### 11.2.6 הכללה ל- $k$ שכבות סמיות

רשת נוירונים FF כללית עם קלט  $\mathbf{x} \in \mathbb{R}^d$ ,  $T$  שכבות, פונק' אקטיבציה  $\sigma$  ופלט ב-  $\mathbb{R}^k$  מוגדרת באופן הבא:

◦ גראף אציקלי מכוכן ( $V = \bigcup_{t=0}^T V_t$  המציגים את  $T$  השכבות).

◦ פונק' משקל  $E \rightarrow \mathbb{R}$  :  $w$  מעיל הצלעות (ניתנת כמטריצות  $(W_0, \dots, W_{T-1})$ ).

◦ פונק' אקטיבציה  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  :  $\sigma$  ופונק' פלט  $\sigma : \mathbb{R}^{|V_t|} \rightarrow \mathbb{R}^{|V_t|}$  עבור  $\phi$ .

נדיר בקורס רקורסיבית  $\mathbf{x} = \mathbf{o}_0$  ו-

$$\forall 1 \leq t \leq T : \mathbf{o}_t = \sigma(W_{t-1}^\top \mathbf{o}_{t-1})$$

ולכן רשת הנוירונים מוגדרת כפונקציה (היפוטזה) המפה  $\mathbf{x} \mapsto \phi(\sigma(\mathbf{o}_T))$ .

### 11.2.7 כמה עשרה מחלקות ההיפוטזות?

נתבונן ברשת נוירונים בولיאנית, בה הפונק'  $\phi(x) = \text{sign}(x) = \sigma(x) = \phi(x) = \begin{cases} 1 & \text{если } x > 0 \\ 0 & \text{если } x = 0 \\ -1 & \text{если } x < 0 \end{cases}$  בלבד.

**משפט:** לכל  $d$  קיימת רשת נוירונים בוליאנית עם שכבה סمية אחת, כך שלכל פונק'  $f : \{\pm 1\}^d \rightarrow \{\pm 1\}^d$  יש השמה למשקלות  $W_1, \dots, W_d$  עבורה רשת נוירונים עם משקלות אלה מוגדרת את  $f$ . עם זאת, המס' המינימלי של נוירונים הנחוצ' בשכבה הסمية הוא אקספוננציאלי ב-  $d$ .

ניתן להסיק ממשפט זה כי אפילו עם רשת נוירונים בעלת שכבה סمية אחת (מוגדל לא חסום), מחלקות ההיפוטזות של רשת הנוירונים היא עצומה בגודלה ועשירה מאוד.

### 11.2.8 אימון רשות עמוק

אימון רשות נוירונים בעזרת עקרון ERM היא בעיה NP-קשה.

נשאלת השאלה - באיזה אופן שונה רשות הנוירונים מכל חלוקת היפותיות סופית אחרת?

התשובה היא שחלוקת היפותיות  $\mathcal{H}_G$  עבור  $(V, E) = G$  מורכבת מוקטורים ב-  $\mathbb{R}^{|E|}$ .

זהי צורה אוקlidית, ובמקרה להשתמש בעקרון ה- ERM אנו משתמשים בעקנון הלמידה של מקסום הנראות.

בדיק כמו ברגRESSED לוגיסטי, עבור סט אימון  $S = \{(x_i, y_i)\}_{i=1}^m$  נסמן ב-  $\ell_w(x, y)$  את ה-NLL,

$$\text{ונרצה למזער את } \ell_w(x, y) = \sum_{(x,y) \in S} \ell_w(x, y) \text{ מעל משקלות הרשות } w.$$

נבחן כי בעוד ברגRESSED לוגיסטי הפונק' הניל' קמורה, ברשות נוירונים בעלת שכבה סمية אחת או יותר הפונק' אינה קמורה.

אז כיצד נפתרו את הבעיה? שימוש ב- GD על מנת למקסם את הנראות עבור בחירת המשקלות של הרשות מסתמן כמובן למדידה חזק מאוד למורות שהפונק' המטריה היא מאוד לא קמורה.

### 11.2.9 כיצד נפעיל את אלג' GD על הרשות?

על פניו, מימוש אלג' GD בעזרת האיטרציה:

$$w^{(t+1)} = w^{(t)} - \eta_{t+1} \nabla L(w^{(t)}; s)$$

על מנת למזער את ה- NLL נראית פשוטה ותativa. בפועל היא לא, שכן  $w \in \mathbb{R}^{|E|}$  מייצג את כל משקלות הרשות.

נבחן כי אם  $\phi, \sigma$  הן פונק' חלקות, אזי פונק' המטריה הינה דפרנציאבילית ויש לנו גרדיאנט. אם הן לא - כמו לדוגמה פונק'  $\sigma(x) = \max(0, x)$  - אז יש לנו תחת גרדיאנט.

למה בעיה זו קשה? מכיוון שאם הרשות עמוקה, הגודל  $|E|$  הוא עצום ולמודל מס' גדול מאוד של פרמטרים. מכאן שסט האימון  $S$  יהיה עצום בגודלו. נשאלת השאלה כיצד נחשב את הגרדיינט  $\nabla L(w; S)$ ?

1. נשתמש ב- SGD ולא ב- GD. במקומות לנסוט ולחשב את  $\nabla L(w; S)$  עבור כל סט האימון, נחשב את  $\nabla \ell_w(x, y)$ .

2. נשתמש באלג' נומרי מהיר לחישוב  $\nabla \ell_w(x, y)$  Back Propagation בשם.

### 11.2.10 אלג' SGD עם Mini-batch

ונכל למשתמש בסט אלג' SGD בעזרת אלג' Backprop שמאפשר לנו לחשב את הגרדיינט בנקודת אימון בודדת:

נניח חלוקה של הדאטא ל-  $S = \sum_{j=1}^{\mu} B_j$ . אזי באופן בודד, איטרציה  $t$  (עבור  $\mu \leq t \leq 1$ ) מעדכנת את משקלות הרשות ע"י:

$$w^{(t+1)} = w^{(t)} - \eta_t \sum_{(x,y) \in B_t} \nabla L(w^{(t)}; B_t)$$

**11.2.11 אלג' SGD עם מומנטום**

מכיוון שכיווני האלג' SGD לרוב בעלי שונות גובהה, נוכל להוסיף את Nesterov's momentum ולהשתמש באיטרציה:

$$\mathbf{v}^{(t+1)} = \alpha \mathbf{v}^{(t)} - \eta_t \sum_{(\mathbf{x},y) \in B_t} \nabla L(\theta^{(t)} + \alpha \mathbf{v}^{(t)}; B_t) \quad \left| \begin{array}{l} \theta^{(t+1)} = \theta^{(t)} + \mathbf{v}^{(t)} \\ \text{עבור היפר-פרמטר } \alpha. \end{array} \right.$$

**11.2.12 אתחול משקלות**

אנו מביצעים אלג' GD או SGD במورد פונק' שאינה קמורה. הנקודה ההתחלתית  $\theta^{(0)}$  שמןנה נתחל לרדת היא חשובה מאוד.

- ניתן לאתחול את המשקלות ע"י דוגמאות איחודות ובלתי תלויות ממשתנה מקרי כלשהו (עם שונות פרופורציונית לגודל השכבה).

- ניתן לאתחול את המשקלות כך שכל נוירון מתחילה עם מס' קבוע של משקלות אקרואיות שונות מ一封ס.

**11.2.13 גודל הצעד**

בקשר של רשתות נוירוניים, גודל הצעד  $\eta$  נקרא **קצב הלמידה**.

נזכר באלג' ה- Backtracking line search אותו ראיינו עבור GD. זהו אלג' אדפטיבי לבחירת גודל הצעד. גם ברשתות נוירוניים אין צורך להשתמש באותו  $\eta$  עבור כל הרכיבים, ולמעשה בחירת אדפטיבית של קצב הלמידה הפכה לאומנות, עם מתודות מפורסמות כמו AdaGrad, RMSProp, ADAM ועוד.

- על קצב הלמידה לגודל עם גודל ה-.mini-batch

**11.2.14 רגולרייזציה**

נזכר ברגולרייזציה רידג' - הוספת פרמטר רגולרייזציה יכול להוריד את המשקלות, ובכך להוריד את השונות ולשפר את הביצועים. דבר זה נכון גם עבור רשתות נוירוניים. קל להוסיף פרמטר רגולרייזציה  $\ell_2$  למשקלות: אם פונק' המטרה היא  $L(\mathbf{w})$  במקום  $L(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$  אז הגרדיאנט הוא פשוט  $\nabla L(\mathbf{w}) + \lambda \mathbf{w}$ .