048843

Itay Hubara and Zohar Rimon

April 2020

1 Testing a simple active algorithm

1.1 Synthetic dataset

a Since $\Sigma_+ = \Sigma_- = \Sigma$ and $P(w_-) = P(w_+) = \frac{1}{2}$ we get that the Bayes Classifier is:

$$\frac{1}{2\pi^{k/2} \cdot |\Sigma|^{1/2}} e^{(x-\mu_-)^T \Sigma^{-1}(x-\mu_-)} = \frac{1}{2\pi^{k/2} \cdot |\Sigma|^{1/2}} e^{(x-\mu_+)^T \Sigma^{-1}(x-\mu_+)} \tag{1}$$

Multiplying by $2\pi^{k/2} \cdot |\Sigma|^{1/2}$ and taking log from both sides results with:

$$(x-\mu_-)^T \Sigma^{-1}(x-\mu_-) = (x-\mu_+)^T \Sigma^{-1}(x-\mu_+)$$

Straightforward derivation leads to:

$$2x^T \Sigma^{-1}(\mu_- - \mu_+) = \mu_-^T \Sigma^{-1} \mu_- - \mu_+^T \Sigma{-1} \mu_+$$

Thus we get

$$W = \Sigma^{-1}(\mu_- - \mu_+) \ \ ; \ \ b = -\frac{1}{2}(\mu_-^T \Sigma^{-1} \mu_- - \mu_+^T \Sigma^{-1} \mu_+)$$

b Here we apply same derivation but for the general case in which $\Sigma_- \neq \Sigma_+$, this results with:

$$\frac{1}{2\pi^{k/2} \cdot |\Sigma_-|^{1/2}} e^{(x-\mu_-)^T \Sigma_-^{-1}(x-\mu_-)} = \frac{1}{2\pi^{k/2} \cdot |\Sigma_+|^{1/2}} e^{(x-\mu_+)^T \Sigma_+^{-1}(x-\mu_+)} \tag{2}$$

$$\frac{|\Sigma_+|^{1/2}}{|\Sigma_-|^{1/2}} e^{(x-\mu_-)^T \Sigma_-^{-1}(x-\mu_-)} = e^{(x-\mu_+)^T \Sigma_+^{-1}(x-\mu_+)} \tag{3}$$

$$\frac{1}{2} log(|\Sigma_+|) - \frac{1}{2} log(|\Sigma_-|) + (x-\mu_-)^T \Sigma_-^{-1}(x-\mu_-) = (x-\mu_+)^T \Sigma_+^{-1}(x-\mu_+) \tag{4}$$

Defining $c1 = \frac{1}{2} log(|\Sigma_+|) - \frac{1}{2} log(|\Sigma_-|)$ and simplifying the equation leads to

$$c1 + x^T(\Sigma_-^{-1} - \Sigma_+^{-1})x + 2x^T(\Sigma_+ \mu_+ - \Sigma_- \mu_-) = \mu_+ \Sigma_+ \mu_+ - \mu_- \Sigma_- \mu_- \tag{5}$$

Thus we results with quadratic classifier: $x^T A x + x^T b + c$ where $c = c1 + \mu_- \Sigma_- \mu_- - \mu_+ \Sigma_+ \mu_+$, $A = \Sigma_-^{-1} - \Sigma_+^{-1}$ and $b = 2(\Sigma_+ \mu_+ - \Sigma_- \mu_-)$

c As can be seen in table 1 ,the first experiment in which the mean is normalized by $sqrt(d)$ the accuracy results is roughly constant while for the second experiment for which $\mu_{+/-} = \pm(1, 1, .., 1)$ we get perfect classification for $d \geq 20$

Experiment	d=1	d=20	d=50	d=200
$\mu_{\pm} = \pm(1, 1, ...1)/\sqrt(d)$	84%	84.5%	83.5%	84.6%
$\mu_{\pm} = \pm(1, 1, ...1)$	84%	100%	100%	100%

In both of the cases we get the same Bayes estimator (as seen in the first section, with this section's parameters):

$$W = (\mu_{-} - \mu_{+}) \;\; ; \;\; b = -\frac{1}{2}(\mu_{-}^{T}\mu_{-} - \mu_{+}^{T}\mu_{+}) = 0$$

*We get different Ws but the result is a same classifier.

We can notice that the classifier, as expected is symmetric and is through the origin. more specifically, it does not change with the dimension.

For the normalized case: The distance between the means of the two distributions and the origin stays the same as the dimension increases

$$dist = \sqrt{\sum_{i=1}^{d}(\mu_{+} - 0)^{2}} = \sqrt{\sum_{i=1}^{d} 1/d} = 1$$

For the non-normalized case: The distance between the means of the two distributions and the origin grows as the dimension increases

$$dist = \sqrt{\sum_{i=1}^{d}(\mu_{+} - 0)^{2}} = \sqrt{\sum_{i=1}^{d} 1} = \sqrt{d}$$

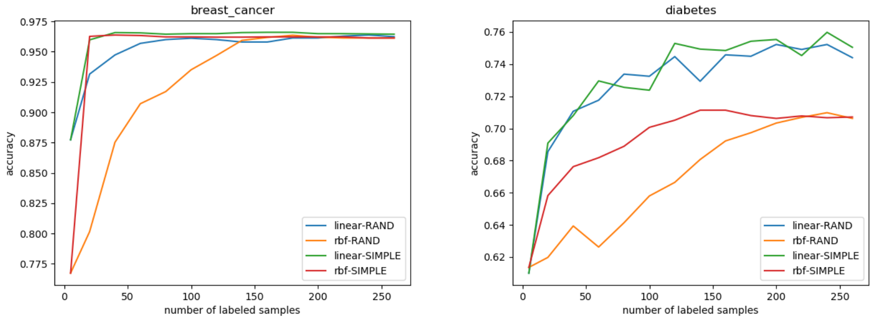
For the normalized case, although the distance of the Gaussian means from the origin is constant and equal to the variance, thus we get constant accuracy of 85% (which makes as for each as for each Gaussian you have 50% that bigger/smaller then the mean and an additional 34% which still on the correct side of the hyper-plane) . For the non-normalized case the distance of the Gaussian's mean from the origin scales with $\sqrt(d)$ thus for $d == 20$ we get that the distance from the origin is between $4\sigma - 5\sigma$ thus the classification is perfect.

1.2 Test Set Results

<u>$Setting$</u>

Throughout the experiments we used sklearn package which includes a method to optimized the parameters, $SearchGridCV$. This method run hyper parameters tuning using a simple exhaustive grid search (nested for loops) and picks the best parameters based on the cross validation score. The results were averaged over 30 trials (runs in parallel) where in each round the pool/test/val dataset are randomly chosen out of 600 samples. We note that we used the scaled breast cancer dataset no further reprocessing was done to the datasets.

As can be seen in Figure 1. SIMPLE algorithm generalize better on the dataset as it improved its hyper-plane based on the points close to the margin.



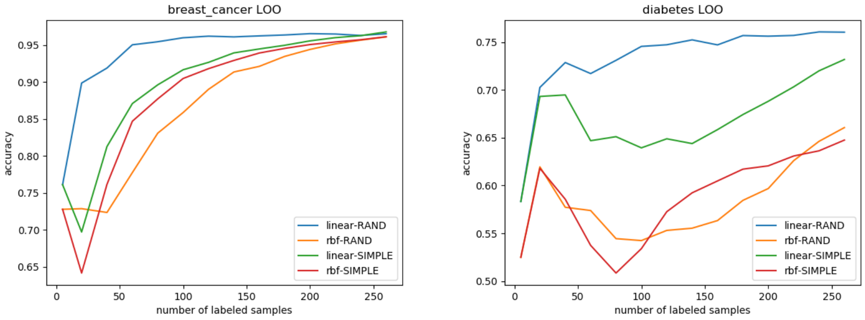
(a) breast cancer dataset

(b) diabetes dataset

Figure 1: Accuracy (over the test set) per number of labeled samples for the ”real” datasets

1.3 LOO Results

To reduce running time we used cross validation with five segments instead LOO. Here, by using SIMPLE algorithm we pick the hardest points to classify (the one closes to the margin, and test our performance on them. Thus, it falsely appears like that the SIMPLE algorithm perform worse then the RAND. However if we would test the SIMPLE algorithm on the subset of the training set chosen by the RAND algorithm we would see that it is actually better.



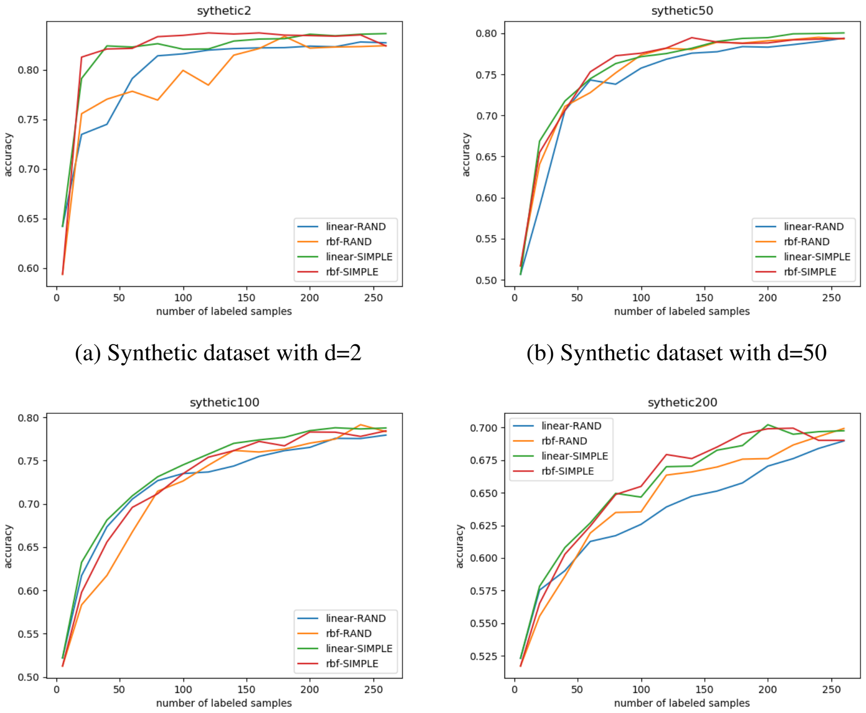
(a) breast cancer dataset

(b) diabetes dataset

Figure 2: Accuracy (using LOO on the labeled samples) per number of labeled samples for the ”real” datasets

1.4 Synthetic dataset results

Here we see the same trends as in the ”real” dataset. The SIMPLE algorithm generalize better on the test dataset and if you test the algorithms on the labeled points, since SIMPLE choose the ones closest to the margin its results are inferior, In addition we observe that the accuracy decreases with the problem dimension suggesting that the estimation error increases with the dimension. This means that ,although the Bayes estimator is a linear classifier thus the approximation error is zero the SVM can’t find it. Since we know that the samples are not noisy, we simply need to increase the numbers of samples to achieve Bayes error.



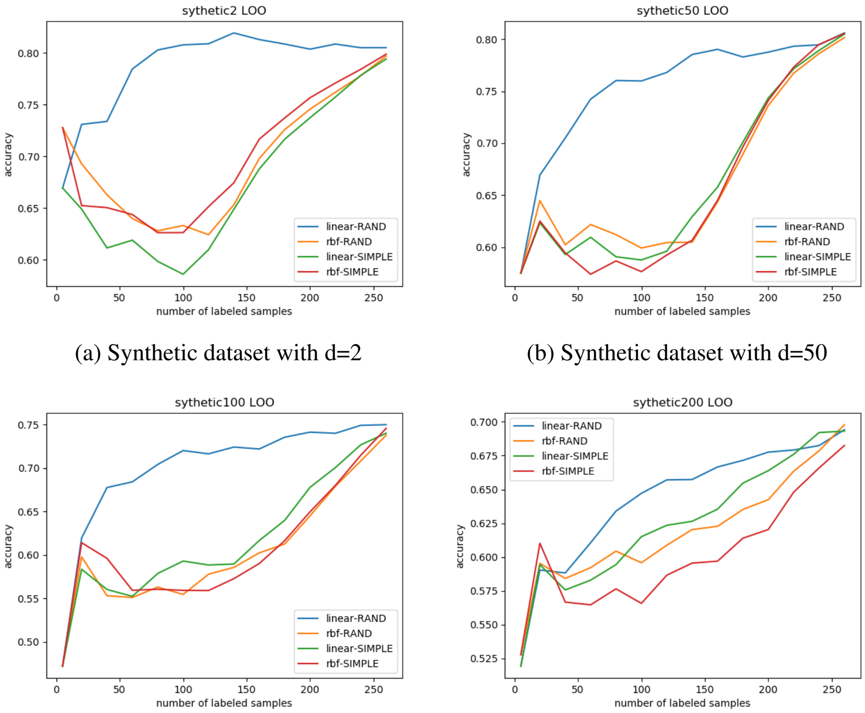
(a) Synthetic dataset with d=2

(b) Synthetic dataset with d=50

(c) Synthetic dataset with d=100

(d) Synthetic dataset with d=200

Figure 3: plots of synthetic dataset accuracy over the test set v.s number of labels sampled for d=2,50,100,200



(a) Synthetic dataset with d=2

(b) Synthetic dataset with d=50

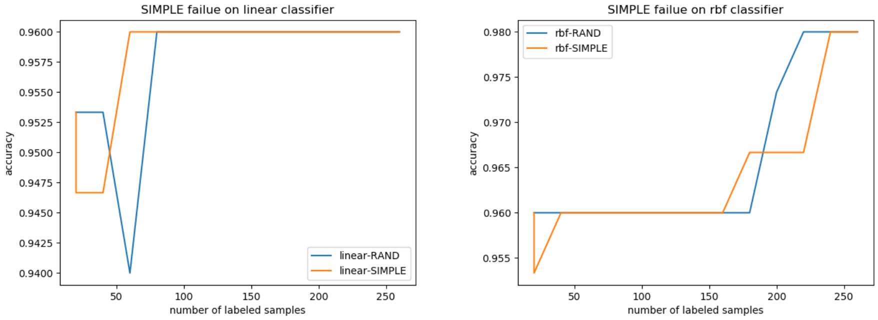
(c) Synthetic dataset with d=100

(d) Synthetic dataset with d=200

Figure 4: plots of synthetic dataset accuracy using LOO over the labels samples v.s number of labels sampled for d=2,50,100,200

1.5 SIMPLE Algorithm Failures

For the linear classifier we'll use the 1D example from the lecture notes in which 95% of the data leads to a hyper-plane on the origin. Since, with high priority we would sample from those examples we wont see the rest of the 5% and if would take a lot of samples until we'll get 97.5 accuracy. For the RBF kernel the example is similar only in 2D with circles. Thus we know that the classifier can reach 97.5% yet it requires almost all the samples to get to that accuracy



(a) breast cancer dataset

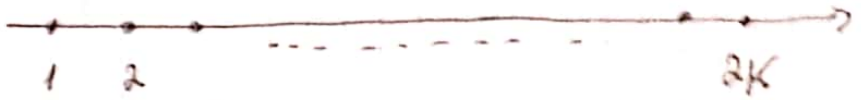
(b) diabetes dataset

Figure 5: Simple failure on RBF and Linear classifier

2) <u>חלק עיוני</u>

I) <u>מימד VC</u>

1) נתבונן ב 2k נקודות על הציר הממשי:



ניתן להשיג ע"י k קטעים כל תיוג אפשרי של הנקודות הללו (ע"י מיקום הקטעים בהתאם).

עבור כל תיוג אחר בו יש נקודות רצופות המתויגות זהה ניתן "לאחד" אותן ולהתבונן בהן כנקודה יחידה ובכך לצמצם גם את הקטעים שלהן.

עבור 2k+1 נקודות, נתבונן בתיוג הבא:

$+ \; - \; + - \_\_\_\_\_ \ldots + - +$

⇐ יש k+1 נקודות + מופרדות ולכן צריך k+1 אינטרוולים בכדי לכסות את הסיווג.

$$\Rightarrow \boxed{V_H = 2k}$$

2) נראה שמימד VC = 3. יהיו 4 נקודות שונות $x_1<x_2<x_3<x_4$. לא ניתן להשיג תיוג $+-+-$.

נניח בשלילה שכן: חייב להתקיים ע"י אינטרוולים:

$$\Rightarrow x_1 \in [\alpha, \alpha+1],\ x_2 \in (\alpha+1, \alpha+2),\ x_3 \in [\alpha+2, \infty)$$

$$\Rightarrow x_4 \in [\alpha+2, \infty)$$

$$\Rightarrow h(x_4) = '+' \neq '-'$$

$V_H < 4$ ⊏

נרצה להראות ש $V_H=3$. נראה קבוצה של 3 נקודות אשר shattered:

$$x_1=\frac{2}{8}, \quad x_2=\frac{9}{8}, \quad x_3=\frac{14}{8}$$

נראה שלכל תיוג קיים מסווג המשיג אותו בצורה הבאה!

$+ + + \Rightarrow \alpha=-5 \Rightarrow [-5,-4]\cup[-3,\infty)$

$- - - \Rightarrow \alpha=5 \Rightarrow [5,6]\cup[7,\infty.)$

$+ + - \Rightarrow \alpha=\frac{2}{8} \Rightarrow [\frac{2}{8}, \frac{10}{8}]\cup[\frac{18}{8},\infty)$

$+ - + \Rightarrow \alpha=-\frac{3}{8} \Rightarrow [-\frac{3}{8}, \frac{5}{8}]\cup[\frac{13}{8},\infty)$

$- + + \Rightarrow \alpha=-1 \Rightarrow [-\frac{7}{8}, \frac{1}{8}]\cup[\frac{9}{8},\infty)$

$- - + \Rightarrow \alpha=-\frac{13}{16} \Rightarrow [-\frac{13}{16}, \frac{3}{16}]\cup[\frac{19}{16},\infty)$

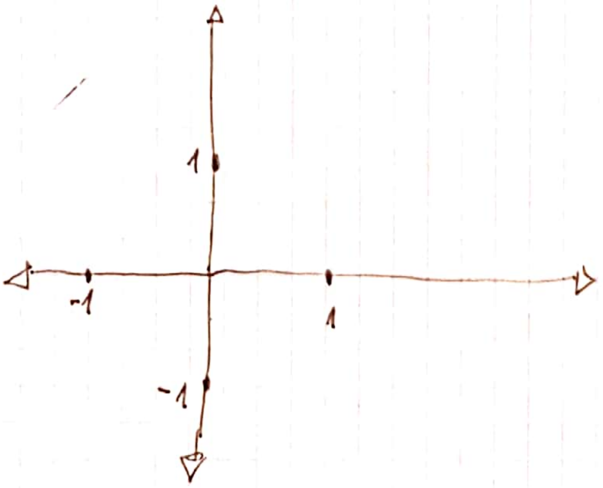
$- + - \Rightarrow \alpha=\frac{1}{2} \Rightarrow [\frac{1}{2}, \frac{3}{2}]\cup[\frac{5}{2},\infty)$

$+ - - \Rightarrow \alpha=0 \Rightarrow [0,1]\cup[2,\infty)$

$\boxed{V_H=3}$

$\sigma=$

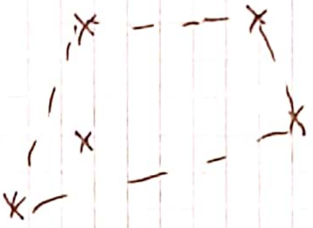
(3) נימוק בנקודות:



$\Rightarrow$ הבחירות שנקודותינו מנופצות ע"י המחלקה

נוכיח כי כל סט של 5 נקודות לא מנופץ ע"י המחלקה ולכן $V_H=5$.

נתבונן ב5 נקודות ובצורה הקונבקסית שלהן:



<u>מקרה I</u>: ישנה נקודה בתוך הצורה הקונבקסית:

. צריך לסווג את הנקודות על הצורה כ ⊕ ואת הפנימית כ ⊖ בטבלת המתאימה. כיצד הייתה מופיעים בה הצורה הקונבקסית?

~~מקרה II: אין נקודה פנימית:~~

<u>הסתירה</u>: כל הקודקודים של הצורה בתוך המשולש ועל כן כל הקווים המתחברים ביניהם בתוך המשולש ולכן כל הצורה הקונבקסית בתוך המשולש ולא ניתן שכל נקודה תהיה מחוברת אליו.

<u>מקרה II</u>: אין נקודה פנימית! $p_{\leftarrow}$ $p_{\rightarrow}$

נתבונן בנקודה השמאלית ביותר והימנית ביותר בין 5 הנקודות (כבלל נה"כ' הן שונות ויחידות)

אותן נקרא צ"ל ב + ולכן בסיס המשולש צריך לעבור מתחת ל $p_{\rightarrow}$ והצד השמאלי משמאל ל $p_{\leftarrow}$. כמו את שאר הנקודות שעברו יהיו מסווגות בטבלת ישר עם שנות שלהן!

נסדר את 3 הנקודות שנותרו כ: p_3, p_2, p_1 בסדר משמאל לימין: $x(p_1)<x(p_2)<x(p_3)$ (וכבל נה"כ' אין שוויון)

נשים לב מכיוון ואין נקודה בתוך הצורה הקונבקסית:

הישר המחבר בין p_1 לp_3 צריך לעבור מתחת ל p_2. (*)

עבור סיווג: p_3: −, p_2: +, p_1: − צריך שהישר יעבור מתחת p_1 ו p_3 ומעל p_2 אך זה לא ייתכן מ(*).

⇐ לא ניתן לסווג סיווג שכזה.

<u>מקרים כרגיל</u>: (i) אם יש נקודה שהיא השמאלית ביותר והימנית ביותר − נצטרך יחידות הנקודות האחרות ונמשיך כרגיל.

(ii) אם יש שתי נקודות מעל שהן הימניות ביותר (ומהן שמאליות ביותר)

נבחר מהמשולשים הנקודות את הזוית יחד ונמשיך כרגיל.
ואם יש שוויון $x(\beta_1)=x(\beta_2)$ ההוכחה נשארת נכונה.

II) מהסתברות לתוחלת!

הוכחה:

$$P(X>t)\leq ce^{-2nt^2}$$

$$E(X^2)=\int_0^\infty P(X^2\geq t)dt=\int_0^a P(X^2\geq t)dt+\int_a^\infty P(X^2\geq t)$$

נוסחת הזנב

$$\underset{P<1}{\leq} a+\int_a^\infty P(X^2\geq t)=a+\int_a^\infty P(X\geq\sqrt{t})dt$$

$$\underset{\text{מנתון}}{\leq} a+c\int_a^\infty e^{-2nt}dt=a+\frac{ce^{-2na}}{2n}$$

עבור $a=\log(c)/n$ נקבל:

$$E[X^2]\leq\frac{\log(c)}{2n}+\frac{1}{2n}=\frac{\log(ce)}{2n}$$

בנוסף:

$$E[X]\underset{\text{מינסן}}{\leq}\sqrt{E[X^2]}\leq\sqrt{\frac{\log(ce)}{2n}}$$

III) י' - הכנאה

(1)נבדוק את כל התכונות!

$$d(h,h')=P\{h(x)\neq h'(x)\}\underset{\text{כי הסתברות}}{\geq}0$$

$$d(h,h')=d(h',h)=P(h(x)\neq h'(x))$$

עבור $h=h'$ אז

$$d(h,h)=P(h(x)\neq h(x))=0$$

עבור $h\neq h'$ אז יש x עבורו $h(x)\neq h'(x)$ אז (צריך לבדוק כאן מספרים

$$d(h,h')\neq0$$

$$d(h,h') + d(h',h'') = P(h(x) \neq h'(x)) + P(h'(x) \neq h''(x))$$

$$\geq P(h'(x) \neq h''(x) \wedge h(x) = h'(x))$$

$$+ P(h(x) \neq h''(x) \wedge h(x) \neq h'(x))$$

$$= P(h(x) \neq h''(x)) = d(h,h'')$$

(ג)

$$DIS(B(h, \varepsilon/\sqrt{d})) =$$

$$= \{x \in \mathbb{R}^d ; \exists h,h' \in B(h, \varepsilon/\sqrt{d}) \text{ with } h(x) \neq h'(x)\}$$

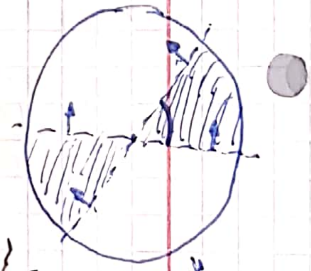
$$B(h, \varepsilon/\sqrt{d}) = \{h' \in H : d(h,h') \leq \frac{\varepsilon}{\sqrt{d}}\}$$

מכיוון ו X מפולג אחיד בתוך כדור היחידה ב $\mathbb{R}^d$:

$$d(h,h') = P\{h(x) \neq h'(x)\} = \frac{\underset{\text{נפח כל המפריד}}{V}(h(x) \neq h'(x) : x \in \text{unit-ball})}{V_d} =$$

נפח
unitball

$$= \frac{V(\text{sign}(\omega \cdot x) \neq \text{sign}(\omega' \cdot x) ; x \in \text{unit-ball})}{V_d} =$$



יחס בין נפחים
=
יחס בזוויות

מכיוון והמשורים עוברים במרכז ⇐ הזווית בין המשורים

$$= \frac{\arccos(\omega \cdot \omega')}{\pi}$$

ה ω המתאים ל h: $\omega = (1,0,0,\dots)$

$$\Rightarrow B(h, \varepsilon/\sqrt{d}) = \{h' \in H ; \frac{\arccos(\omega \cdot \omega')}{\pi} \leq \frac{\varepsilon}{\sqrt{d}}\} =$$

$$= \{h' \in H ; \frac{\arccos(\omega'_1)}{\pi} \leq \frac{\varepsilon}{\sqrt{d}}\}$$

$$= \{h' \in H : \omega'_1 \geq \cos(\frac{\pi\varepsilon}{\sqrt{d}})\}$$

$$\Rightarrow \boxed{DIS[B(h, \frac{\varepsilon}{\sqrt{d}})] = \{x \in \text{unit-ball} : |x_1 \omega_1| \leq \sin(\pi \frac{\varepsilon}{\sqrt{d}})\}}$$