Homework 1

1. LFD Problem 1.3

- **(a) Let $\rho = min_{1 \leq n \leq N}\ y_n(\mathbf{w^*}^T\mathbf{x}_n)$. Show that $\rho > 0$.**

If $\mathbf{w}$ is the optimal weight vector, $\mathbf{w^*}$, then:

$y_n = sign(\mathbf{w^*}^T\mathbf{x}_n)$.

Therefore, if $\mathbf{w^*}^T\mathbf{x}_n$ is positive, then y_n will be positive. This makes ρ positive.
Also, if $\mathbf{w^*}^T\mathbf{x}_n$ is negative, then y_n will be negative as well. This makes ρ positive.
These are the only possible cases. Neither combination can generate a negative value.

This assumes that $\mathbf{w^*}$ and $\mathbf{x}$ are not $\mathbf{0}$.

- **(b) Show that $\mathbf{w}^T(t)\mathbf{w}^* \geq \mathbf{w}^T(t-1)\mathbf{w^*} + \rho$, and conclude that $\mathbf{w}^T(t)\mathbf{w}^* \geq t\rho$.**

1. Base case $t = 1$.

$\mathbf{w}^T(1)\mathbf{w^*} \geq \mathbf{w}^T(0)\mathbf{w^*} + \rho$
Notice that $\mathbf{w}^T(0) = \mathbf{0}$ and that $\mathbf{w}(t) = \mathbf{w}(t-1) + y(t-1)\mathbf{x}(t)$ and get:
$y_i\mathbf{x}_i^T\mathbf{w^*} \geq \rho$. This will always hold because ρ is the minimum of $y_i\mathbf{x}_i^T\mathbf{w^*}$ across all i, so it will at the very least always be equal to the left side.

1. Assume t, prove for $t+1$.

$\mathbf{w}^T(t+1)\mathbf{w}^* \geq \mathbf{w}^T(t)\mathbf{w^*} + \rho$
By definition, the update rule (with both sides multiplied by $\mathbf{w*}$ is:
$\mathbf{w}^T(t+1)\mathbf{w}^* = \mathbf{w}^T(t)\mathbf{w^*} + y(t)\mathbf{x}^T\mathbf{w^*}$
Subsitute this into both sides to get:
$\mathbf{w}^T(t)\mathbf{w^*} + y(t)\mathbf{x}^T\mathbf{w^*} \geq \mathbf{w}^T(t-1)\mathbf{w^*} + y(t-1)\mathbf{x}^T\mathbf{w^*} + \rho$
Since $y(t-1) = y(t)$ we can eliminate $y(t-1)\mathbf{x}^T\mathbf{w^*}$ and $y(t)\mathbf{x}^T\mathbf{w^*}$ from the equation resulting in:

$\mathbf{w}^T(t)\mathbf{w}^* \geq \mathbf{w}^T(t-1)\mathbf{w^*} + \rho$
This is the assumption for t. With (1) and (2) we have proved by induction.

Extending what we just proved many steps back gives us:
$\mathbf{w}^T(t)\mathbf{w}^* \geq \mathbf{w}^T(t-1)\mathbf{w^*} + \rho \geq \mathbf{w}^T(t-2)\mathbf{w^*} + 2\rho \geq \ldots \geq \mathbf{0}^T\mathbf{w^*} + t\rho$
$\mathbf{w}^T(t)\mathbf{w}^* \geq t\rho$

- **(c) Show that $||\mathbf{w}(t)||^2 \leq |\mathbf{w}(t-1)||^2 + |\mathbf{x}(t-1)||^2$**

$||\mathbf{w}(t)||^2 = ||\mathbf{w}(t-1) + y(t-1)\mathbf{x}(t-1)||^2$
$= (\mathbf{w}(t-1) + y(t-1)\mathbf{x}(t-1))^T(\mathbf{w}(t-1) + y(t-1)\mathbf{x}(t-1))$
$= ||\mathbf{w}(t-1)||^2 + 2y(t-1)\mathbf{w}^T(t-1)\mathbf{x}(t-1) + ||y(t-1)\mathbf{x}(t-1)||^2$
The $y(t-1)$ in the last term has no effect because it is +/- 1. So, we get:

$= ||\mathbf{w}(t-1)||^2 + ||\mathbf{x}(t-1)||^2 + 2y(t-1)\mathbf{w}^T(t-1)\mathbf{x}(t-1)$
The first two terms in this expression correspond with the inequality presented in the problem, the final term is always negative because $\mathbf{x}(t-1)$ was misclassified. Therefore, it guarantees that $||\mathbf{w}(t)||^2 \leq |\mathbf{w}(t-1)||^2 + |\mathbf{x}(t-1)||^2$

- **(d) Show by induction that $||\mathbf{w}(t)||^2 \leq tR^2$ where $R = max_{1 \leq n \leq N}||\mathbf{x}_n||$.**

1. $t = 0, 0 \leq 0$
2. Assume t, prove $t + 1$.

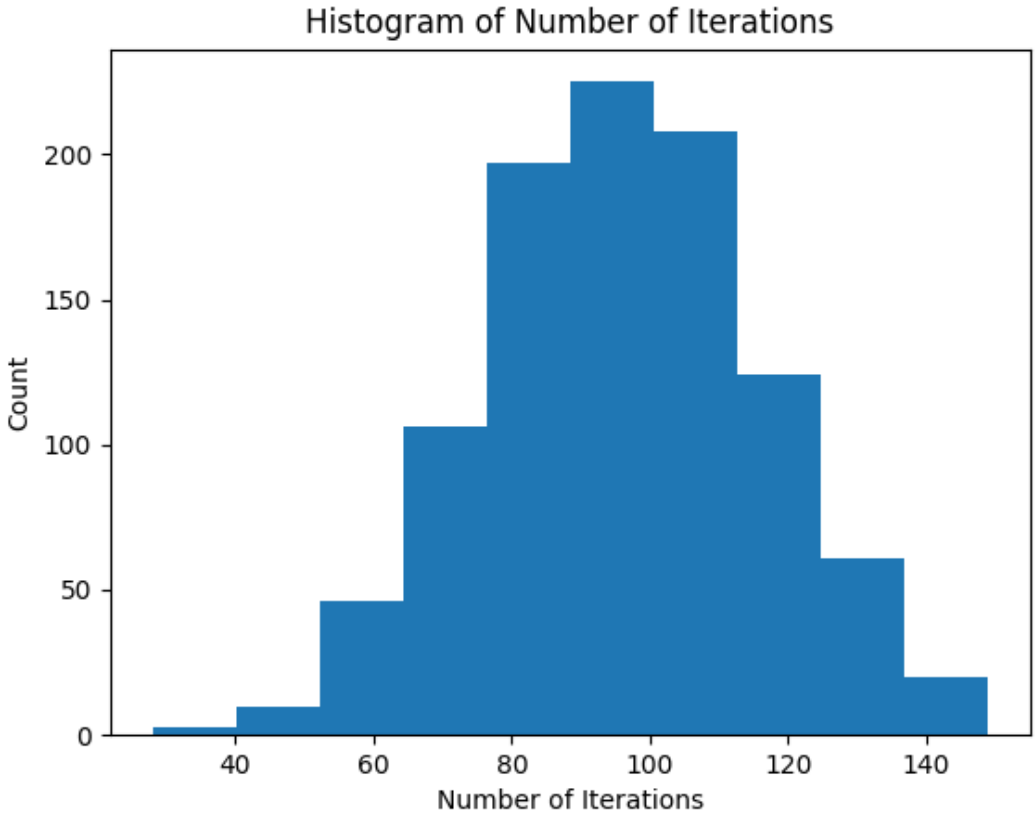
From (c) we have:
$||\mathbf{w}(t)||^2 \leq |\mathbf{w}(t-1)||^2 + |\mathbf{x}(t-1)||^2$
Subsistute $(t-1)R^2$ in from the problem definition, and $R = max_{1 \leq n \leq N}||\mathbf{x}_n||$. This gives us:
$||\mathbf{w}(t)||^2 \leq (t-1)R^2 + R^2$
$||\mathbf{w}(t)||^2 \leq tR^2$

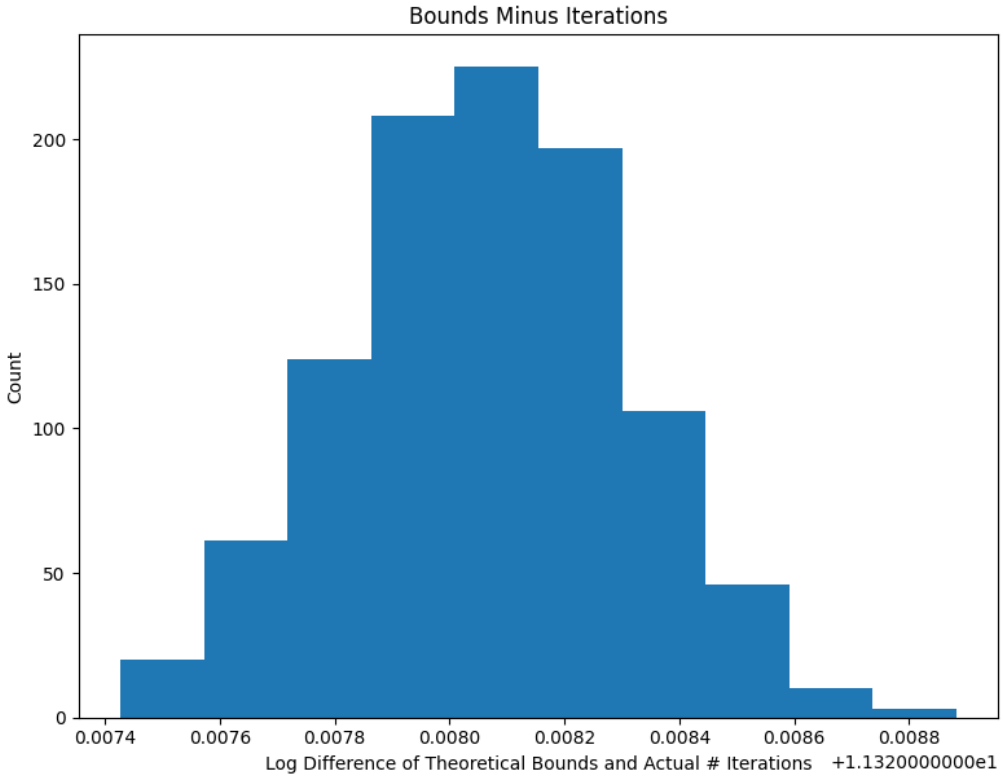
- **(e) Using (b) and (d) show that $\frac{\mathbf{w}^T(t)}{||\mathbf{w}(t)||}\mathbf{w}^* \geq \sqrt{t} \cdot \frac{\rho}{R}$ and hence prove that $t \leq \frac{R^2||\mathbf{w}^*||^2}{\rho^2}$**

Dividing part (b) ($\mathbf{w}(t)\mathbf{w}^* \geq t\rho$) by the square root of part (d) ($||\mathbf{w}(t)|| \leq \sqrt{t}R$) gives us:

$\frac{\mathbf{w}^T(t)}{||\mathbf{w}(t)||}\mathbf{w}^* \geq \sqrt{t} \cdot \frac{\rho}{R}$
Multiplying both sides by $\frac{R}{\rho}$ gives:
$\sqrt{t} \leq \frac{R}{\rho}\frac{\mathbf{w}^T(t)\mathbf{w}^*}{||\mathbf{w}(t)||}$
Squaring both sides and distributing the top weight vectors gives:
$t \leq \frac{R^2||\mathbf{w}^*||^2}{\rho^2}$
This means that PLA will converge if there is an optimal seperator $\mathbf{w}^*$.

2.



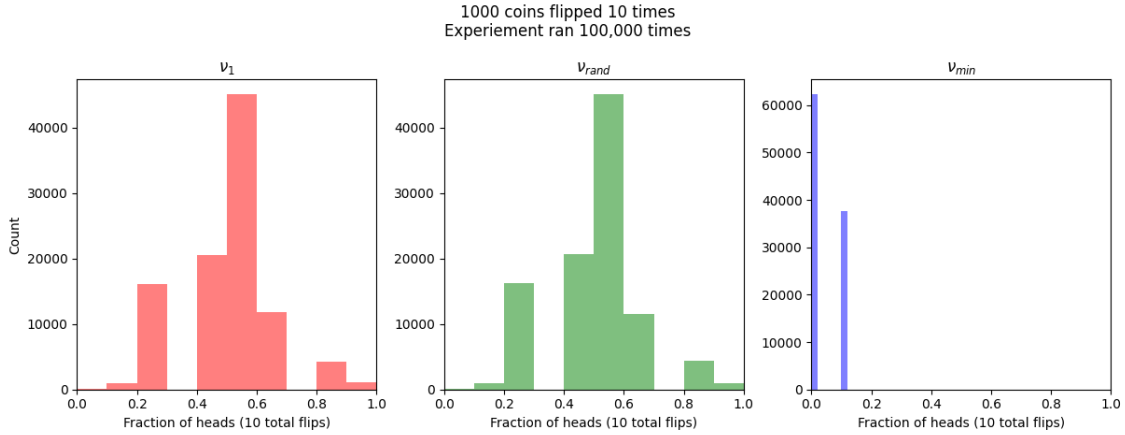


The bound derived in question 1 represents the algorithm making the worst update possible every time it iterates. As long as we randomly choose points to check for misclassification, our algorithm will converge faster than this bound, on average. With the uniformly selected weight

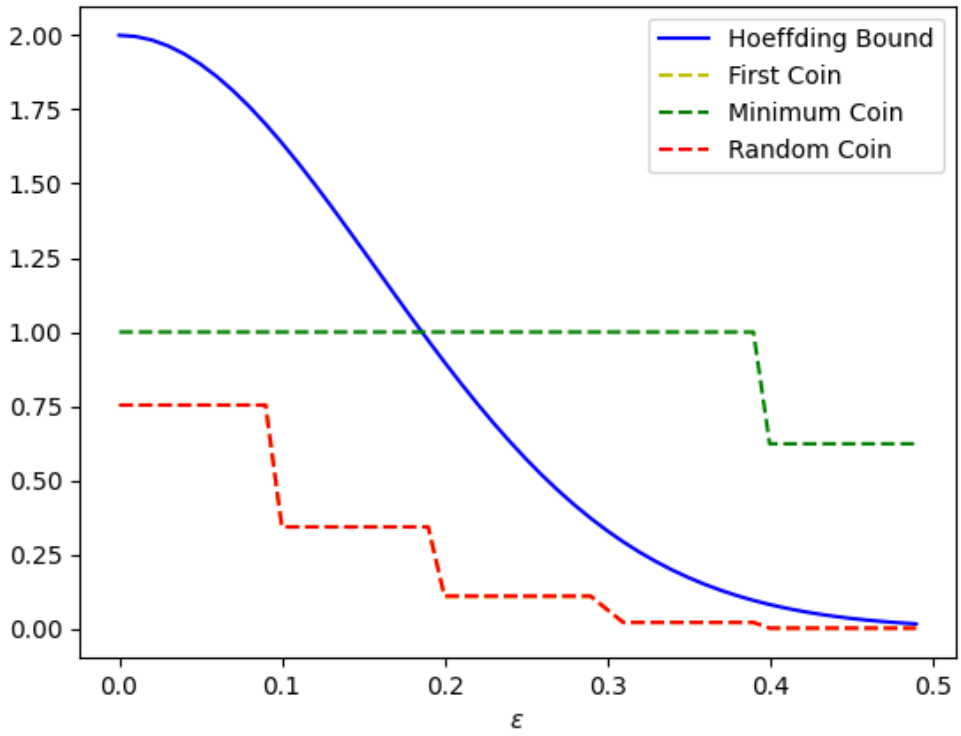
vector and dataset, the number of iterations to converge follows a normal distribution centered around 100. However, this does not correspond with the dimensionality of the dataset.

3.

- **(a)** μ for each of the three coins is 0.5.
- **(b)**



- **(c)**



- **(d)** The random coin and first coin follow the assumptions made in the Hoeffding bound. The minimum coin does not. This is because the minimum coin is not a fixed hypothesis h, instead it is chosen every time after the experiment has occured. This violates the assumption that the Hoeffding bound makes.

Problem 1.8

$t \rightarrow$ non negative random var

(a) $\alpha > 0$, Prove $\mathbb{P}[t \geq \alpha] \leq \mathbb{E}(t)/\alpha$

If we take $\mathbb{I}(t \geq \alpha) \in \{0, 1\}$

Then this will always be true:

$$\alpha \mathbb{I}(t \geq \alpha) \leq t$$

Taking the expected value of both sides:

$$E[\alpha \mathbb{I}(t \geq \alpha)] \leq E[t]$$

$$\alpha E[\mathbb{I}(t \geq \alpha)] \leq E[t]$$

The expected value of $\mathbb{I}$ is just the probability, rearrange to get:

$$\underline{Pr[t \geq \alpha] \leq E[t]/\alpha}$$

(b) If v is any random variable with μ, σ^2, prove

$$\mathbb{P}[(v-\mu)^2 \geq \alpha] \leq \sigma^2/\alpha$$

- $(v-\mu)^2$ is always positive
- subsistute $t = (v-\mu)^2$:

$$\mathbb{P}[(v-\mu)^2 \geq \alpha] \leq \frac{E[(v-\mu)^2]}{\alpha} \rightarrow \text{this is the variance}$$

$$\underline{\mathbb{P}[(v-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}}$$

(c) $v, \ldots v_N$ are iid with μ, σ^2. $v = \frac{1}{N}\sum_{n=1}^{N} v_N$, prove

$$\mathbb{P}[(v-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}$$

- $E[v] = \frac{1}{N}\sum_{n=1}^{N} \mu = \mu$
- $Var[v] = \frac{1}{N^2}\sum_{n=1}^{N} \sigma^2 = \frac{\sigma^2}{N}$
- Subsistute $Var[v]$ for σ^2 in (b) gets us

$$\underline{\mathbb{P}[(v-\mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}}$$

Problem 1.12

N data points $y_1 \leq \dots \leq y_N$

(a) Algorithm uses $E_{in}(h) = \sum_{n=1}^{N} (h - y_n)^2$

Show this equals $h_{mean} = \frac{1}{N} \sum_{n=1}^{N} y_n$

We want to minimize $E_{in}(h)$, so take derative

$$\frac{\partial E_{in}(h)}{\partial h} = \sum_{n=1}^{N} 2(h - y_n) \cdot 1$$

Setting equal to 0:

$$\sum_{n=1}^{N} (h - y_n) = 0$$

$$\sum_{n=1}^{N} h - \sum_{n=1}^{N} y_n = 0$$

$$N(h_{mean}) = \sum_{n=1}^{N} y_n$$

$$h_{mean} = \frac{1}{N} \sum_{n=1}^{N} y_n$$

(b) $E_{in}(h) = \sum_{n=1}^{N} |h - y_n| = \sum_{n=1}^{N} \sqrt{(h - y_n)^2}$

$$\frac{\partial E_{in}(h)}{\partial h} = \sum_{n=1}^{N} \frac{h - y_n}{|h - y_n|} = \sum_{n=1}^{N} \text{sign}(h - y_n) = 0$$

→ to accomplish this, exactly half of h must be positive and half of h must be negative. This means h must equal h_{med}.

(c) $y_N = y_N + \varepsilon$ where $\varepsilon \to \infty$

$h_{mean} \to \infty$ because $h_{mean} = \frac{1}{N} \sum_{n=1}^{N} y_N + \varepsilon$ → goes to ∞

h_{med} stays the same so long as half $> y$, and half $< y$

Problem 2.3

Compute max number of dichotomies $m_H(N)$ for models:

(a) **Positive / Negative ray**

Positive: $m_H(N) = N+1$

Negative: $m_H(N) = N+1$

But, all + and all − are always repeated

so $m_H(N) = \underbrace{N+1}_{\text{Positive ray}} + \underbrace{N+1}_{\text{Negative ray}} \underbrace{-2}_{\text{repeats}} = \boxed{2N}$

$\boxed{d_{VC} = 2}$

(b) **Positive / Negative Interval**

Positive/Negative intervals: $m_H(N) = \frac{1}{2}N^2 + N/2 + 1 = \binom{N+1}{2} + 1$

What is repeated? all +

all −

But can get $N-2$ new dichotomies w/ negative

so $m_H(N) = \binom{N+1}{2} + 1 + \binom{N-1}{2}$

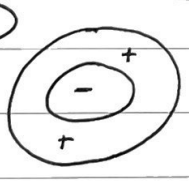
$= \frac{1}{2}N^2 + \frac{1}{2}N + 1 + \frac{1}{2}N^2 - \frac{3}{2}N + 1$

$= \boxed{N^2 - N + 2}$

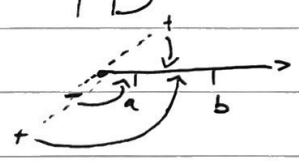
$\boxed{d_{VC} = 3}$ as $m_H(4) = 14$

(c) **Spheres in $\mathbb{R}^d$** $+1$ for $a \le \sqrt{x_1^2 + \dots x_d^2} \le b$

2D



1D



Mapping from ND → 2D → 1D shows that this is just like a positive interval. So long you can map the interval back to ND, we use the same growth function $m_H(N) = \binom{N+1}{2} + 1$ and $\boxed{d_{VC} = 3}$

Problem 2.8

Which are possible growth functions?

✓ $1+N$: Formula 2.10 from LFD gives us $m_H(N) \leq N^{d_{VC}}+1$

$d_{VC}=1$ here so gives us:

$$1+N \leq N^1+1$$

This will always hold, so $1+N$ is a possible function.

✓ $1+N+\frac{N(N-1)}{2}$: $d_{VC}=2$

so $1+N+\frac{N^2}{2}+N \leq N^2+1$

The N^2 coefficient is smaller on the left so this is possible.

✓ 2^N: $d_{VC}=\infty$, but this is the growth function if d_{VC} is ∞ by definition so this is possible.

✗ $2^{\lfloor\sqrt{N}\rfloor}$: $d_{VC}=1$ so:

$$2^{\lfloor\sqrt{N}\rfloor} \leq N^1+1$$

This is fine at $N=10, 20$, bot at $N=30$

$2^5 \not\leq 31$ so this is impossible.

✗ $2^{\lfloor N/2\rfloor}$: $d_{VC}=0$ so bounded by $N^0+1=2$. This will never be true above 4. so not possible

✗ $1+N+\frac{N(N+1)(N-2)}{6}$: $d_{VC}=1$

can't be bounded by N^1+1, so not possible.