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 $t + 1$.

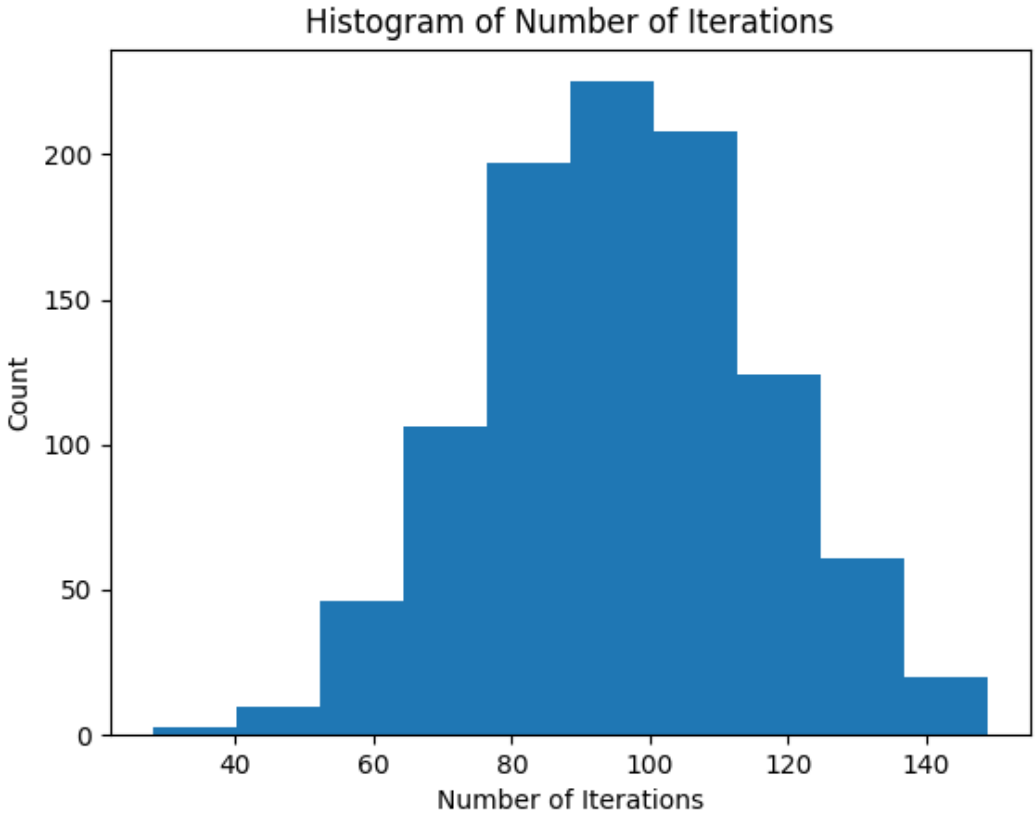
From (c) we have:
$||\mathbf{w}(t)||^2 \leq |\mathbf{w}(t-1)||^2 + |\mathbf{x}(t-1)||^2$
Subsistute $(t-1)R^2$ in from the problem definition, and $R = max_{1 \leq n \leq N}||\mathbf{x}_n||$. This gives us:
$||\mathbf{w}(t)||^2 \leq (t-1)R^2 + R^2$
$||\mathbf{w}(t)||^2 \leq tR^2$

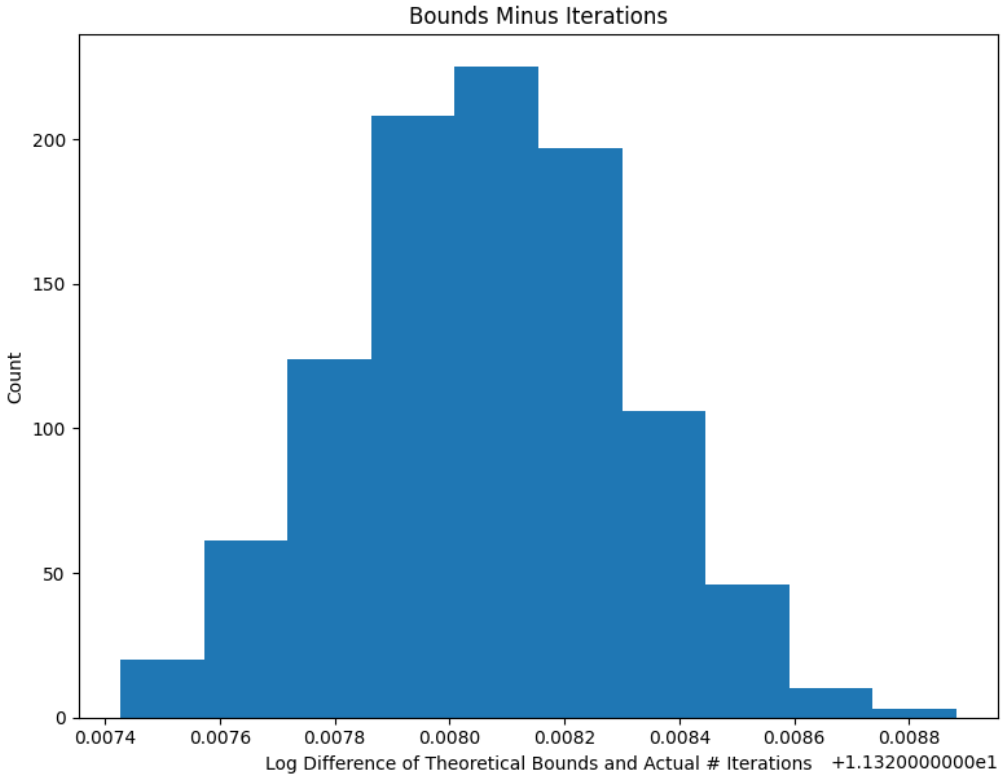
- **(e) Using (b) and (d) show that $\frac{\mathbf{w}^T(t)}{||\mathbf{w}(t)||}\mathbf{w}^* \geq \sqrt{t} \cdot \frac{\rho}{R}$ and hence prove that $t \leq \frac{R^2||\mathbf{w}^*||^2}{\rho^2}$**

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.



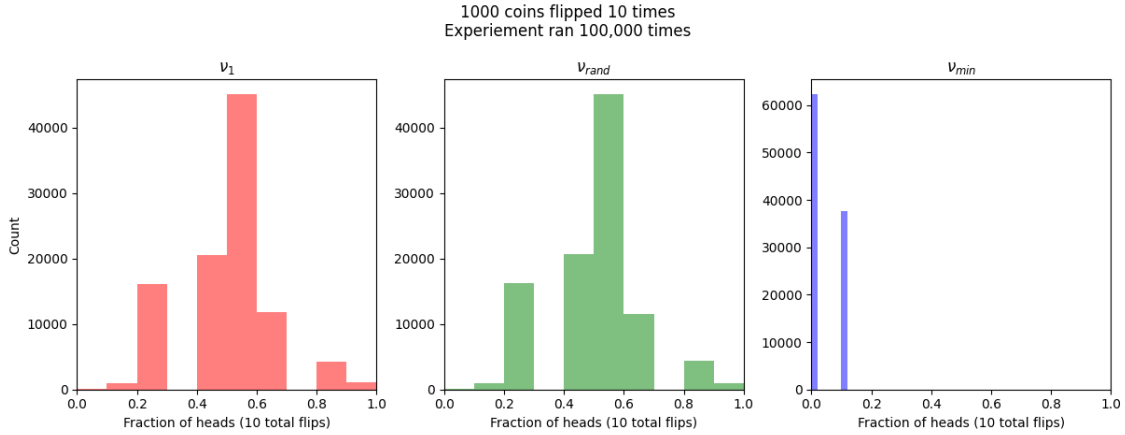


The bound derived in question 1 represents the algorithm making the worst update possible every time it iterates. As long as we randomly choose points to check for misclassification, our algorithm will converge faster than this bound, on average. With the uniformly selected weight

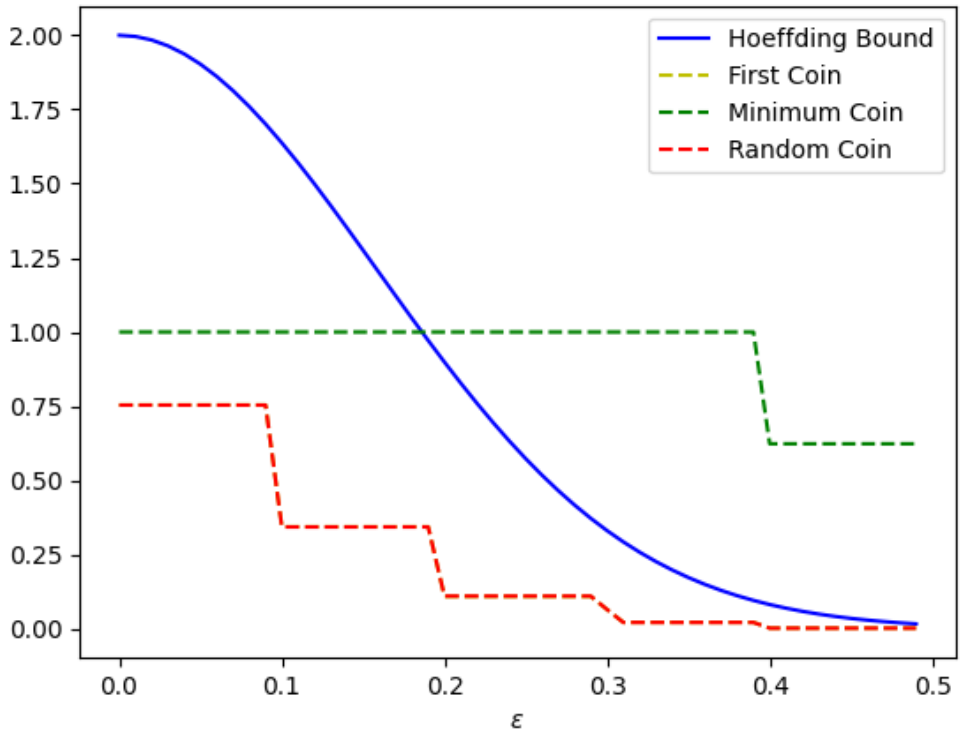
vector and dataset, the number of iterations to converge follows a normal distribution centered around 100. However, this does not correspond with the dimensionality of the dataset.

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.