CSE 417A: Homework 1 Solution Sketch

September 5, 2022

Note: These are not intended to be comprehensive, just to help you see what the answers should be.

Problem 1.3

(a) By definition, $\boldsymbol{w}^*$ separates the data. Therefore, $y_n(\boldsymbol{w}^{*T}\boldsymbol{x}_n) > 0$ for every n, and so $\rho > 0$.

(b) We first show $\boldsymbol{w}^T(t)\boldsymbol{w}^* \geq \boldsymbol{w}^T(t-1)\boldsymbol{w}^* + \rho$.
Let $\boldsymbol{x}(t-1), y(t-1)$ be the misclassified example at $t-1$, from PLA, we have $\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + y(t-1)\boldsymbol{x}(t-1)$. Therefore,

$$\boldsymbol{w}^T(t)\boldsymbol{w}^* = \boldsymbol{w}^T(t-1)\boldsymbol{w}^* + y(t)\boldsymbol{x}^T(t)\boldsymbol{w}^* \geq \boldsymbol{w}^T(t-1)\boldsymbol{w}^* + \rho$$

Next we prove $\boldsymbol{w}^T(t)\boldsymbol{w}^* \geq t\rho$ by induction
Since $\boldsymbol{w}(0) = \boldsymbol{0}$. The statement is true at $t = 0$. Assume $\boldsymbol{w}^T(t)\boldsymbol{w}^* \geq t\rho$ is true at t. Using the above result, we have

$$\boldsymbol{w}^T(t+1)\boldsymbol{w}^* \geq \boldsymbol{w}^T(t)\boldsymbol{w}^* + \rho \geq (t+1)\rho$$

(c) Again, we have $\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + y(t-1)\boldsymbol{x}(t-1)$, so

$$\|\boldsymbol{w}(t)\|^2 = \|\boldsymbol{w}(t-1)\| + \|\boldsymbol{x}(t-1)\|^2 + 2y(t-1)\boldsymbol{w}^T(t-1)\boldsymbol{x}(t-1)$$

Since $\boldsymbol{x}(t-1)$, $y(t-1)$ is the misclassified example at $t-1$, $y_t\boldsymbol{w}^T(t-1)\boldsymbol{x}_t \leq 0$. Therefore $\|\boldsymbol{w}(t)\|^2 \leq \|\boldsymbol{w}(t-1)\| + \|\boldsymbol{x}(t-1)\|^2$.

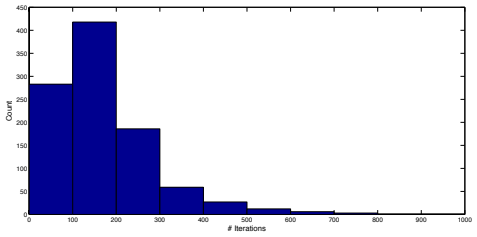
(d) We prove by induction. Since $\boldsymbol{w}(0) = \boldsymbol{0}$. The statement is true at $t = 0$. Assume the statement is true at $t-1$, i.e., $\|\boldsymbol{w}(t-1)\|^2 \leq (t-1)R^2$. Since $\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + y(t-1)\boldsymbol{x}(t-1)$, we have

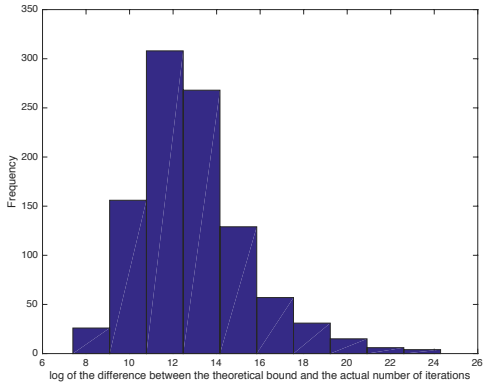
$$\|\boldsymbol{w}(t)\|^2 \leq \|\boldsymbol{w}(t-1)\|^2 + \|\boldsymbol{x}(t-1)\|^2 \leq \|\boldsymbol{w}(t-1)\|^2 + R^2 \leq (t-1)R^2 + R^2 \leq tR^2$$

(e) Since $\|\boldsymbol{w}(t)\|\|\boldsymbol{w}^*\| \geq \boldsymbol{w}^T(t)\boldsymbol{w}^*$. From (b), we know $\|\boldsymbol{w}(t)\|\|\boldsymbol{w}^*\| \geq t\rho$. From (d), we know $\|\boldsymbol{w}(t)\| \leq R\sqrt{t}$. Therefore,

$$R\sqrt{t}\|\boldsymbol{w}^*\| \geq \|\boldsymbol{w}(t)\|\|\boldsymbol{w}^*\| \geq t\rho$$

PLA Expt An example histogram of the number of iterations is shown below, without representing four cases that took more than 1000 iterations (the maximum was 3834). If we compare with the bound derived in class, we see that the bound is **very** loose, in fact, orders of magnitude so (see histogram of the log differences below).

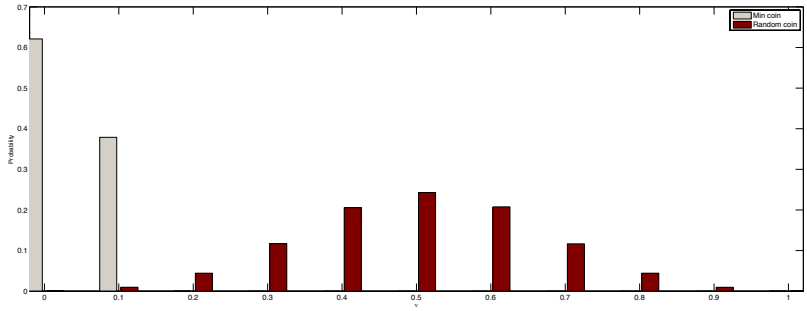


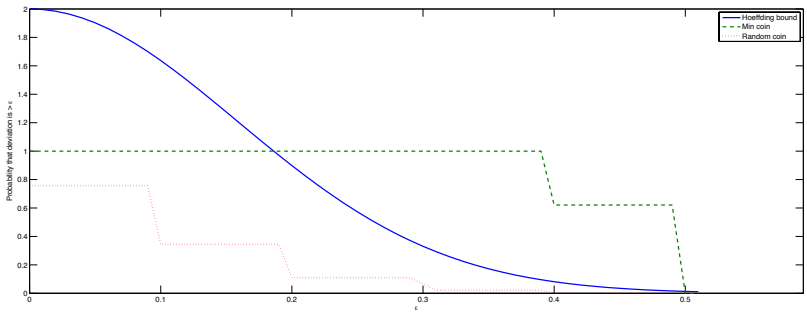


Exercise 1.10 For part (a), $\mu = 0.5$ for all the coins. A single line of matlab code that works is

```
mean(randi([0,1],10,1000),1)
```

For parts (b) and (c), your graphs should look something like this (the histograms for the random coin and coin 1 should look almost identical, so we only show one of them here):





For parts (d) and (e), the main thing is that coin 1 and the random coin are both the equivalent of hypotheses that are selected before looking at the data, so the Hoeffding bound applies (it's hypothesis verification). On the other hand the data is used to choose the "minimum" coin out of many, so the

bound doesn't apply, equivalent to the multiple (1000) bins model in the text that we discussed in class.

Problem 1.8 For part (a) note that $E(t) = E(t|t < \alpha)\Pr(t < \alpha) + E(t|t \geq \alpha)\Pr(t \geq \alpha)$. Since the first term is non-negative and $E(t|t \geq \alpha) \geq \alpha$, we have

$$E(t) \leq \alpha \Pr(t \geq \alpha).$$

For part (b), let $t = (u - \mu)^2$, which is non-negative. Now, $E(t) = \sigma^2$. Apply part (a) to t:

$$\Pr((u - \mu)^2 \geq \alpha) = \Pr(t \geq \alpha) \leq \frac{E(t)}{\alpha} = \frac{\sigma^2}{\alpha}$$

For part (c), since u is the sum of N i.i.d. u_n, each with variance σ^2, the variance of u is σ^2/N (standard result). Now, applying part (b) with the variance σ^2/N gives us the result.

Problem 1.12 Part (a) is a standard, simple minimization. $\frac{\partial E_{\text{in}}(h)}{\partial h} = \sum_{n=1}^{N}(2h - 2y_n)$. Setting to zero, we get $h = \frac{\sum_{n=1}^{N} y_n}{N}$ and since the second derivative is positive, this is a minimum.

Part (b) can be done in a couple of different ways: one is to split it up piecewise for minimization. Alternatively, here's a nice proof. If you choose any point that is not a median, you must have more points either on the left or on the right of that point. Assume you have more points on the left, that is $\sum_n [\![y_n \geq h]\!] < \sum_n [\![y_n \leq h]\!]$. Then, if you decrease h by δ without changing how many points are on each side, you decrease the error by $\delta(\sum_n [\![y_n \leq h]\!] - \sum_n [\![y_n \geq h]\!]) > 0$. A similar argument applies if you have more points on the right. Therefore, E_{in} can only be minimized if $\sum_n [\![y_n \geq h]\!] = \sum_n [\![y_n \leq h]\!]$.

Finally, for part (c), $h_{\text{mean}} \to \infty$ but h_{med} is not affected.

Problem 2.3

(a) First count dichotomies that have at least one positive and one negative point. There are $N - 1$ intervals between N points where one can start a positive or negative ray, so there are $2(N - 1)$ such dichotomies. There are two more that are all positive and all negative, giving a total of $2N$, thus $d_{\text{VC}} = 2$ (3 is a break point, 2 is not).

(b) Positive or negative intervals alone can implement $1 + \binom{N+1}{2}$. Just using $2(1 + \binom{N+1}{2})$ overcounts those dichotomies that can be implemented by both – that number is in fact exactly equal to the number that can be implemented by the positive or negative ray, so the total number that can be implemented is $2(1 + \binom{N+1}{2}) - 2N = N^2 - N + 2$. $k = 3$ is not a break point, but $k = 4$ is, so $d_{\text{VC}} = 3$.

(c) In terms of dichotomies, this is equivalent to positive intervals on the positive real numbers (think about it as choosing the radius). So this is again $1 + \binom{N+1}{2}$. $d_{\text{VC}} = 2$.