① $f: \{0,1\}^d \to \{0,1\}$

$AND_S(x) = \bigwedge_{i \in S} x_i \quad S \subseteq [d]$

$(x^1, y_1), \ldots, (x^m, y_m)$ where $x^l \in \{0,1\}^d$ & $y_l = AND_{S*}(x^l)$ for non-empty $S*$

Initialize set of features S to be $[d]$, full set of features

For each example (x^l, y_l):

① $\hat{y}_l = AND_S(x^l)$ where S is current set of features

② If $\hat{y}_l \neq y_l$ UPDATE set of features. Remove all features x_i such that $x_i = 0$ in x_l, Continue to next example

At end, output set of remaining features, $\hat{S}$.

Predict $\hat{y}(x) = AND_{\hat{S}}(x)$ for any new input, x

- Since we removed only features that we know cannot be in $S*$, the above algorithm shows that the true set of features $S*$ will be a subset of remaining features at end.
- Since the algorithm removes at least one feature per mistake, it will make at most d mistakes in entire run-time, which is a polynomial bound
- Run-time is also polynomial in d since each example can be processed in $O(d)$. There are at most d examples seen before true set of features is identified. So, the overall run-time is $O(d^2)$

(2.) Consider the following scenario with 2 experts A & B.

- A predicts 1 for each day, t
- B predicts 0 for each day, t
- Since the algorithm is deterministic, an adversary can fix all outcomes such that the predictions are always wrong.

  Then, at least 1 of A & B will have an error rate of ≤ 0.5, & the algorithm error rate is 1.

- Thus, showing that no deterministic algorithm can do better than a factor of 2 compared to the best expert

③ Theres some $W*$ & we see $(x^1, y^1), \dots, (x^T, y^T)$:

$y^T = \text{sign}(\langle W*, x_i \rangle)$, $w_i^* \geq 0$, $\sum_{i=1}^{d} w_i^* = 1$

MWM:

① $w(0,i) = 1 \quad \forall i \in [d]$, $u_i^0 = \frac{w(0,i)}{\sum_{j=1}^{d} w(0,j)}$

② for each step $t = 1, \dots, T$

Predict: $\text{sign}(\langle u^{t-1}, x^t \rangle)$

Update: if correct $w(t,i) = w(t-1,i) \quad \forall i$

$u^t = u^{t-1}$

else $L(t,i) \in [0,1]$

$w(t,i) = (1 - \varepsilon \cdot L(t,i))\, w(t-1,i)$

u^t = normalized vector of $w(t)$

from office hours: $L(t,i) = -y^t x_i^t$

Regret Bound: $\mathbb{E}[L(T)] \leq L_*(T) + O(\sqrt{T \ln d})$

$$\sum_{t=1}^{T} \left[\sum_{i=1}^{d} \Pr[\text{pick expert } i \text{ at step } t] \cdot L(t,i) \right] = \sum_{t=1}^{T} \sum_{i=1}^{d} -y^t u_i^t x_i^t = \sum_{t=1}^{T} -y^t \langle u^t, x^t \rangle$$

Loss of Best Expert: let L^t be vector of losses on step t, $L^t \in [-1,1]^d$

$\sum_{t=1}^{T} L^t$ vector of losses for each expert

$$L_*(T) = \min_i \sum_{t=1}^{T} L^T \leq \langle W^*, \sum_{t=1}^{T} L^t \rangle$$

$$\Rightarrow \mathbb{E}[L(T)] \leq \sum_{t=1}^{T} \langle W^*, L^t \rangle + O(\sqrt{T \ln d})$$ // sum term = 0 if correct

$L_i^t = -y^t x_i^t$ if not

$$\mathbb{E}[L(T)] \leq (\#\text{ mistakes})(-\gamma) + O(\sqrt{T \ln d})$$

$$\Rightarrow (\#\text{ mistakes}) \cdot \gamma + \mathbb{E}[L(T)] \leq O(\sqrt{T \ln d})$$

④ a.) $\{(0,1), (1,0)\}$

To determine the line of best fit through the origin we know that we want to maximize the following expression:

$$\max_{v} \sum_{x \in X} \| \mathrm{Proj}_L(x) \|_2^2 \equiv \max_{v:\|v\|_2 = 1} \sum_{x \in X} \langle x, v \rangle^2$$

We know: $\| \mathrm{Proj}_v(x) \|_2 = \dfrac{\langle x, v \rangle}{\|v\|_2}$

So, the problem boils down to:

$$\max_{v:\|v\|_2=1} \langle (x_1, y_1), \vec{v} \rangle^2 + \langle (x_2, y_2), \vec{v} \rangle^2$$

For the points $(0,1)$ $ $(1,0)$ we get

$$\max_{v:\|v\|_2=1} \langle (1,0), \vec{v} \rangle^2 + \langle (0,1), \vec{v} \rangle^2 \Rightarrow v_1^2 + v_2^2 = 1$$

- So, the best fit line is <u>not</u> unique for this problem. Any line that passes through the origin will be the line of best fit.

b.) $\cdot \{(0,1),(2,0)\}$

- Following the same analysis as before, but with the new points we get:

$$\max_{V:\|V\|_2=1} \langle (0,1),\vec{V}\rangle^2 + \langle (2,0),\vec{V}\rangle^2 \Rightarrow V_2^2 + 4V_1^2$$

$\Rightarrow \quad \vec{V} = \langle 1,0\rangle$

Thus the line of best fit is the X-axis. In this case the best-fit line <u>is</u> unique. If we think about it geometrically this makes sense as moving away from the X-axis will have adverse effects because the (2,0) point has more "power" than the (0,1) point.

c) $\circ \{(0,-1), (2,0)\}$

- Following the same analysis as before, but with the new points we get:

$$\max_{V:\|V\|_2=1} \langle (0,-1), \vec{V} \rangle^2 + \langle (2,0), \vec{V} \rangle^2 \Rightarrow V_2^2 + 4V_2^2$$

$$\Rightarrow \quad \vec{V} = \langle 1,0 \rangle$$

Again, the line of best-fit is the x-axis. The line of best-fit <u>is</u> unique for the same reasons as before

(5) If we were given a dataset & asked to find the best-fit line that doesn't necessarily need to go through the origin, I would do the following:

- "Center" the data by subtracting the mean of each dimension from each data point, which basically shifts the origin to the center of the data
- Use the same approach as discussed in class to find the line of best-fit
- Re-shift the line to the correct position by using the means again