1. ALGORITHM:

1. Consider S to be the set of all features.

2. For $t = 1 \ldots T$

2.1. For the given x_t, $\hat{y}_t = AND_S(x_t)$

2.2. If the predicted output is correct

$$\hat{y}_t = y_t,$$

do nothing.

2.3. In case of a mistake, the mistake has to be that we predicted 0 and answer was 1.

In this case remove all the i in S, for which $x_{(t,i)}$ was false (0).

If we predicted 1 and answer is 0, this means solution does not exist. But as the question says solution always exist, we can assume that we will never reach this state.

MISTAKE-BOUND:

For each time a mistake occurs (i.e. we predicted 0 when the answer was 1), atleast one $i \in S$ existed which gave $x_{(t,i)} = \text{FALSE}$ and this will be weeded out. So, for each mistake, atleast one i is weeded out.

$M \Rightarrow$ total mistakes.

$N_t \Rightarrow$ Number of $i \in S$ removed till time t.

$$M \leq N_t \qquad - (1)$$

At time T, we know that finally there exists an answer S_* and it cannot be empty. So

$$|S_*| \geq 1$$

$N_0 = 0$ and totally there are d features

So $N_t \leq N_T < d \quad - (2)$

From (1) and (2)

$$\boxed{M < d}$$

RUNNING TIME:

For each input x_t, if it makes a mistake, we check all $i \in S$ and perform AND operation and remove/keep accordingly. At any time $|S| \leq d$,

So running time is $O(d)$ //

POLYNOMIAL.

2. We can take an example of 2 experts and prove that no deterministic algorithm can do better than a factor 2 approximation to the best expert loss.

Assume 2 experts, where expert 1 always chooses 0 and expert 2 always chooses 1. Now, whatever we predict for each day, the adversary could decide the final answer as opposite of it, i.e. it is always possible that our predictions were 100% wrong. So if there are T days,

$$L(T) = T.$$

Now whatever the final output be, atleast one of 0 or 1 has to occur $\geq 50\%$ of the days. In other words, as each expert predicts only 0 or only 1, atleast one of the expert is correct $\geq 50\%$ of the days.

$$L_*(T) \leq T/2 \qquad T \geq 2L_*(T)$$

Therefore $L(T) \geq 2L_*(T)$

NO DETERMINISTIC ALGORITHM can do better than factor 2 approximation to the best expert loss.

As it is impossible for 2 experts, it is impossible for the more generic n experts as 2 experts is a subset. Hence proved //

→ We can extrapolate this to n experts. Lets divide them into 2 groups. One group predicts 0, other 1.

So whatever our output $L(T) = T$

Some experts, either group 1 or 2 will have

$$L_*(T) \leq T/2$$

Therefore $L(T) \geq 2L_*(T)$ //.

3. ALGORITHM:

1. $w[0,i] = 1 \quad , i = 1, \dots, n$

2. On each day, $t = 1, \dots, T$:

2.1. Pick an expert i with probability $\propto w(t-1,i)$

$$Pr[\text{Expert } i \text{ is picked}] = \frac{w(t-1,i)}{\sum_{j=1}^{n} w(t-1,j)}$$

* Then follow expert i's output $\in [0,1]$.

2.2. Update weight of every expert:

* Compute loss of each expert

$L(t,j) = |$Prediction of expert j on day t − True value on day $t|$

* $w(t,j) = (1-\varepsilon)^{L(t,j)} \cdot w(t-1,j)$

To PROVE:

$$E[L(T)] \leq (1+\varepsilon) A_*(T) + (\ln n)/\varepsilon$$

$A_* T \Rightarrow$ loss of the best expert after T days.

We know

$$L(t) = E[\text{loss incurred on day } t]$$

$$= \sum_{i=1}^{n} \Pr[\text{we picked expert } i] \cdot L(t,i)$$

$$= \sum_{i=1}^{n} \frac{\omega(t-1,i)}{\underbrace{\sum_{j=1}^{n} \omega(t-1,j)}_{\omega(t-1)}} \cdot L(t,i)$$

$$L(t) = \frac{1}{\omega(t-1)} \cdot \sum_{i=1}^{n} \omega(t-1,i) \cdot L(t,i) \quad - (1)$$

We know $(1-a)^x \le (1-ax)$ for $0 < x < 1$

$$\omega(t) = \sum_{i=1}^{n} \omega(t,i) = \sum_{i=1}^{n} \omega(t-1,i) \cdot (1-\varepsilon)^{L(t,i)}$$

$$\le \sum_{i=1}^{n} \omega(t-1,i) \cdot (1 - \varepsilon \cdot L(t,i))$$

$$= \sum_{i=1}^{n} \omega(t-1,i) - \varepsilon \sum_{i=1}^{n} \omega(t-1,i) \cdot L(t,i) - (2)$$

(1) in (2)

$$\omega(t) \le \omega(t-1) - \varepsilon\, \omega(t-1) L(t)$$

$$\le \omega(t-1)(1 - \varepsilon L(t))$$

We know

$$1 - x \leq e^{-x} \quad \forall x$$

So $$w(t) \leq w(t-1)\left(1 - \varepsilon L(t)\right) \leq w(t-1) \cdot e^{-\varepsilon L(t)}$$

Therefore

$$w(T) \leq w(0) \cdot e^{-\varepsilon L(1)} \cdot e^{-\varepsilon L(2)} \cdots e^{-\varepsilon L(T)}$$

$$= w(0) \cdot e^{-\varepsilon \underbrace{(L(1) + L(2) + \cdots + L(T))}_{\text{Our total loss}}}$$

$$w(T) \leq w(0) \cdot e^{-\varepsilon A(T)} \quad - \quad ③$$

where $A(T) =$ Our total expected loss.

Given $A_*(T)$ is loss of the best expert after T days,

we know that

$$w(T) \geq (1-\varepsilon)^{A_*(T)} \quad - \quad ④$$

From ③ and ④

$$(1-\varepsilon)^{A_*(T)} \leq w(0) \cdot e^{-\varepsilon A(T)}$$

$$w(0) = n$$

So $$(1-\varepsilon)^{A_*(T)} \leq n \cdot e^{-\varepsilon A(T)}$$

$$A_*(T)\ln(1-\varepsilon) \le \ln(n) - \varepsilon A(T)$$

$$\varepsilon A(T) \le \left(-\ln(1-\varepsilon)\right) \cdot A_*(T) + \ln(n)$$

$$A(T) \le \left(\frac{-\ln(1-\varepsilon)}{\varepsilon}\right) A_*(T) + \frac{\ln n}{\varepsilon}$$

Given for

$$0 \le \varepsilon \le 1/2$$

$$\frac{-\ln(1-\varepsilon)}{\varepsilon} \le 1+\varepsilon$$

So $0 \le \varepsilon \le 1/2$

$$A(T) \le (1+\varepsilon) A_*(T) + \frac{\ln n}{\varepsilon}$$

$$E[L(T)] \le (1+\varepsilon) A_*(T) + \frac{\ln n}{\varepsilon} \quad //$$

HENCE PROVED//

4. Assume each feature of x is an expert. So d experts.

ALGORITHM:

1. $w(0,i) = 1/d \quad , i = 1, \dots, d$

2. On each day, $t = 1, \dots, T$:

2.1. Give each expert i a weight $w(t-1,i)$

Let $$p(t,i) = \frac{w(t-1,i)}{\sum_{j=1}^{d} w(t-1,j)}$$

* Then predict

$$\hat{y}_t = \text{sign}(\langle p_t \cdot x_t \rangle)$$

$$\hat{y}_t = \text{sign}\left(\sum p(t,i)\, x(t,i) \right)$$

y_t is $\text{sign}(\langle w_*, x_t \rangle)$

2.2. Update weight when we are wrong:

- Compute loss of each expert

$$\boxed{L(t,i) = -y_t\, x(t,i)}$$

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

Our loss function is $-y_t x(t,i)$. If $x(t,i)$ matches in sign to y_t, we are rewarding the coordinate. Else, we are penalizing them.

From MWM, we have

$$A(T) \le L_*(T) + 2\sqrt{T \ln d} \quad \sim \text{①}$$

$$L(t,i) = -y_t x(t,i)$$

$$L_*(T) = \min_i \sum_{t=1}^{T} \sum_{i=1}^{d} L(t,i)$$

$$= \min_i \sum_{t=1}^{T} -y_t * x(t,i)$$

Minimum across all i will be less than the weighted average.

$$L_*(T) \le \sum_i w_i^* \sum_{t=1}^{T} -y_t * x(t,i)$$

$$\le -\sum_{t=1}^{T} y_t \sum_i w_i^* \cdot x(t,i)$$

$$\le -\sum_t y_t \langle w^*, x(t) \rangle$$

We know that $y_t \langle w^*, x(t) \rangle > \gamma$

$$L_*(T) < -\sum_t \gamma$$

$$L_*(T) < -M\gamma \quad - ①$$

where M is the number of mistakes.

$A(T)$ = Total expected loss

$$A(T) = \sum_{t=1}^{T} \sum_{i=1}^{d} w(t,i)\,|-y_t \cdot x(t,i)|$$

$$= \sum_{t=1}^{T} -y_t \cdot \langle w(t), x(t) \rangle$$

So, it is a loss only when y_t and $\hat{y}_t = \langle w(t), x(t) \rangle$ are opposite in signs. So $A(T)$ is always +ve.

$$A(T) \geq 0 \quad - ③$$

Using ①, ② and ③

$$0 < -M\gamma + 2\sqrt{T \ln d}$$

$$M\gamma < 2\sqrt{T \ln d}$$

$$M < \frac{2\sqrt{T \ln d}}{\gamma}$$

$$M \equiv O\left(\sqrt{T \ln d} / \gamma\right)$$

HENCE PROVED //