

1 The Mistake Bound Model

Suppose there are N experts who make predictions about a certain event every day – for example, whether it rains today or not. At the beginning of each time step t , the experts give their predictions in a vector \mathcal{E}^t . We then also make a prediction about the outcome. Finally, we see the actual outcome o^t . The goal is to minimize the number of times our prediction differs from the outcome.

Example 12.1. Suppose $\mathcal{E}^t = (0, 1, 0, 0, 0, 1, 1, 1, 0)$. We predict 0, but the actual outcome is $o^t = 1$, so this is a mistake.

Fact 12.2. *If there is a perfect expert, then there is an algorithm that makes at most $\log_2 N$ mistakes.*

Proof. At each step, look at the all the experts who have made no mistakes so far. Predict what the majority of them predict. Note that everytime we make a mistake, the number of experts who have not been wrong yet is cut in half. Since there is at least one perfect expert, we can make at most $\log N$ mistakes. \square

Fact 12.3. *If the best expert makes m mistakes, there is an algorithm that makes at most $m(\log_2 N + 1) + \log_2 N$ mistakes.*

Proof. Think of time as divided into “epochs”. In each epoch, we proceed as in the perfect expert scenario, and keep track of all experts who have not made a mistake in that epoch. This set halves with every mistake the algorithm makes. When this set becomes empty, we have made at most $\log_2 N + 1$ mistakes, and every expert has made a mistake. This epoch is over and we start the next epoch.

Note that in each epoch, every expert makes a mistake. Therefore the number of completed epochs is at most m , and the number of mistakes our algorithm makes in these epochs is at most $m(\log_2 N + 1)$. In the last (current) epoch, the algorithm makes at most $\log_2 N$ mistakes as in Fact 12.2. \square

2 The Weighted Majority Algorithm

- Assign a weight w_i to expert i . Let $w_i^{(t)}$ denote the weight of expert i at time t . Initially, all weights are 1 ($w_i^{(1)} = 1$).
- At each time t , predict according to the weighted majority of experts using their weights. In other words, choose the outcome that maximizes the sum of weights of experts that predicted it.
- When we see the outcome, set $w_i^{(t+1)} = w_i^{(t)} \cdot \begin{cases} 1 & \text{if } i \text{ was correct} \\ \frac{1}{2} & \text{if } i \text{ was incorrect} \end{cases}$.

Theorem 12.4. *For all times t and experts i , the number of mistakes the weighted majority algorithm (WM) makes is at most $2.41(m_i + \log_2 N)$, where m_i is the number of mistakes expert i makes.*

Proof. The proof uses a potential function argument. Let $\Phi^t = \sum_{i \in [N]} w_i^{(t)}$. Note that

- $\Phi_1 = N$
- $\Phi_{t+1} \leq \Phi_t$ for all t
- If WM makes a mistake at time t , then the weight of the wrong experts is higher than the weight of the correct experts. Then

$$\begin{aligned}
 \Phi^{t+1} &= \sum_{i \text{ wrong}} w_i^{(t+1)} + \sum_{i \text{ correct}} w_i^{(t+1)} \\
 &= \frac{1}{2} \sum_{i \text{ wrong}} w_i^{(t)} + \sum_{i \text{ correct}} w_i^{(t)} \\
 &= \Phi^t - \frac{1}{2} \sum_{i \text{ wrong}} w_i^{(t)} \\
 &\geq \frac{3}{4} \Phi^t
 \end{aligned}$$

Therefore if expert i makes m_i mistakes and WM makes M mistakes, then

$$\begin{aligned}
 \left(\frac{1}{2}\right)^{m_i} &= w_i^{(t+1)} \leq \Phi^{t+1} \leq \Phi^1 \left(\frac{3}{4}\right)^M = N \left(\frac{3}{4}\right)^M \\
 \implies m_i \log_2 \left(\frac{1}{2}\right) &\leq \log_2 N + M \log_2 \frac{3}{4} \\
 \iff M &\leq \frac{m_i + \log_2 N}{\log_2 \frac{4}{3}} \leq 2.41(m_i + \log_2 N)
 \end{aligned}$$

□

Corollary 12.5. *By changing the re-weighting process to*

$$w_i^{(t+1)} = w_i^{(t)} \cdot \begin{cases} 1 & \text{if } i \text{ was correct} \\ 1 - \varepsilon & \text{if } i \text{ was incorrect} \end{cases}$$

the bound in Theorem 12.4 is

$$2(1 + \varepsilon)m_i + O\left(\frac{\log N}{\varepsilon}\right)$$

Remark 12.6. Any deterministic algorithm must have the 2 coefficient in front of the m_i . Consider $N = 2$ experts and suppose the algorithm happened to predict wrong every time. Then it has made at least twice the number of mistakes as the better expert.

3 Randomized Weighted Majority

The randomized weighted majority algorithm (RMW) proceeds as MW, except the prediction is probabilistic based on the current weights of the experts.

- As before, set $w_i^{(1)} = 1$ for all i .

- After time t , predict

$$\begin{cases} 0 & \text{with probability } \frac{\sum_{i:\mathcal{E}_i^t=0} w_i^{(t)}}{\sum_i w_i^{(t)}} \\ 1 & \text{otherwise} \end{cases}$$

- After seeing o^t , set $w_i^{(t+1)} \leftarrow w_i^{(t)} \cdot \begin{cases} 1 & \text{if } i \text{ was correct} \\ 1 - \varepsilon & \text{otherwise} \end{cases}$

Theorem 12.7. *Given any input sequence of \mathcal{E} and o , for any prefix of length T , and for all i , if RWM makes M mistakes and expert i makes m_i mistakes then*

$$\mathbf{E}[M] \leq (1 + \varepsilon)m_i + O\left(\frac{\log N}{\varepsilon}\right)$$

Remark 12.8. The $\varepsilon m_i + O\left(\frac{\log N}{\varepsilon}\right)$ gap from the best expert is called the *regret*.

Remark 12.9. When describing randomized algorithms, we must be careful in defining what adversaries can do.

Oblivious Adversary Plans entire sequence up front: The inputs $\mathcal{E}^1, o^1, \mathcal{E}^2, o^2, \dots$ are pre-determined.

Adaptive Adversary Sees our prediction *before* producing output: In order, it creates $\mathcal{E}^1 \rightarrow$ sees prediction $\rightarrow o^t \rightarrow \mathcal{E} \dots$

Semi-Adaptive Adversary Like the adaptive adversary, but our prediction happens in parallel with the actual outcome; neither depends on the other.

The adversaries are equivalent on deterministic algorithms, because it always outputs the same prediction and the oblivious adversary could have calculated that in advance when creating \mathcal{E}^{t+1} .

RWM works in the semi-adaptive model because predictions are not affected by the future. ¹

Proof of Theorem 12.7. Define the potential $\Phi^t = \sum_i w_i^{(t)}$ as before. Let

$$F^t = \frac{\sum_i \text{incorrect } w_i^{(t)}}{\sum_i w_i^{(t)}}$$

be the fraction of weight on incorrect experts at time t .

Because we predict proportionally to the weights of the experts, the probably that RWM makes a mistake at time t is F_t . Therefore

$$\mathbf{E}[M] = \sum_{t \in [T]} F_t$$

By our re-weighting rules,

$$\Phi^{t+1} = \Phi^t ((1 - F_t) + F_t(1 - \varepsilon)) = \Phi^t(1 - \varepsilon F_t)$$

¹In the literature, what we call the semi-adaptive adversary here is known as an adaptive adversary. There is no common name for what we call the adaptive adversary.

Bounding the size of the potential after T steps,

$$\begin{aligned}
(1 - \varepsilon)^{m_i} = w_i^{T+1} &\leq \Phi^{T+1} = \Phi^1 \prod_{t=1}^T (1 - \varepsilon F_t) \leq N e^{-\varepsilon \sum F_t} = N e^{-\varepsilon \mathbf{E}[M]} \\
\implies m_i \ln(1 - \varepsilon) &\leq \ln N - \varepsilon \mathbf{E}[M] \\
\implies \mathbf{E}[M] &\leq \frac{m_i \ln\left(\frac{1}{1-\varepsilon}\right) + \ln N}{\varepsilon}
\end{aligned}$$

where we used the inequality $1+x \leq e^x$ in the first line. Finally, note that $\ln\left(\frac{1}{1-\varepsilon}\right) \approx \varepsilon + \frac{\varepsilon^2}{2} \leq \varepsilon + \varepsilon^2$ for $\varepsilon \in [0, 1]$. So we get the bound

$$\mathbf{E}[M] \leq m_i(1 + \varepsilon) + \frac{\ln N}{\varepsilon}$$

□

4 Extending the “game”

We can generalize the notion of experts who are right or wrong to arbitrary gain/loss functions. At each time t , suppose the algorithm produces a vector (of probabilities) $\bar{p}^{(t)} = (p_1^t p_2^t \cdots p_N^t) \in \Delta_N$ (Δ_N is the probability simplex so $p_i \geq 0$ and $\sum_i p_i^t = 1$). In parallel, the adversary produces loss vector $\bar{\ell}^{(t)} = (\ell_1^t \cdots \ell_N^t) \in [-1, 1]^N$. Our loss or cost is $\langle \bar{p}^{(t)}, \bar{\ell}^{(t)} \rangle$.

Theorem 12.10. *Consider a fixed $\varepsilon \leq 1$. For all sequences of loss vectors, for all times T , and for all indices $i \in [N]$, there exists a deterministic algorithm such that*

$$\sum_{i=1}^T \langle \bar{p}^{(t)}, \bar{\ell}^{(t)} \rangle \leq \sum_{t=1}^T \bar{\ell}_i^{(t)} + \varepsilon T + \frac{\ln N}{\varepsilon}$$

Corollary 12.11. *For all $T \geq \frac{4 \log N}{\varepsilon^2}$, the average loss is bounded by*

$$\frac{1}{T} \sum_t \langle \bar{p}^{(t)}, \bar{\ell}^{(t)} \rangle \leq \frac{1}{T} \sum_t \bar{\ell}_i^{(t)} + \varepsilon$$

Remark 12.12. In the scenario with experts, the loss vector can be thought of as a vector of 0 and 1 representing whether they are wrong. When the loss vector is non-negative, we can use the multiplicative weights algorithm to get a slightly stronger bound.

Proof of Theorem 12.10. The Hedge algorithm [Freund, Schapire] proceeds as follows for a fixed ε .

As usual, consider weights $w_i^{(t)}$ initialized to $w_i^1 = 1$, and set $\Phi^t = \sum_i w_i^{(t)}$. Choose probabilities $p_i^{(t)} = \frac{w_i^t}{\Phi^t}$, so that $\sum_i p_i^{(t)} = 1$.

After each round, set $w_i^{(t+1)} \leftarrow w_i^{(t)} \cdot e^{-\varepsilon \ell_i^{(t)}}$.

Note that $\Phi^1 = N$, and

$$\begin{aligned}
\Phi^{t+1} &= \sum w_i^{(t+1)} = \sum_i w_i^{(t)} e^{-\varepsilon \bar{\ell}_i^{(t)}} \\
&\leq \sum_i w_i^{(t)} (1 - \varepsilon \bar{\ell}_i^{(t)} + \varepsilon^2 (\bar{\ell}_i^{(t)})^2) \quad (\text{using the inequality } e^x \leq 1 + x + x^2 \forall x \in [-1, 1]) \\
&\leq \sum_i w_i^{(t)} (1 + \varepsilon^2) - \varepsilon \sum_i w_i^{(t)} \bar{\ell}_i^{(t)} \quad (\text{because } |\bar{\ell}_i^{(t)}| \leq 1) \\
&= \Phi^t (1 + \varepsilon^2) - \varepsilon \Phi^t \langle \bar{p}^{(t)}, \bar{\ell}_i^{(t)} \rangle \\
&= \Phi^t \left(1 + \varepsilon^2 - \varepsilon \langle \bar{p}^{(t)}, \bar{\ell}_i^{(t)} \rangle \right) \\
&\leq \Phi^t e^{\varepsilon^2 - \varepsilon \langle \bar{p}^{(t)}, \bar{\ell}_i^{(t)} \rangle} \quad (\text{using } 1 + x \leq e^x)
\end{aligned}$$

So

$$\begin{aligned}
e^{-\varepsilon \sum \bar{\ell}_i^{(t)}} = w_i^{(t+1)} &\leq \Phi^{T+1} \leq \Phi^1 e^{\varepsilon^2 T - \varepsilon \sum \langle \bar{p}^{(t)}, \bar{\ell}_i^{(t)} \rangle} \\
\Rightarrow -\varepsilon \sum \bar{\ell}_i^{(t)} &\leq \ln N + \varepsilon^2 T - \varepsilon \sum_t \langle \bar{p}^{(t)}, \bar{\ell}_i^{(t)} \rangle \\
\Rightarrow \sum_t \langle \bar{p}^{(t)}, \bar{\ell}_i^{(t)} \rangle &\leq \sum \bar{\ell}_i^{(t)} + \varepsilon T + \frac{\ln N}{\varepsilon}
\end{aligned}$$

□