Multiplicative-Weight Update Method

Anil Maheshwari

anil@scs.carleton.ca
School of Computer Science
Carleton University
Canada

Outline

Problem

Expert Investor Without an Expertise

Model:

1. Access to n experts (newspapers, stock briefs, ...)
2. Predict whether TSX will go ↑ or ↓ at the end of each day.
3. Reward of 0/per day for correct prediction.
4. Costs us 1/per day for wrong prediction.

Problem: Devise an algorithm that makes prediction for each day. Suppose we are at day t, where $t \in \{1, \ldots, T\}$. Algorithm can use previous predictions + experts advises for days $1, \ldots, t-1$

Objective: At the end of T days we want to be competitive with respect to the best expert. Our cost (loss) is not significantly higher than the cost of any expert (including the best expert).

Warmup

Warmup: ≥ 1 Real Experts

Extra Knowledge: Among all the n experts there is at least one (unknown) expert who is always correct.

Let the set of experts be $E = \{1, \ldots, n\}$

For each day $t := 1$ to T do:

Step 1: Among all the remaining experts in E, poll them to find the prediction of the majority of them for that day. Record that as the prediction of the algorithm.

Step 2: Observe the true outcome at the end of the day. Discard all those experts that predicted wrong from E from future considerations.

#Wrong Predictions

Algorithm makes at most $O(\log n)$ wrong predictions

Extra Knowledge: The best expert (unknown) makes $\leq m$ mistakes over the period of T days.

Let the set of experts be $E = \{1, \ldots, n\}$
For each expert i, set its weight $w_i^1 = 1$

For each day $t := 1$ to T do:

    Step 1: Find the weighted majority prediction of the experts. Sum total the weights of all the experts that predict ↑ (respectively, ↓). Whichever of the two sums is higher is the prediction for day t.

    Step 2: Observe the true outcome at the end of the day t

    Step 3: For all the experts i that predicted wrongly, their weight is set to $w_i^{t+1} = w_i^t/2$. For all others, $w_i^{t+1} = w_i^t$.

Potential Function

Potential Function Φ^t

Define the potential function Φ^t for day $t \in \{1, \ldots, T\}$ to be the sum total of the weights of all the experts at the start of day t, i.e. $\Phi^t = \sum\limits_{i=1}^{n} w_i^t$

Observations:

1. $\Phi^1 = n$

2. If the algorithm makes a wrong prediction on day t, $\Phi^{t+1} \leq \frac{3}{4}\Phi^t$

3. If the algorithm has made M mistakes in T days, its total weight at the end of day T, $\Phi^{T+1} \leq \left(\frac{3}{4}\right)^M \Phi^1 = \left(\frac{3}{4}\right)^M n$

Assume that the best Expert is i (we don't know it's identity).

Bounds on Potential Function

Weight of the expert i at the end of day T is $\geq (\frac{1}{2})^m$. Therefore,
$\left(\frac{1}{2}\right)^m \leq \Phi^{T+1} \leq \left(\frac{3}{4}\right)^M n$

Bounding number of mistakes

The algorithms makes at most $2.41(m + \log n)$ mistakes.

Proof: We had $\left(\frac{1}{2}\right)^m \leq \Phi^{T+1} \leq \left(\frac{3}{4}\right)^M n$. Take $\log$'s:

$$
\begin{aligned}
-m &\leq M \log\left(\frac{3}{4}\right) + \log n \\
-M \log\left(\frac{3}{4}\right) &\leq m + \log n \\
M \log\left(\frac{4}{3}\right) &\leq m + \log n \\
M &\leq 2.41(m + \log n)
\end{aligned}
$$

$\square$

Replacing $\frac{1}{2}$

Choose $\eta \in (0, \frac{1}{2}]$

Let the set of experts be $E = \{1, \ldots, n\}$
For each expert i, set its starting weight $w_i^1 = 1$

For each day $t := 1$ to T do:

  Step 1: Find the weighted majority prediction of the experts. Sum total the weights of all the experts that predict $\uparrow$ (respectively, $\downarrow$). Whichever of the two sums is higher is the prediction for day t.

  Step 2: Observe the true outcome at the end of the day t

  Step 3: For all the experts i that predicted wrongly, their weight is set to $w_i^{t+1} = (1 - \eta)w_i^t$. For all others, $w_i^{t+1} = w_i^t$.

Analysis

New Bound

For any $\eta \in (0, \frac{1}{2}]$, we have $M \leq 2(1+\eta)m + \frac{2}{\eta}\log n$

Proof: Potential function equation: $(1-\eta)^m \leq \Phi^{T+1} \leq (1-\frac{\eta}{2})^M n$

Take $\log$'s and use $-\eta - \eta^2 \leq \ln(1-\eta) \leq -\eta$ for $\eta \in [0, \frac{1}{2}]$:

$$
\begin{aligned}
m \ln(1-\eta) &\leq M \ln(1-\frac{\eta}{2}) + \ln n \\
-m(\eta + \eta^2) &\leq -M\frac{\eta}{2} + \ln n \\
M\frac{\eta}{2} &\leq \ln n + (\eta + \eta^2)m \\
M &\leq \frac{2}{\eta}\ln n + 2(1+\eta)m
\end{aligned}
$$

$\square$

Randomization

An Example

For any $\eta \in (0, \frac{1}{2}]$, we have $M \leq 2(1 + \eta)m + \frac{2}{\eta} \log n$

Example: Two experts A and B.
A predicts $(\uparrow, \downarrow)$ correctly only on even numbered days
B predicts $(\uparrow, \downarrow)$ correctly only on odd numbered days

Best Expert: Wrong half the times.

How many times the Algorithm predicts wrongly?

How to remove the multiplicative factor of 2 in $M \leq \mathbf{2}(1 + \eta)m + \frac{2}{\eta} \log n$?

Use Randomization

Loss/Cost: Assume that the loss (costs) are real numbers in the interval $[0, 1]$.

Let $m_i^t \in [0, 1]$ denote the loss of expert $i \in \{1, \ldots, n\}$ on day $t \in \{1, \ldots, T\}$

Objective: Algorithm to be competitive against the cost incurred by the best expert. Assume i is the (unknown) best expert.

$M^t =$ Expected cost that the algorithm incurs on day t.

Our algorithm should have the following guarantee:

$$\sum_{i=1}^{T} M^t \leq \frac{\ln n}{\eta} + (1 + \eta) \sum_{i=1}^{T} m_i^t$$

Total expected cost over T days is within an additive factor $\frac{\ln n}{\eta}$ and a multiplicative factor $(1 + \eta)$ of the best expert.

Multiplicative Weight Update Method

Set of experts $E = \{1, \ldots, n\}$.

Let η be any real number in $[0, \frac{1}{2}]$

For each expert i, set its initial weight $w_i^1 = 1$

For each day $t := 1$ to T do:

> **Step 1:** Define $\Phi^t = \sum\limits_{i=1}^{n} w_i^t$
>
> For each expert i, compute $p_i^t = \frac{w_i^t}{\Phi^t}$
>
> **Step 2:** Choose an expert based on their probabilities and predict
> ($\uparrow, \downarrow$) according to the chosen expert
>
> **Step 3:** Update Weights: For each expert i set $w_i^{t+1} = w_i^t(1 - \eta m_i^t)$

Expected loss M^t

The expected loss M^t that the algorithm incurs on day t is given by

$$M^t = \sum_{i=1}^{n} p_i^t m_i^t = \langle p^t \cdot m^t \rangle, \tag{1}$$

where $p^t = (p_1^t, p_2^t, \ldots, p_n^t)$ and $m^t = (m_1^t, m_2^t, \ldots, m_n^t)$ and their dot product is $\langle p^t \cdot m^t \rangle$

Potential Function

For any $t \in \{1, \ldots, T\}$: $\Phi^{t+1} \leq \Phi^t e^{-\eta M^t}$

Moreover, $\Phi^{T+1} \leq n e^{-\eta \sum\limits_{t=1}^{T} M^t}$

Proof:

$$
\begin{aligned}
\Phi^{t+1} &= \sum_{i=1}^{n} w_i^{t+1} \\
&= \sum_{i=1}^{n} w_i^t (1 - \eta m_i^t) \\
&= \sum_{i=1}^{n} w_i^t - \eta \sum_{i=1}^{n} w_i^t m_i^t \\
&= \Phi^t - \eta \sum_{i=1}^{n} \Phi^t p_i^t m_i^t
\end{aligned}
$$

$$
\begin{aligned}
\Phi^{t+1} &= \Phi^t - \eta \Phi^t \sum_{i=1}^{n} p_i^t m_i^t \\
&= \Phi^t (1 - \eta \langle p^t \cdot m^t \rangle) \\
&\leq \Phi^t e^{-\eta \langle p^t \cdot m^t \rangle} \\
&= \Phi^t e^{-\eta M^t}
\end{aligned}
$$

Using induction on t, we obtain

$$
\begin{aligned}
\Phi^{T+1} &\leq \Phi^0 e^{-\eta \sum_{t=1}^{T} M^t} \\
&= n e^{-\eta \sum_{t=1}^{T} M^t}
\end{aligned}
$$

$\square$

Lower Bound

For any expert i, $\Phi^{T+1} \geq (1-\eta)^{\sum\limits_{t=1}^{T} m_i^t}$

Proof: Since all $m_i^t \in [0,1]$, $w_i^t \geq 0$.

$\implies \Phi^{t+1} \geq w_i^{t+1}$ for any individual weight as $\Phi^{t+1} = \sum\limits_{i=1}^{n} w_i^{t+1}$

Note: For $\epsilon \in [0,1]$, $1 - \epsilon x \geq (1-\epsilon)^x$ if $x \in [0,1]$

Using the update rule of w_i^{t+1}, we have

$\Phi^{T+1} \geq w_i^{T+1} = w_i^1 \prod\limits_{t=1}^{T} (1 - \eta m_i^t) \geq (1-\eta)^{\sum\limits_{t=1}^{T} m_i^t}$

$\square$

19

Putting Lower & Upper Bounds Together

Putting upper and lower bounds for Φ^{T+1} we have

$$ne^{-\eta \sum\limits_{t=1}^{T} M^t} \geq \Phi^{T+1} \geq (1-\eta)^{\sum\limits_{t=1}^{T} m_i^t} \tag{2}$$

Take $\log$'s and divide by η:

$$\frac{\ln n}{\eta} - \sum_{t=1}^{T} M^t \geq \frac{\ln(1-\eta)}{\eta} \sum_{t=1}^{T} m_i^t \tag{3}$$

This is equivalent to

$$\sum_{t=1}^{T} M^t \leq \frac{\ln n}{\eta} - \frac{\ln(1-\eta)}{\eta} \sum_{t=1}^{T} m_i^t \tag{4}$$

Putting Lower & Upper Bounds Together (contd.)

Use $\eta \in [0, \frac{1}{2}]$, $-\eta - \eta^2 \leq \ln(1 - \eta)$ and we obtain:

$$\sum_{t=1}^{T} M^t \leq \frac{\ln n}{\eta} + \frac{\eta + \eta^2}{\eta} \sum_{t=1}^{T} m_i^t \tag{5}$$

or

$$\sum_{t=1}^{T} M^t \leq \frac{\ln n}{\eta} + (1 + \eta) \sum_{t=1}^{T} m_i^t$$

± 1 **Costs**

The costs of each expert can be positive or negative, i.e. $m_i^t \in [-1, 1]$. Use the same algorithm as for the cost $[0, 1]$:

MWU with costs in $[-1, 1]$

Set of experts $E = \{1, \ldots, n\}$.
Let η be any real number in $[0, \frac{1}{2}]$
For each expert i, set its initial weight $w_i^1 = 1$

For each day $t := 1$ to T do:

    Step 1: Define $\Phi^t = \sum\limits_{i=1}^{n} w_i^t$

             For each expert i, compute $p_i^t = \frac{w_i^t}{\Phi^t}$

    Step 2: Choose an expert based on their probabilities and predict $(\uparrow, \downarrow)$ according to the chosen expert

    Step 3: Update Weights: For each expert i set $w_i^{t+1} = w_i^t(1 - \eta m_i^t)$

Upper Bound

$$\Phi^{T+1} \leq n e^{-\eta \sum_{t=1}^{T} M^t}$$

Proof: Expected cost of algorithm on day t: $M^t = \sum_{i=1}^{n} p_i^t m_i^t = \langle p^t \cdot m^t \rangle$

The upper bound for Φ^{t+1} follows the same analysis:
$$\Phi^{t+1} = \sum_{i=1}^{n} w_i^{t+1} = \sum_{i=1}^{n} w_i^t (1 - \eta m_i^t) \leq \Phi^t e^{-\eta M^t}$$

And using induction on t we have $\Phi^{T+1} \leq n e^{-\eta \sum_{t=1}^{T} M^t}$ $\square$

Lower Bound

$$\Phi^{T+1} \geq (1-\eta)^{\sum\limits_{m_i^t \geq 0} m_i^t} (1+\eta)^{-\sum\limits_{m_i^t < 0} m_i^t}$$

Proof:

Note: For $\epsilon \in [0,1]$, $(1-\epsilon)^x \leq 1 - \epsilon x$ if $x \in [0,1]$ and $(1+\epsilon)^{-x} \leq 1 - \epsilon x$ if $x \in [-1, 0]$

Since $m_i^t \in [-1, 1]$, we have $1 - \eta m_i^t \geq 0$

Thus, $w_i^t \geq 0 \implies \Phi^{t+1} \geq w_i^{t+1}$ for any individual weight.

From the update rule of w_i^{t+1}:

$$\Phi^{T+1} \geq w_i^{T+1} = \prod_{t=1}^{T} (1 - \eta m_i^t).$$

Group for each day the positive m_i^t's and the negative m_i^t's:

$$\Phi^{T+1} \geq (1-\eta)^{\sum\limits_{m_i^t \geq 0} m_i^t} (1+\eta)^{-\sum\limits_{m_i^t < 0} m_i^t}$$

$\square$

$$ne^{-\eta \sum_{t=1}^{T} M^t} \geq (1-\eta)^{\sum_{m_i^t \geq 0} m_i^t} (1+\eta)^{-\sum_{m_i^t < 0} m_i^t}$$

Take $\log$'s and divide by η:

$$\frac{\ln n}{\eta} - \sum_{t=1}^{T} M^t \geq \frac{\ln(1-\eta)}{\eta} \sum_{m_i^t \geq 0} m_i^t - \frac{\ln(1+\eta)}{\eta} \sum_{m_i^t < 0} m_i^t$$

$$\sum_{t=1}^{T} M^t \leq \frac{\ln n}{\eta} - \frac{\ln(1-\eta)}{\eta} \sum_{m_i^t \geq 0} m_i^t + \frac{\ln(1+\eta)}{\eta} \sum_{m_i^t < 0} m_i^t$$

Upper+Lower Bound (contd.)

Use, $\eta + \eta^2 \geq -\ln(1 - \eta)$ and $\ln(1 + \eta) \geq \eta - \eta^2$, for $\eta \in [0, \frac{1}{2}]$.

$$\sum_{t=1}^{T} M^t \leq \frac{\ln n}{\eta} + (1 + \eta) \sum_{m_i^t \geq 0} m_i^t + (1 - \eta) \sum_{m_i^t < 0} m_i^t$$

Note: $(\eta - \eta^2) \sum_{m_i^t < 0} m_i^t \geq \ln(1 + \eta) \sum_{m_i^t < 0} m_i^t$ because of negative values. On expanding we obtain

$$\sum_{t=1}^{T} M^t \leq \frac{\ln n}{\eta} + \eta \sum_{t=1}^{T} |m_i^t| + \sum_{t=1}^{T} m_i^t$$

Since $|m_i^t| \leq 1$, we have

$$\sum_{t=1}^{T} M^t \leq \frac{\ln n}{\eta} + \eta T + \sum_{t=1}^{T} m_i^t \tag{6}$$

26

Cost of MWU

By setting $\eta = \sqrt{\frac{\ln n}{T}}$ in $\sum\limits_{t=1}^{T} M^t \leq \frac{\ln n}{\eta} + \eta T + \sum\limits_{t=1}^{T} m_i^t$, we obtain

$\sum\limits_{t=1}^{T} M^t \leq 2\sqrt{T \ln n} + \sum\limits_{t=1}^{T} m_i^t$

Interpretation: Cost of MWU algorithm is off by an additive factor that is proportional to the square root of the product of the number of days and the number of experts as compared to the best expert.

Average Error: Consider the average error on each day (divide by T):

$$\frac{1}{T} \sum\limits_{t=1}^{T} M^t \leq 2\sqrt{\frac{\ln n}{T}} + \frac{1}{T} \sum\limits_{t=1}^{T} m_i^t$$

Observe that as T increases the average error drops down. Therefore, MWU method is able to learn from experts reasonably well when executed over a number of days.

Generalization

The costs of each expert is in $m_i^t \in [-l, \rho]$, where $l < \rho$.

MWU with costs in $[-l, \rho]$

Set of experts $E = \{1, \ldots, n\}$.

Let η be any real number in $[0, \frac{1}{2}]$

For each expert i, set its weight $w_i^1 = 1$

For each day $t := 1$ to T do:

Step 1: Define $\Phi^t = \sum\limits_{i=1}^{n} w_i^t$. For each expert i, compute $p_i^t = \frac{w_i^t}{\Phi^t}$

Step 2: Choose an expert based on their probabilities and predict according to the chosen expert

Step 3: Update Weights: For each expert i set

$$w_i^{t+1} = \begin{cases} w_i^t(1 - \eta)^{\frac{m_i^t}{\rho}}, & \text{if } m_i^t \geq 0 \\ w_i^t(1 + \eta)^{-\frac{m_i^t}{\rho}}, & \text{if } m_i^t < 0 \end{cases}$$

As $m_i^t/\rho \in [-1, +1]$, we have

Upper Bound

$$\Phi^{T+1} \leq n e^{-\eta \sum\limits_{t=1}^{T} \frac{M^t}{\rho}}$$

Lower Bound

$$\Phi^{T+1} \geq (1 - \eta)^{\sum\limits_{m_i^t \geq 0} \frac{m_i^t}{\rho}} (1 + \eta)^{-\sum\limits_{m_i^t < 0} \frac{m_i^t}{\rho}}$$

Results

Theorem

$$\sum_{t=1}^{T} M^t \leq \frac{\rho \ln n}{\eta} + (1+\eta) \sum_{m_i^t \geq 0} m_i^t + (1-\eta) \sum_{m_i^t < 0} m_i^t$$

Corollary

(A) If $\eta \leq \min(\frac{1}{2}, \frac{\epsilon}{4l})$ for some error parameter ϵ, after $T = \frac{2\rho \ln n}{\eta \epsilon}$ time, the average expected loss is given by

$$\frac{1}{T} \sum_{t=1}^{T} M^t \leq \epsilon + \frac{(1+\eta)}{T} \sum_{m_i^t \geq 0} m_i^t + \frac{(1-\eta)}{T} \sum_{m_i^t < 0} m_i^t$$

(B) If $\eta \leq \min(\frac{1}{2}, \frac{\epsilon}{4\rho})$ for some error parameter ϵ, after $T = \frac{16\rho^2 \ln n}{\epsilon^2}$ time, the average expected loss is given by

$$\frac{1}{T} \sum_{t=1}^{T} M^t \leq \epsilon + \frac{1}{T} \sum_{t} m_i^t$$

- Arora, Hazan and Kale, The multiplicative weights update method: a meta-algorithm and applications, Theory of Computing 8(1): 121-164, 2012.

- Several Lecture Notes