CSCI-B609: A Theorist's Toolkit, Fall 2016 Dec 01

Lecture 24: The Multiplicative Weights Algorithm

Lecturer: Yuan Zhou *Scribe: Kaiyuan Zhu*

1 Introduction

In the next two lectures, we will present an alternative algorithm for solving LPs and SDPs. *Multiplicative weights* is a term for the simple iterative rule underlying these algorithms; it is known by different names in the various fields where it was (re)discovered. The detailed content can be found in a survey by Arora, Hazan and Kale [AHK05]; and our discussion is mainly based on their treatment.

Due to its broad appeal, we first consider multiplicative weights in more generality than solving LPs and SDPs, in the form of playing a prediction game. Then, we'll investigate the tweaked game with a strategy called Hedge.

2 Warmup: Prediction with Expert Advice

The following sequential game is played between an omniscient adversary and an aggregator who is advised by N experts. In total, there are T sequential predictions to be made. (Special cases of this game include predicting if it will rain tomorrow, or if the stock market will rise or decline.)

For $t = 1, 2, \cdots, T$

- 1. Each expert $i \in [N]$ makes his/her own prediction (Yes/No)
- 2. The aggregator predicts either Yes or No based on expert advices
- 3. The adversary decides the real Yes/No outcome
- 4. Aggregator observes the outcome and suffers if the prediction was incorrect.

Naturally, the aggregator wants to make as few mistakes as possible. If there is a perfect expert who always makes the correct prediction (and the other $N - 1$ experts may make

arbitrarily many incorrect predictions), then the following simple strategy can be used to dismiss the experts who are not perfect.

Aggregator Strategy: Start with a set S of active experts, say $|S| = N$, for each $t \in [T]$

- Decide Yes/No using majority vote among expert votes in S;
- If wrong decision made by aggregator, dismiss the experts in S who made a wrong decision.

Fact. Each time the aggregator makes a wrong decision, at least $\frac{|S|}{2}$ experts are dismissed.

Claim 1. *The aggregator suffers at most* $\lceil \log_2 N \rceil$ *times.*

Now consider the case that the best expert even makes at most M mistakes. The similar strategy of majority vote and dismiss an expert who has made $M+1$ mistakes can lead to the aggregator making at most $(M+1)\lceil \log_2 N \rceil$ mistakes. This bound is rather poor since it depends multiplicatively on M.

Improved Strategy (Weighted Majority): We may obtain an additive mistake bound by softening the penalties, i.e., instead of dismissing him/her after an expert made enough mistakes, discount that expert's advice [LW89]. Start with a weight vector $w_i^{(1)}$ that corresponds to each expert $i \in [N]$.

For each $t = 1, 2, \cdots, T$

- Predict Yes/No based on a weighted majority vote per $\mathbf{w}^{(t)} = (w_t^{(t)}, \cdots, w_N^{(t)})$
- After observing the outcome, for each expert i who makes a mistake, set $w_i^{(t+1)} = \frac{1}{2} w_i^{(t)}$; for other experts i set $w_i^{(t+1)} = w_i^{(t)}$

Theorem 1. *Let* m_j *be the number of mistakes made by expert* j*, then for any sequence of outcomes and predictions by experts*

$$\#\ \textit{Aggregator errors}\ \leq (\log_{\frac{4}{3}} 2) \cdot m_j + \log_{\frac{4}{3}} N$$

Proof. Let $\Phi^{(t)} = \sum_{i=1}^{N} w_i^{(t)}$ be a potential function, then we have: a) $\Phi^{(1)} = N$; b) $\Phi^{(T+1)} \geq w_j^{(T+1)} \geq 2^{-m_j}$ and c) At any t when the aggregator makes a mistake, at least half of the weights in $\Phi^{(t)}$ get halved. In other words, $\Phi^{(t+1)} \leq \frac{3}{4}\Phi^{(t)}$. Therefore, let E denote the number of aggregator errors, $2^{-m_j} \leq \Phi^{(t+1)} \leq \left(\frac{3}{4}\right)^E \cdot \Phi^{(1)}$. Hence $E \leq \log_{\frac{4}{3}}(N \cdot 2^{m_j}) = (\log_{\frac{4}{3}} 2) \cdot m_j + \log_{\frac{4}{3}} N$. □

Corollary 2. *If instead of halving the weights of mistaking experts, we let* $w_i^{(t+1)} = \dfrac{w_i^{(t)}}{1+\epsilon}$, *then we can achieve # Aggregator errors* $\leq 2(1+\epsilon)\cdot m_j + O\left(\dfrac{\log N}{\epsilon}\right)$.

Proof. Following a similar analysis, when the aggregator errors at time t, $\Phi^{(t+1)} \leq \left(\frac{1}{2} + \frac{1}{2(1+\epsilon)}\right)\Phi^{(t)}$, and the final $\Phi^{(T+1)} \geq \left(\frac{1}{1+\epsilon}\right)^{m_j}$. Therefore $\left(\frac{1}{1+\epsilon}\right)^{m_j} \leq \left(\frac{2+\epsilon}{2+2\epsilon}\right)^{E}\cdot\Phi^{(1)}$, where E is the number of aggregator errors again, and we can obtain $E \leq \log_{1+\frac{\epsilon}{2+\epsilon}}[N\cdot(1+\epsilon)^{m_j}] \leq 2(1+\epsilon)\cdot m_j + O\left(\frac{\log N}{\epsilon}\right)$. □

3 The Hedge Strategy

Now, we modify the game with a view to solving LPs and SDPs. Specifically, the new game's rule introduces weighting and costs:

For each $t = 1, 2, \cdots, T$

- 1. Allocator picks some distribution $\mathbf{p}^{(t)} = (p_1^{(t)}, \cdots, p_N^{(t)})$ over experts.
- 2. Adversary, knowing all expert advices and $\mathbf{p}^{(t)}$, determines a cost vector $\mathbf{c}^{(t)} = (c_1^{(t)}, \cdots, c_N^{(t)}) \in [-1, 1]^N$.
- 3. Allocator observes $\mathbf{c}^{(t)}$ and suffers $\mathbf{p}^{(t)}\cdot\mathbf{c}^{(t)}$.

Remark. An intuitive analogy is, each expert corresponds to some investment, and at each day t, the allocator has to distribute his money to all N investments. Negative cost can be interpreted as benefit.

The new game is played with the Hedge strategy. Its exponential update rule distinguishes it from Weighted Majority above: assign each expert i a weight $w_i^{(1)}$ initialized to 1.

The Hedge Strategy [FS97] At each time t,

- Pick the distribution $p^{(t)} = w_i^{(t)}/\Phi^{(t)}$, where $\Phi^{(t)} = \sum_i w_i^{(t)}$
- After observing the cost vector, set $w_i^{(t+1)} = w_i^{(t)}\cdot\exp(-\epsilon c_i^{(t)})$, $\forall i \in [N]$.

Theorem 3. *For $\epsilon < 1$ and every $i \in [N]$, the Hedge strategy guarantees*

$$\sum_{t=1}^{T} \mathbf{p}^{(t)} \cdot \mathbf{c}^{(t)} \leq \sum_{t=1}^{T} c_i^{(t)} + \frac{\ln N}{\epsilon} + \epsilon T$$

Remark. In the analogy of investments, the above theorem may be interpreted as "the total expected cost of the Hedge strategy is not much worse than the total cost of any individual investment". The proof of Theorem 3 is given as follows.

Proof. Consider again the potential function Φ, we still have a)$\Phi^{(1)} = N$ and b)$\Phi^{(T+1)} \geq w_i^{(T+1)} \exp(-\epsilon \cdot \sum_t c_i^{(t)})$. For each t,

$$\Phi^{(t+1)} = \sum_{j=1}^{N} w_j^{(t+1)} = \sum_j w_j^{(t)} \cdot \exp(-\epsilon c_j^{(t)}) \quad (*)$$

For $x \in [-1, 1]$ we have $\exp(x) = 1 + x + \frac{x^2}{2} + O(x^3) \leq 1 + x + x^2$. Therefore,

$$\begin{aligned}
(*) &\leq \sum_j w_j^{(t)}(1 - \epsilon c_j^{(t)} + \epsilon^2 c_j^{(t)2}) \\
&\leq \sum_j w_j^{(t)}(1 + \epsilon^2 - \epsilon c_j^{(t)}) \\
&= (1 + \epsilon^2)\Phi^{(t)} - \epsilon \cdot \sum_j (p_j^{(t)} \cdot c_j^{(t)})\Phi^{(t)} \\
&= \Phi^{(t)}(1 + \epsilon^2 - \epsilon \mathbf{p}^{(t)} \cdot \mathbf{c}^{(t)}) \\
&\leq \Phi^{(t)} \cdot \exp(-\epsilon \mathbf{p}^{(t)} \mathbf{c}^{(t)} + \epsilon^2)
\end{aligned}$$

which gives $\exp\left(-\epsilon \cdot \sum_{t=1}^{T} c_i^{(t)}\right) \leq \Phi^{(T+1)} \leq \Phi^{(1)} \cdot \exp\left(\epsilon T - \epsilon \sum_{t=1}^{T} \mathbf{p}^{(t)} \cdot \mathbf{c}^{(t)}\right)$, i.e. $-\epsilon \cdot \sum_{t=1}^{T} c_i^{(t)} \leq \ln N + \epsilon^2 T - \epsilon \cdot \sum_{t=1}^{T} \mathbf{p}^{(t)} \mathbf{c}^{(t)}$ and the final inequality follows after rearranging. □

Corollary 4. *(Similar result also holds for average cost.) If $T > \frac{\ln N}{\epsilon^2}$, we have (of course, for $\epsilon < 1$ and all $i \in [N]$)*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbf{p}^{(t)} \mathbf{c}^{(t)} \leq \frac{1}{T} \sum_{t=1}^{T} c_i^{(t)} + 2\epsilon$$

Furthermore, we can extend the result to cost vectors in $[-\rho, \rho]^N$.

Corollary 5. *If the cost vectors* $\mathbf{c}^{(t)} \in [-\rho, \rho]^N$ *and* $T > \dfrac{\rho^2 \ln N}{\epsilon^2}$*, then for all* $i \in [N]$ *and* $\epsilon < \rho$*, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{p}^{(t)}\mathbf{c}^{(t)} \leq \frac{1}{T}\sum_{t=1}^{T} c_i^{(t)} + 2\epsilon$$

Reference

[AHK05] https://www.cs.princeton.edu/~arora/pubs/MWsurvey.pdf

[LW89] http://homepages.math.uic.edu/~lreyzin/f16_mcs548/littlestone94.pdf

[FS97] http://www.face-rec.org/algorithms/Boosting-Ensemble/decision-theoretic_generalization.pdf

*Acknowledgement: This lecture https://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15859-f11/www/notes/lecture16.pdf is going to be useful.