Online Learning with Experts & Multiplicative Weights Algorithms

CS 159 lecture #2

Stephan Zheng

April 1, 2016

Caltech

Table of contents

Today

What should you take away from this lecture?

- How should you base your prediction on expert predictions?
- What are the characteristics of multiplicative weights algorithms?

References

- Online Learning [Ch 2-4], *Gabor Bartok, David Pal, Csaba Szepesvari, and Istvan Szita.*
- The Multiplicative Weights Update Method: A Meta-Algorithm and Applications, *Sanjeev Arora, Elad Hazan, and Satyen Kale. Theory of Computing, 8, 121-164, 2012.*
- `http://www.yisongyue.com/courses/cs159/`

Online Learning with Experts

Leo at the Oscars: an online binary prediction problem

Will Leo win an Oscar this year? (running question since $\sim$1997...)

...	No	
2005	No	
2006	No	
2007	No	
2008	No	
2009	No	
2010	No	

Leo at the Oscars: an online binary prediction problem

2011	No	
2012	No	
2013	No	
2014	No	
2015	No	
2016	Yes!	

Leo at the Oscars: an online binary prediction problem

A frivolous extrapolation:

2016	Yes!	
2017	No	
2018	Yes!	
2019	No	
2020	Yes	
2021	Yes	

Leo at the Oscars: an online binary prediction problem

Imagine you're a showbiz fan and want to **predict** the answer every year t.

But, you don't know all the ins and outs of Hollywood.

Instead, you are lucky enough to have access to experts $\{i\}_{i \in \mathcal{I}}$, that each make a (possibly wrong!) prediction $p_{i,t}$.

- How do you use the experts for your own prediction?
- How do you incorporate the feedback from Nature?
- How do we achieve sublinear regret?

Formal description

Formal setup: there are T rounds in which we predict a binary label $\hat{p}_t \in \mathcal{Y} = \{0, 1\}$.

At every timestep t:

1. Each expert $i = 1 \dots n$ predicts $p_{i,t} \in \mathcal{Y}$
2. You make a prediction $\hat{p}_t \in \mathcal{Y}$
3. Nature reveals $y_t \in \mathcal{Y}$
4. We suffer a loss $l(\hat{p}_t, y_t) = \mathbf{1}[\hat{p}_t \neq y_t]$
5. Each expert suffers a loss $l(p_{i,t}, y_t) = \mathbf{1}[p_{i,t} \neq y_t]$

Loss: $\hat{L}_T = \sum_{t=1}^{T} l(\hat{p}_t, y_t), L_{i,T} = \sum_{t=1}^{T} l(p_{i,t}, y_t)$ *# of mistakes made*

Regret: $R_T = \hat{L}_T - \min_i L_{i,T}$

How do you decide what $\hat{p}_t$ to predict? How do you incorporate the feedback from Nature? How do we achieve sublinear regret?

- ”Experts” is a way to abstract a hypothesis class.
- For the most part, we’ll deal with a finite, discrete number of experts, because that’s easier to analyze.
- In general, there can be a *continuous* space of experts = using a standard hypothesis class.

- Boosting uses a collection of n weak classifiers as ”experts”. At every time t it adds **1** weak classifier with weight α_t to the ensemble classifier $h_{1:t}$. The prediction $\hat{p}_t$ is based on the ensemble $h_{1:t}$ that we have collected so far.

Weighted Majority: the HALVING algorithm

Let's assume that there is a perfect expert i^*, that is guaranteed to know the right answer (say a mind-reader that can read the minds of the voting Academy members). That is, $\forall t : p_{i^*,t} = y_t$ and $l_{i^*,t} = 0$.

Keep regret (= # mistakes) small $\longrightarrow$ find the perfect expert quickly with few mistakes.

- Eliminate experts that make a mistake
- Take a majority vote of "alive" experts

Let's define some groups of experts:

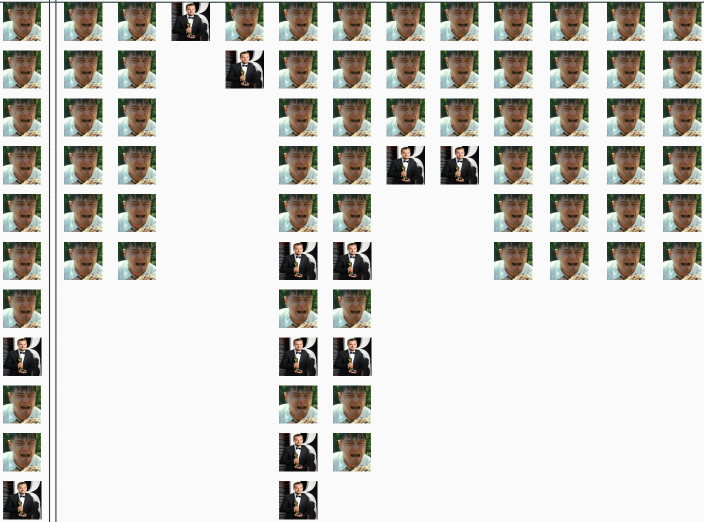
The "alive" set:
$E_t = \{i : i \text{ did not make a mistake until time } t\}, E_0 = \{1, \ldots, n\}$

The nay-sayers: $E^0_{t-1} = \{i \in E_{t-1} : p_{i,t} = 0\}$

The yay-sayers: $E^1_{t-1} = \{i \in E_{t-1} : p_{i,t} = 1\}$

Weighted Majority: the HALVING algorithm

t	$\hat{l}_t$	y_t	$\hat{p}_t$	$p_{1,t}$	$p_{2,t}$	$p_{3,t}$	$p_{4,t}$	$p_{5,t}$	$p_{6,t}$	$p_{7,t}$	$p_{8,t}$	$p_{9,t}$	$p_{10,t}$	$p_{11,t}$	$p_{12,t}$
2011	0	No													
2012	0	No													
2013	0	No													
2014	0	No													
2015	0	No													
2016	1	Yes!													
2017	0	No													
2018	0	Yes!													
2019	0	No													
2020	1	Yes													
2021	0	Yes													



The HALVING algorithm

for $t = 1 \ldots T$ **do**
    Receive expert predictions $(p_{1,t} \ldots p_{n,t})$
    Split E_{t-1} into $E^0_{t-1} \cup E^1_{t-1}$
    if $|E^1_{t-1}| > |E^0_{t-1}|$ **then** ▷ follow the majority
        Predict $\hat{p}_t = 1$
    else
        Predict $\hat{p}_t = 0$
    end if
    Receive Nature's answer y_t (and incur loss $l(\hat{p}_t, y_t)$)
    Update E_t with experts that continue to be right
end for

Weighted Majority: the HALVING algorithm

Notice that if HALVING makes a mistake, then at least half of the experts in E_{t-1} were wrong:

$$W_t = |E_t| = |E_{t-1}^{y_t}| \leq |E_{t-1}|/2.$$

Since there is always a perfect expert, the algorithm makes no more than $\lfloor \log_2 |E_0| \rfloor = \lfloor \log_2 n \rfloor$ mistakes $\longrightarrow$ sublinear regret.

Qualitatively speaking:

- W_t multiplicatively decreases when HALVING makes a mistake. *If W_t doesn't shrink too much, then HALVING can't make too many mistakes.*
- There is a lower bound on W_t for all t, since there is an expert. *W_t can't shrink "too much" starting from its initial value.*

We'll see similar behavior later.

So what should we do if we don't know anything about the experts? The majority vote might be always wrong!

Give each of them a weight $w_{i,t}$, initialized as $w_{i,1} = 1$.

Decide based on a weighted sum of experts:

$$\hat{p}_t = \mathbf{1}\left[\text{weighted sum of yay-sayers} > \text{weighted sum of nay-sayers}\right]$$

$$\hat{p}_t = \mathbf{1}\left[\sum_{i=1}^{n} w_{i,t-1} p_{i,t} > \sum_{i=1}^{n} w_{i,t-1}(1 - p_{i,t})\right]$$

Weighted Majority

Recall: with a perfect expert we had

The HALVING algorithm

for $t = 1 \dots T$ **do**
  Receive expert predictions $(p_{1,t} \dots p_{n,t})$
  Split E_{t-1} into $E^0_{t-1} \cup E^1_{t-1}$
  if $|E^1_{t-1}| > |E^0_{t-1}|$ **then** ▷ follow the majority
    Predict $\hat{p}_t = 1$
  else
    Predict $\hat{p}_t = 0$
  end if
  Receive Nature's answer y_t (and incur loss $l(\hat{p}_t, y_t)$)
  Update E_t with experts that continue to be right
end for

Without knowledge about the experts: choose a *decay factor* $\beta \in [0, 1)$

The Weighted Majority algorithm

```
Initialize w_{i,1} = 1, W_1 = n.
for t = 1 ... T do
    Receive expert predictions (p_{1,t} ... p_{n,t})
    if sum_{i=1}^n w_{i,t-1} p_{i,t} > sum_{i=1}^n w_{i,t-1}(1 - p_{i,t}) then
        Predict p̂_t = 1
    else
        Predict p̂_t = 0
    end if
    Receive Nature's answer y_t (and incur loss l(p̂_t, y_t))
    w_{i,t} <- w_{i,t-1} β^{1(p_{i,t} ≠ y_t)}
end for
```

1. Initialize $w_{i,1} = 1, W_1 = n$.
2. **for** $t = 1 \dots T$ **do**
    - Receive expert predictions $(p_{1,t} \dots p_{n,t})$
    - **if** $\sum_{i=1}^{n} w_{i,t-1} p_{i,t} > \sum_{i=1}^{n} w_{i,t-1}(1 - p_{i,t})$ **then**
        - Predict $\hat{p}_t = 1$
    - **else**
        - Predict $\hat{p}_t = 0$
    - **end if**
    - Receive Nature's answer y_t (and incur loss $l(\hat{p}_t, y_t)$)
    - $w_{i,t} \leftarrow w_{i,t-1} \beta^{\mathbb{1}(p_{i,t} \neq y_t)}$
3. **end for**

Weighted Majority

- HALVING is equivalent to choosing $w_{i,t} = 1$ if i is "alive" ($\beta = 0$).
- What happens as $\beta \longrightarrow 1$? Faulty experts are not punished that much.

Define the *potential* $W_t = \sum_{i=1}^{n} w_{i,t}$ (the weighted size of the set of experts).

Theorem 1:

- $W_t \leq W_{t-1}$
- If $\hat{p}_t \neq y_t$ then $W_t \leq \frac{1+\beta}{2} W_{t-1}$.

Compare with before:

- W_t multiplicatively decreases when HALVING makes a mistake.
- Is there always a lower non-zero bound on W_t for all t? *Yes, with finite T. No, infinite time all weights could decay towards 0. Note that we didn't normalize the $w_{i,t}$ - we'll fix this in the next part.*
- What about the regret behavior? Is it sublinear? *We'll leave this for now.*

Multiplicative Weights algorithms

High-level intuition:

- More general case: choose from n options.
- Stochastic strategies allowed.
- Not always experts present.

Setup: at each timestep $t = 1 \dots T$:

1. you want to take a decision $d \in \mathcal{D} = \{1, \dots, n\}$
2. you choose a distribution $\mathbf{p}_t$ over $\mathcal{D}$ and sample randomly from it.
3. Nature reveals the cost vector $\mathbf{c}_t$, where each cost $c_{i,t}$ is bounded in $[-1, 1]$
4. the expected cost is then $\mathbf{E}_{i \sim \mathbf{p}_t}[\mathbf{c}_t] = \mathbf{c}_t \cdot \mathbf{p}_t$.

Goal: minimize regret with respect to the best decision in hindsight, $\min_i \sum_{t=1}^{T} c_{i,t}$.

In HALVING, decision = ”choose expert”, cost = ”mistake”, and $\mathbf{p}_t$ was 1 for the majority of alive experts (note that $\mathbf{p}$ is not ”prediction” here).

The Multiplicative Weights algorithm

Comes in many forms!

The MULTIPLICATIVE WEIGHTS algorithm

Fix $\eta \leq \frac{1}{2}$.
Initialize $w_{i,1} = 1$
Initialize $W_1 = n$
for $t = 1 \ldots T$ **do**
    $W_t = \sum_i w_{i,t}$.
    Choose a decision $i \sim \mathbf{p}_t = \frac{w_{i,t}}{W_t}$
    Nature reveals $\mathbf{c}_t$
    $w_{i,t+1} \leftarrow w_{i,t} \left(1 - \eta c_{i,t}\right)$
end for

Regret of MULTIPLICATIVE WEIGHTS

Assume cost $c_{i,t}$ is bounded in $[-1, 1]$ and $\eta \leq \frac{1}{2}$. Then after T rounds, for any decision i:

$$\sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t \leq \sum_{t=1}^{T} c_{i,t} + \eta \sum_{t=1}^{T} |c_{i,t}| + \frac{\log n}{\eta}$$

Regret of Multiplicative Weights

Regret of MULTIPLICATIVE WEIGHTS

Assume cost $c_{i,t}$ is bounded in $[-1, 1]$ and $\eta \leq \frac{1}{2}$. Then after T rounds, for any decision i:

$$\sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t \leq \sum_{t=1}^{T} c_{i,t} + \eta \sum_{t=1}^{T} |c_{i,t}| + \frac{\log n}{\eta}$$

- The cost of *any* fixed decision i
- Relates to how quickly the MWA updates itself after observing the costs at time t
- How conservative the MWA should be - if η is too small, we ”overfit” too quickly

Regret of Multiplicative Weights

Regret of multiplicative weights

Assume cost $c_{i,t}$ is bounded in $[-1, 1]$ and $\eta \leq \frac{1}{2}$. Then after T rounds, for any decision i:

$$\sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t \leq \sum_{t=1}^{T} c_{i,t} + \eta \sum_{t=1}^{T} |c_{i,t}| + \frac{\log n}{\eta}$$

- $\eta \leq \frac{1}{2}$ is needed to make some inequalities work.
- Generality: no assumptions about the sequence of events (may be correlated, costs may be adversarial)!
- Holds for all i: taking a min over decisions, we see:

$$R_T \leq \frac{\log n}{\eta} + \eta \min_i \sum_{t=1}^{T} |c_{i,t}|$$

Regret of Multiplicative Weights

$$R_T \leq \frac{\log n}{\eta} + \eta \min_i \sum_{t=1}^{T} |c_{i,t}|$$

What is the regret behavior?

Since the sum is $O(T)$, it's sub-linear with appropriate choice $\eta \sim \sqrt{\frac{\log n}{T}}$. Then

$$R_T \leq O\left(\sqrt{T \log n}\right)$$

What is the issue with this?

We need to know the horizon T up front!

If T is unknown, use $\eta_t \sim \min\left(\sqrt{\frac{\log n}{t}}, \frac{1}{2}\right)$ or the *doubling trick*.

Regret of Multiplicative Weights

$$R_T \leq \frac{\log n}{\eta} + \eta \min_i \sum_{t=1}^{T} |c_{i,t}|$$

What if we don't choose $\eta \sim \sqrt{\frac{1}{T}}$?

- What if η is constant w.r.t. T?
- What if $\eta < \frac{1}{\sqrt{T}}$?
- Fundamental tension between fitting to expert and learning from Nature
- Optimality of MW: in the online setting you can't do better: the regret is lower bounded as $R_T \geq \Omega\left(\sqrt{T \log n}\right)$. *See Theorem 4.1 in Arora.*

Regret of Multiplicative Weights

$$R_T \leq \frac{\log n}{\eta} + \eta \min_i \sum_{t=1}^{T} |c_{i,t}|$$

Why can't you achieve $O(\log T)$ regret as with FTL in this case?

- Recall for Follow The Leader with strongly convex loss, the difference between the two consecutive decisions and losses scaled as $p_t^* - p_{t-1}^* = O(\frac{1}{t})$
- In MW, the loss-differential scales with $\mathbf{p}_t - \mathbf{p}_{t-1} = O(\eta)$, which is $O(\frac{1}{\sqrt{t}})$, so the MW algorithm needs to be prepared to make much bigger changes than FTL over time.

Proof of MW

The steps in the proof are:

1. Upper bound W_t in terms of cumulative decay factors
2. Lower bound W_t by using convexity arguments
3. Combine the upper and lower bounds on W_t to get the answer

Let's go through the proof carefully.

Proof of MW: Getting the upper bound on W_t

$$W_{t+1} = \sum_i w_{i,t+1}$$

Proof of MW: Getting the upper bound on W_t

$$
\begin{aligned}
W_{t+1} &= \sum_i w_{i,t+1} \\
&= \sum_i w_{i,t} \left(1 - \eta c_{i,t}\right)
\end{aligned}
$$

Proof of MW: Getting the upper bound on W_t

$$
\begin{aligned}
W_{t+1} &= \sum_i w_{i,t+1} \\
&= \sum_i w_{i,t}\left(1 - \eta c_{i,t}\right) \\
&= W_t - \eta W_t \sum_i c_{i,t} p_{i,t} \qquad\qquad w_{i,t} = W_t p_{i,t}
\end{aligned}
$$

Proof of MW: Getting the upper bound on W_t

$$
\begin{aligned}
W_{t+1} &= \sum_i w_{i,t+1} \\
&= \sum_i w_{i,t}\left(1 - \eta c_{i,t}\right) \\
&= W_t - \eta W_t \sum_i c_{i,t} p_{i,t} && w_{i,t} = W_t p_{i,t} \\
&= W_t\left(1 - \eta \mathbf{c}_t \cdot \mathbf{p}_t\right)
\end{aligned}
$$

Proof of MW: Getting the upper bound on W_t

$$
\begin{aligned}
W_{t+1} &= \sum_i w_{i,t+1} \\
&= \sum_i w_{i,t}\left(1 - \eta c_{i,t}\right) \\
&= W_t - \eta W_t \sum_i c_{i,t} p_{i,t} && w_{i,t} = W_t p_{i,t} \\
&= W_t\left(1 - \eta \mathbf{c}_t \cdot \mathbf{p}_t\right) \\
&\leq W_t \exp\left(-\eta \mathbf{c}_t \cdot \mathbf{p}_t\right) && \text{convexity}
\end{aligned}
$$

Proof of MW: Getting the upper bound on W_t

$$\begin{aligned} W_{t+1} &= \sum_i w_{i,t+1} \\ &= \sum_i w_{i,t}\left(1 - \eta c_{i,t}\right) \\ &= W_t - \eta W_t \sum_i c_{i,t} p_{i,t} && w_{i,t} = W_t p_{i,t} \\ &= W_t\left(1 - \eta \mathbf{c}_t \cdot \mathbf{p}_t\right) \\ &\leq W_t \exp\left(-\eta \mathbf{c}_t \cdot \mathbf{p}_t\right) && \text{convexity} \end{aligned}$$

Recursively using this inequality:

$$\begin{aligned} W_{T+1} &\leq W_1 \exp\left(-\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t\right) \\ &= n \cdot \exp\left(-\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t\right) && \forall x \in \mathbb{R} : 1 - x \leq e^{-x} \end{aligned}$$

Proof of MW: Getting the lower bound on W_t

$$W_{T+1} \geq w_{i,T+1}$$

Proof of MW: Getting the lower bound on W_t

$$
\begin{aligned}
W_{T+1} &\geq w_{i,T+1} \\
&= \prod_{t \leq T} (1 - \eta c_{i,t}) && \text{multipl. updates and } w_{i,1} = 1
\end{aligned}
$$

Proof of MW: Getting the lower bound on W_t

$$
\begin{aligned}
W_{T+1} &\geq w_{i,T+1} \\
&= \prod_{t \leq T} (1 - \eta c_{i,t}) && \text{multipl. updates and } w_{i,1} = 1 \\
&= \prod_{t: c_{i,t} \geq 0} (1 - \eta c_{i,t}) \prod_{t: c_{i,t} < 0} (1 - \eta c_{i,t}) && \text{split positive + negative costs}
\end{aligned}
$$

Proof of MW: Getting the lower bound on W_t

$$
\begin{aligned}
W_{T+1} &\geq w_{i,T+1} \\
&= \prod_{t \leq T} (1 - \eta c_{i,t}) && \text{multipl. updates and } w_{i,1} = 1 \\
&= \prod_{t: c_{i,t} \geq 0} (1 - \eta c_{i,t}) \prod_{t: c_{i,t} < 0} (1 - \eta c_{i,t}) && \text{split positive + negative costs}
\end{aligned}
$$

Now we use the fact that:

- $\forall x \in [0, 1] : (1 - \eta)^x \leq (1 - \eta x)$
- $\forall x \in [-1, 0] : (1 + \eta)^x \leq (1 - \eta x)$

Proof of MW: Getting the lower bound on W_t

$$\begin{aligned} W_{T+1} &\geq w_{i,T+1} \\ &= \prod_{t \leq T} (1 - \eta c_{i,t}) && \text{multipl. updates and } w_{i,1} = 1 \\ &= \prod_{t: c_{i,t} \geq 0} (1 - \eta c_{i,t}) \prod_{t: c_{i,t} < 0} (1 - \eta c_{i,t}) && \text{split positive + negative costs} \end{aligned}$$

Now we use the fact that:

- $\forall x \in [0, 1] : (1 - \eta)^x \leq (1 - \eta x)$
- $\forall x \in [-1, 0] : (1 + \eta)^x \leq (1 - \eta x)$

Since $c_{i,t} \in [-1, 1]$, we can apply this to each factor in the product above:

$$W_{T+1} \geq (1 - \eta)^{\sum_{\geq 0} c_{i,t}} (1 + \eta)^{-\sum_{<0} c_{i,t}}$$

This is the lower bound we want.

Proof of MW: Combine the upper and lower bounds

We get:

$$(1-\eta)^{\sum_{\geq 0} c_{i,t}} (1+\eta)^{-\sum_{<0} c_{i,t}} \leq W_{T+1} \leq n \cdot \exp\left(-\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t\right)$$

Take logs:

$$\sum_{\geq 0} c_{i,t} \log(1-\eta) - \sum_{<0} c_{i,t} \log(1+\eta) \leq \log n - \eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t$$

Proof of MW: Combine the upper and lower bounds

$$\sum_{\geq 0} c_{i,t} \log(1-\eta) - \sum_{<0} c_{i,t} \log(1+\eta) \leq \log n - \eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t$$

Now we'll massage this into the form we want:

Proof of MW: Combine the upper and lower bounds

$$\sum_{\geq 0} c_{i,t} \log(1-\eta) - \sum_{<0} c_{i,t} \log(1+\eta) \leq \log n - \eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t$$

Now we'll massage this into the form we want:

$$\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t \leq \log n - \sum_{\geq 0} c_{i,t} \log(1-\eta) + \sum_{<0} c_{i,t} \log(1+\eta)$$

Proof of MW: Combine the upper and lower bounds

$$\sum_{\geq 0} c_{i,t} \log(1-\eta) - \sum_{<0} c_{i,t} \log(1+\eta) \leq \log n - \eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t$$

Now we'll massage this into the form we want:

$$\begin{aligned}
\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t &\leq \log n - \sum_{\geq 0} c_{i,t} \log(1-\eta) + \sum_{<0} c_{i,t} \log(1+\eta) \\
&= \log n + \sum_{\geq 0} c_{i,t} \log\left(\frac{1}{1-\eta}\right) + \sum_{<0} c_{i,t} \log(1+\eta)
\end{aligned}$$

Proof of MW: Combine the upper and lower bounds

Since $\eta \leq \frac{1}{2}$, we can use $\log\left(\frac{1}{1-\eta}\right) \leq \eta + \eta^2$ and $\log(1+\eta) \geq \eta - \eta^2$:

$$\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t$$

Proof of MW: Combine the upper and lower bounds

Since $\eta \leq \frac{1}{2}$, we can use $\log\left(\frac{1}{1-\eta}\right) \leq \eta + \eta^2$ and $\log(1+\eta) \geq \eta - \eta^2$:

$$\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t$$

$$\leq \log n + \sum_{\geq 0} c_{i,t} \log\left(\frac{1}{1-\eta}\right) + \sum_{<0} c_{i,t} \log(1+\eta)$$

Proof of MW: Combine the upper and lower bounds

Since $\eta \leq \frac{1}{2}$, we can use $\log\left(\frac{1}{1-\eta}\right) \leq \eta + \eta^2$ and $\log(1+\eta) \geq \eta - \eta^2$:

$$\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t$$

$$\leq \log n + \sum_{\geq 0} c_{i,t} \log\left(\frac{1}{1-\eta}\right) + \sum_{<0} c_{i,t} \log(1+\eta)$$

$$\leq \log n + \sum_{\geq 0} c_{i,t}\left(\eta + \eta^2\right) + \sum_{<0} c_{i,t}\left(\eta - \eta^2\right)$$

use inequalities

Proof of MW: Combine the upper and lower bounds

Since $\eta \leq \frac{1}{2}$, we can use $\log\left(\frac{1}{1-\eta}\right) \leq \eta + \eta^2$ and $\log(1+\eta) \geq \eta - \eta^2$:

$$\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t$$

$$\leq \log n + \sum_{\geq 0} c_{i,t} \log\left(\frac{1}{1-\eta}\right) + \sum_{<0} c_{i,t} \log(1+\eta)$$

$$\leq \log n + \sum_{\geq 0} c_{i,t} \left(\eta + \eta^2\right) + \sum_{<0} c_{i,t} \left(\eta - \eta^2\right) \qquad \text{use inequalities}$$

$$= \log n + \eta \sum_{t=1}^{T} c_{i,t} + \eta^2 \sum_{\geq 0} c_{i,t} - \eta^2 \sum_{<0} c_{i,t}$$

Proof of MW: Combine the upper and lower bounds

Since $\eta \leq \frac{1}{2}$, we can use $\log\left(\frac{1}{1-\eta}\right) \leq \eta + \eta^2$ and $\log(1+\eta) \geq \eta - \eta^2$:

$$
\begin{aligned}
&\eta \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t \\
&\leq \log n + \sum_{\geq 0} c_{i,t} \log\left(\frac{1}{1-\eta}\right) + \sum_{<0} c_{i,t} \log(1+\eta) \\
&\leq \log n + \sum_{\geq 0} c_{i,t}\left(\eta + \eta^2\right) + \sum_{<0} c_{i,t}\left(\eta - \eta^2\right) && \text{use inequalities} \\
&= \log n + \eta \sum_{t=1}^{T} c_{i,t} + \eta^2 \sum_{\geq 0} c_{i,t} - \eta^2 \sum_{<0} c_{i,t} \\
&= \log n + \eta \sum_{t=1}^{T} c_{i,t} + \eta^2 \sum_{t=1}^{T} |c_{i,t}| && \text{combine sums}
\end{aligned}
$$

Proof of MW: Combine the upper and lower bounds

Dividing by η, we get what we want:

$$\sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{p}_t \leq \frac{\log n}{\eta} + \sum_{t=1}^{T} c_{i,t} + \eta \sum_{t=1}^{T} |c_{i,t}|$$

Some remarks

- Matrix form of MW
- Gains instead of losses

See Arora's paper for more details

Learning with Experts revisited

Multiplicative Weights with continuous label spaces.

Exponential Weighted Average: continuous prediction case

Formal setup: there are T rounds in which you predict a label $\hat{p}_t \in \mathcal{C}$.

$\mathcal{C}$ is a convex subset of some vector space.

At every timestep t:

1. Each expert $i = 1 \ldots n$ predicts $p_{i,t} \in \mathcal{C}$
2. You make a prediction $\hat{p}_t \in \mathcal{C}$
3. Nature reveals $y_t \in \mathcal{Y}$
4. We suffer a loss $l(\hat{p}_t, y_t)$
5. Each expert suffers a loss $l(p_{i,t}, y_t)$

Loss: $\hat{L}_t = \sum_{t=1}^{T} l(\hat{p}_t, y_t), L_{i,t} = \sum_{t=1}^{T} l(p_{i,t}, y_t)$

Regret: $R_t = \hat{L}_t - \min_i L_{i,t}$

Exponential Weighted Average: continuous prediction case

We assume that the loss $l(\hat{p}_t, y_t)$ as a function $l : \mathcal{C} \times \mathcal{Y} \longrightarrow \mathbb{R}$:

- is **bounded**: $\forall p \in \mathcal{C}, y \in \mathcal{Y} : l(p, y) \in [0, 1]$
- $l(., y)$ is **convex** for any fixed $y \in \mathcal{Y}$

Exponential Weighted Average

This brings us to a familiar algorithm.

Choose an $\eta > 0$ (we'll make this more directed later).

Exponential Weighted Average algorithm

Initialize $w_{i,0} = 1, W_0 = n$.
for $t = 1 \dots T$ **do**
    Receive expert predictions $(p_{1,t} \dots p_{N,t})$
    Predict $\hat{p}_t = \frac{\sum_{i=1}^{N} w_{i,t-1} p_{i,t}}{W_{t-1}}$
    Receive Nature's answer y_t (and incur loss $l(\hat{p}_t, y_t)$)
    $w_{i,t} \leftarrow w_{i,t-1} e^{-\eta l(p_{i,t}, y_t)}$
    $W_t \leftarrow \sum_{i=1}^{N} w_{i,t}$
end for

Regret of EWA

Assume that the loss l is a function $l : \mathcal{C} \times \mathcal{Y} \longrightarrow [0,1]$, where $\mathcal{C}$ is convex and $l(.,y)$ is convex for any fixed $y \in \mathcal{Y}$. Then

$$R_T = \hat{L}_T - \min_i L_{i,T} \leq \frac{\log n}{\eta} + \frac{\eta T}{8}$$

Hence, if we choose $\eta = \sqrt{\frac{8 \log n}{T}}$, the regret $R_T \leq \sqrt{\frac{T}{2} \log n} = O(\sqrt{T})$.

The proof follows from an application of Jensen's inequality and Hoeffding's lemma. See chapter 3 of *Bartok* for details.

Again note that we need to know what T is in advance.

So how about the regret in the discrete case?

Regret of Weighted Majority

Let $\mathcal{C} = \mathcal{Y}$, $\mathcal{Y}$ have at least 2 elements and $l(p, y) = \mathbf{1}(p \neq y)$. Let $L^*_{i,T} = \min_i L_{i,T}$.

$$R_T = \hat{L}_T - \min_i L_{i,T} \leq \frac{\left(\log_2\left(\frac{1}{\beta}\right) - \log_2\left(\frac{2}{1+\beta}\right) L^*_{i,T}\right) + \log_2 N}{\log_2\left(\frac{2}{1+\beta}\right)}$$

This bound is of the form $R_T = aL^*_{i,T} + b = O(T)$!

- The discrete case is harder than the continuous case if we stick to **deterministic** algorithms.
- The worst-case regret is achieved by a set of 2 experts and an outcome sequence y_t such that $\hat{L}_T = T$.

See chapter 4 of *Bartok* for details.

Today

What should you (at least) take away from this lecture?

How should you base your prediction on expert predictions?

- Can use a weighted majority algorithm, which is a type of MWA.
- General framework for online learning with experts (e.g. boosting).

What are the characteristics of multiplicative weights algorithms? ...

- Few assumptions on costs / Nature (can be adversarial), therefore broadly applicable.
- Fundamental tension between fitting to decision and learning from Nature.
- Tends to have instantenous loss-differential $O(\frac{1}{\sqrt{t}})$, worse than the version of FTL $O(\frac{1}{t})$ that we saw in lecture 1.

Questions?