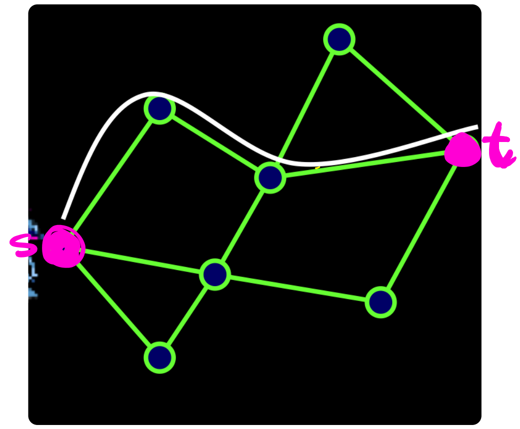
Repeated Online Decision Making and the Multiplicative Weights Algorithm



A set of possible actions $|A|=n$

T a time horizon $A=\{1,2,\dots,n\}$

Setup:

At each time step $t=1..T$

- a decision maker picks an action $a_t \in A$
  where $p^t_i = Pr(a_t = i)$
  $$\vec{p}^{\,t} = (p^t(1), p^t(2), \dots, p^t(n))$$
- an adversary picks reward vector
  $$\vec{r}^{\,t} = (r^t(1), r^t(2), \dots, r^t(n))$$
  where $r^t(i)$ = reward to alg. if picked action i
- decision maker learns r^t

Goal of alg: maximize total exp reward.

$$= \sum_{t=1}^{T} \underbrace{\sum_{i=1}^{n} p^t(i)\, r^t(i)}_{\vec{p}^{\,t} \cdot \vec{r}^{\,t}}$$

Examples:

1. Choosing a route
2. Choosing stocks to buy

	Expert 1		Expert 2		Expert 3		
	$p^t(1)$	$r^t(1)$	$p^t(2)$	$r^t(2)$	$p^t(3)$	$r^t(3)$	$\vec{p}^{\,t} \cdot \vec{r}^{\,t}$
$t=1$							
$t=2$							
$t=3$							

Alg total reward $= \sum_{t=1}^{3} \vec{p}^{\,t} \cdot \vec{r}^{\,t} =$

Totals

Exp 1 Exp 2 Exp 3

Best possible result

$$\text{total reward} = \sum_{t=1}^{T} \max_i r^t(i) \quad (*)$$

Observation: this benchmark is too strong

Ex: $A = \{1,2\}$

Adv: if $p^t(1) \geq \frac{1}{2} \Rightarrow r^t(1) = -1$, $r^t(2) = 1$

$p^t(1) + p^t(2) = 1$

if $p^t(1) < \frac{1}{2} \Rightarrow r^t(1) = 1$, $r^t(2) = -1$

$E(\text{reward}) \leq 0$ whereas $(*) = T$

To make progress, weaken benchmark

$$\underset{T \text{ step regret}}{\text{Regret}(\vec{p}^1, \ldots, \vec{p}^T)} = \underbrace{\max_{a \in A} \sum_{t=1}^{T} r_t(a)}_{\text{best reward possible if you use same action every day}} - \underbrace{\sum_{t=1}^{T} \vec{p}^t \cdot \vec{r}^t}_{\text{alg total exp reward}}$$

$$\text{Avg Regret} = \frac{1}{T}\text{Regret}$$

Goal: Avg regret $\to 0$

Most obvious thing to try to minimize Regret

"Follow the Leader": set $p^t(i) = \begin{cases} 1 & \sum_{\tau=1}^{t-1} r_\tau(i) > \sum_{\tau=1}^{t-1} r_\tau(j) \quad \forall j \neq i \\ 0 & \text{o.w.} \end{cases}$

(break ties arbitrarily)

Claim: no good. In fact, no det alg good

adversary sets $r^t(a) = \begin{cases} 0 & \text{if alg chooses action } a \\ 1 & \text{o.w.} \end{cases}$

Alg total reward = $\qquad \max_{a \in A} \sum_{t=1}^{T} r^t(a) = ?$

How well can we do with a randomized alg?

simple lower bound.

$n = 2 \qquad \vec{r}^{\,t} = \begin{cases} (1,-1) & \text{wp. } \frac{1}{2} \\ (-1,1) & \text{wp } \frac{1}{2} \end{cases}$

E[reward of any alg] =

What about best action in hindsight?

if $H > T \Rightarrow$ choose action 1
else action 2

reward of best action

can extend to n actions & show

every alg has regret $\Omega\left(\sqrt{T \log n}\right)$

avg regret $\Omega\left(\sqrt{\frac{\log n}{T}}\right)$

$\boxed{\textbf{Theorem:}}$ $\exists$ online alg (MWU, Hedge, ...) s.t.

$$\max_a \sum_t r_t(a) - E(\text{reward of online alg}) \le 2\sqrt{T \log n}$$

$\equiv$ avg per-step reward of online alg $\geq$ avg reward of best action $- 2\sqrt{\frac{\log n}{T}}$

MWU algorithm

initialize $w^1(a) = 1 \quad \forall a \in A$

for $t = 1$ to T

pick action a with probability proportional to $w^t(a)$

given r^t, update wts as follows:

$$w^{t+1}(a) = w^t(a) \cdot (1 + \eta r^t(a))$$

Another interesting setting: "bandits".

$\xrightarrow{p^t}$ sample $a_t \sim p^t$

$\xleftarrow[r^t(a_t)]{}$

Can still get avg regret. $\longrightarrow 0$

Application domain	Action	Reward
medical trials	which drug to prescribe	health outcome.
web design	*e.g.,* font color or page layout	#clicks.
content optimization	which items/articles to emphasize	#clicks.
web search	search results for a given query	1 if the user is satisfied.
advertisement	which ad to display	revenue from ads.
recommender systems	*e.g.,* which movie to watch	1 if follows recommendation.
sales optimization	which products to offer at which prices	revenue.
procurement	which items to buy at which prices	#items procured
auction/market design	*e.g.,* which reserve price to use	revenue
crowdsourcing	which tasks to give to which workers, and at which prices	1 if task completed at sufficient quality.
datacenter design	*e.g.,* which server to route the job to	job completion time.
Internet	*e.g.,* which TCP settings to use?	connection quality.
radio networks	which radio frequency to use?	1 if successful transmission.
robot control	a "strategy" for a given task	job completion time.

Minimax Theorem

Let A be $m \times n$ payoff matrix for zero-sum game

$\left[a_{ij} : \; = \begin{array}{l} \text{gain of row player} \\ \text{loss of col player} \end{array} \right.$ when row player plays i & col player plays j

Nature flips coins
submit distn

Let $\vec{x} \in \mathbb{R}^m$ $\sum_{i=1}^{m} x_i = 1$, $x_i \geq 0$ be mixed strategy for row player

$\vec{y} \in \mathbb{R}^n$ $\sum_{j=1}^{n} y_j = 1$, $y_j \geq 0$ be mixed strategy for col player

$$\max_{\vec{x} \in \Delta_m} \min_{\vec{y} \in \Delta_n} x^T A y = \min_{\vec{y} \in \Delta_n} \max_{\vec{x} \in \Delta_m} x^T A y$$

$V = V_R = V_C$ called "value" of game

x^*, y^* called "optimal" strategies

FIGURE 2.6. Von Neumann explaining duality to Dantzig.

Proof of Minimax Thm using MWU thm (sketch)

[assume $a_{ij} \in [-1,1]$ wlog scale]

Thought experiment

each player is "adversary" for other

Fix $\varepsilon > 0$ (eventually we'll take $\varepsilon \to 0$)

for $t = 1 \ldots T = \frac{4 \ln(\max(n,m))}{\varepsilon^2}$

Imagine row player & col player using MWU alg to play T rounds of game

$\Rightarrow \vec{p}^t, \vec{q}^t, \quad t = 1..T$

where in each round

row player rewards $r^t = Aq^t$

col player rewards $r^t = -(p^t)^T A$

Let $\hat{x} = \frac{1}{T} \sum_{t=1}^{T} \vec{p}^t$

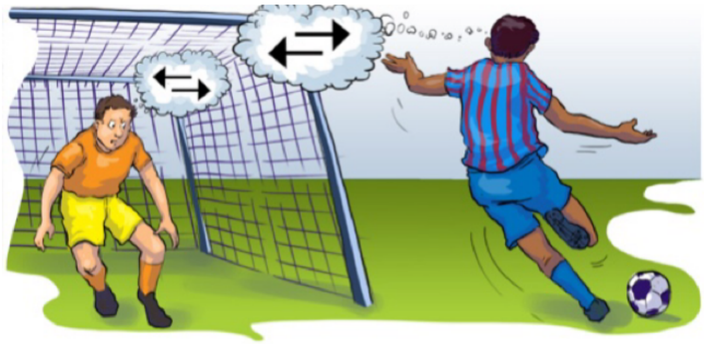
$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \vec{q}^t$

Let $V = \frac{1}{T} \sum_{t=1}^{T} (\vec{p}^t)^T A \vec{q}^t$

avg exp payoff of row player over T rounds

Is it predictive?

Penalty Kicks.



col player goalee

row player kicker	L	R
L	0.58	0.95
R	0.93	0.7

Based on actual data on 1417 penalty kicks from professional games in Europe

	Kicker	Goalee
Optimal strategies	(0.38, 0.62)	(0.42, 0.58)
Observed frequencies	(0.40, 0.60)	(0.423, 0.577)