Algorithms, Games, and Networks January 22, 2013

Lecture 3

Lecturer: Avrim Blum *Scribe: Will Frankenstein*

Readings: 4.4-4.6 of *AGT*

Today's lecture:

1. Recap nice properties of Randomized Weighted Majority (RWM)
2. Bandit algorithms
3. Internal and swap regret, correlated equilibria

1 Recap of RWM

- Algo has N options (can view as rows in a matrix game). Picks an action. World picks what life is going to be that day (can view as columns in the matrix).
- Algo pays cost / gets benefit for action
- Column is given as feedback
- Goal: do nearly as well as best fixed row in hindsight.
- RWM Algo: Begin with weights $w_i^0 = 1$. At time t, penalize as $(1\text{-}\varepsilon c_i^t)$, where we have scaled so that all costs $\in [0, 1]$.

Weights (after 2 iterations)	World		
$(1\text{-}\varepsilon c_1^2)(1\text{-}\varepsilon c_1^1)1$	$\lceil c_1 \rceil$	...	$\lceil c_2 \rceil$
$\vdots$	$\vert c_1 \vert$	...	$\vert c_2 \vert$
$(1\text{-}\varepsilon c_n^2)(1\text{-}\varepsilon c_n^1)1$	$\lfloor c_1 \rfloor$	...	$\lfloor c_2 \rfloor$

If we set ϵ to balance out the two loss terms, we get:

$$E[cost] \quad \leq OPT + 2(OPT * \log n)^{\frac{1}{2}}$$

$\therefore$ as OPT $\leq$ T, the upper limit of $\frac{regret}{time}$ is bounded by $O\left(\left(\frac{\log n}{T}\right)^{\frac{1}{2}}\right)$.
Furthermore, we can use this to argue that if we want this regret to be below a threshold

δ, we can determine the minimum number of time steps T necessary:

$$\begin{aligned} \left(\frac{\log n}{T}\right)^{\frac{1}{2}} &\leq \delta \\ \frac{\log n}{\delta^2} &\leq T \end{aligned}$$

2 Bandits

What is the bandit setting? How do we apply the RWM to a bandit setting?

Bandit setting: assume that you're only going to get your own cost/benefit information - not the cost information associated with your other options. (Originally - one armed bandit referred to slot machines.)

Two versions of the bandit problem:

- Stochastic bandit problem: Each bandit is like a coin of some bias. Outcomes for bandit i are independent draws from some probability distribution (uses martingales)
- Adversarial bandit problem: Non-independent, arbitrary sequence of events

We focus on the adversarial bandit. Auer, Cesa-Bianchi, Freund, and Schapire [2002] show how to use RWM as a subroutine to get an algorithm with cumulative regret $O(\sqrt{TN \log N})$ and therefore average regret on the order of $O(((N \log N)/T)^{0.5})$. We will examine their algorithm, called exp^3, but prove a weaker bound where the dependence on T is $T^{2/3}$ rather than $T^{1/2}$.

We will do this in the context of online pricing, e.g., selling lemonade. We can think of each possible price as a row or expert. Our goal is to do nearly as well as best fixed price in hindsight. In particular, the setting is: for $t = 1, 2, 3. \ldots$

- Seller sets price p^t
- Buyer has valuation v^t (assume $\forall t,\ v^t \leq h$)
- If $v^t \geq p^t$ buyer purchases and pays p^t; else not
- Repeat

If we knew v^t, then we could run RWM, with $E[gain] \geq OPT(1-\epsilon) - O(\epsilon^{-1} h \log n)$. How can we adapt RWM for this multi-armed bandit problem?

Definition 1 *exp^3: Exponential Weights for Exploration and Exploitation. This algorithm uses RWM as a subroutine and operates as follows. Let n denote the number of experts:*

- At time t the algorithm receives probability distribution p^t over experts (actions) $1, \ldots, n$ proposed by the RWM algorithm.
- Our algorithm will convert p^t to a new distribution q^t and select an expert (action) i from q^t. Specifically, $q^t = (1-\gamma)p^t + \gamma$ unif.
- Our algorithm will then receive a gain g_i^t. It then converts it to a gain vector $\hat{g}^t$ to give to RWM. Specifically, $\hat{g}^t = (0, \ldots, 0, g_i^t/q_i^t, 0, \ldots, 0)$.
- Note that $g_i^t/q_i^t \leq nh/\gamma$, since $q_i^t \geq \gamma/n$. This is mathematically the reason we mix p^t with the uniform distribution in defining q^t.

The analysis is based on the following four steps. Let $\widehat{OPT}$ denote the gain of the best expert according to the made-up gain vectors $\hat{g}^t$. I.e., this is what the RWM subroutine believes is the gain of the best expert.

1. RWM believes gain to be $p^t \cdot \hat{g}^t = p_i^t(g_i^t/q_i^t) = g_{RWM}^t$
2. $\sum_t g_{RWM}^t \geq \widehat{OPT}(1-\epsilon) - O(\epsilon^{-1}(nh/\gamma)\log n)$. This is by the standard analysis of RWM for gains, where the "nh/γ" term is due to the range of the payoffs.
3. True gain: $g_i^t = g_{RWM}^t(q_i^t/p_i^t) \geq g_{RWM}^t(1-\gamma)$
4. $E[\widehat{OPT}] \geq OPT$. Why?$E[\hat{g}_j^t] = (1-q_j^t) * 0 + q_j^t(g_jt/q_j^t) = g_j^t$.

   $\therefore\ E[\max_j[\sum_t \hat{g}_j^t]] \geq \max_j[E[\sum_t \hat{g}_j^t]] = OPT$.

   The expected value of a max is always greater than or equal to the max of the expected values. That's because you can think of the former as repeating the process 1000 times, each time picking the best j and then averaging. On the other hand, the latter corresponds to the case where you have to pick the *same* j each of the 1000 times, so you pick the j of highest expectation.

In conclusion (where $\gamma = \epsilon$), summing item (3) over all times t, plugging into (2), and taking expectation, we have, using (4):

$$E[Exp3] \ \geq\ OPT(1-\epsilon)^2 - O(\epsilon^{-2}nh\log(n)).$$

Summary:

- We have algorithms for online decision-making with strong guarantees on performance compared to best fixed choice.

  - Can be affected by discretization choices (which/how many n?)

  - Application: If we play a repeated game, we can do nearly as well as a fixed strategy in hindsight
- Algo can be applied even with limited feedback.

  - Application: Which way to drive to work; online pricing, even if there's only buy/no buy feedback

3 Internal regret, swap regret, correlated equilibria

What happens when players minimize regret? We know:

- In zero-sum games, empirical frequencies → minimax optimal
- In general sum games, however, the empirical frequencies need not approach a Nash equilibrium (NE). (If they converge at all, they have to converge to a NE else some player would have regret, but they might not converge).

For example, Zinkevich'04 presents a 4-action game that can be viewed as a game of rock-paper-scissors-foosball with a single, pure-strategy Nash equilibrium (both playing foosball) but where the RWM will cycle among the first three actions and do worse than the NE.

Balcan, Constantin, and Mehta show that failure to converge can happen even in rank-1 games (games where $R + C$ has rank 1 — i.e. only one independent column vector). This is interesting because in rank 1 games, Nash equilibria can be found in polynomial time.

Here we will show that if algorithms minimize "internal" or "swap" regret, the empirical distribution of play will be an approximate *correlated* equilibrium.

Definition 2 *External Regret ("best expert" regret): Given n strategies, compete with best strategy in hindsight.*

Definition 3 *Sleeping Expert Regret (regret with time intervals): Given n strategies, k properties. Let S_i be set of days satisfying property i (can overlap). Simultaneously achieve low regret over each S_i. Think of as judging over several sets: rainy days, Mondays, etc.*

Definition 4 *Internal Regret / Swap Regret: like sleeping experts, except that S_i is the set of days in which we choose strategy i.*

For example, if buying stocks then this corresponds to regret of the form: "every time I bought MSFT, should have bought APPL."

Formally, *swap regret* is regret wrt the optimal "rewiring" function $f : \{1, \ldots, , n\} \longrightarrow \{1, \ldots, n\}$ such that every time action j is chosen by the algorithm, the comparator plays $f(j)$. External regret would correspond to just considering the n constant rewiring functions. Internal regret is regret with respect to the optimal rewiring of just one action: i.e., the optimal f such that $f(j) = j$ for all but one j.

Definition 5 *Correlated equilibrium: Distribution p_{ij} over entries in the game matrix such that if a trusted party chooses one at random and tells you your part, you have no incentive to deviate.*

Any Nash equilibrium (p, q) induces a correlated equilibrium $p_{ij} = p_i q_j$. However, there are also correlated equilibria that are not Nash equilibria. An interesting example is the Shapley game, which is like Rock-Paper-Scissors except if both players choose the same action, they both lose. I.e.,

	R	P	S
R	-1,-1	-1,1	1,-1
P	1,-1	-1,-1	-1,1
S	-1,1	1,-1	-1,-1

This has a correlated equilibrium that is better than any NE, namely the uniform distribution over the six entries not on the diagonal. Games can also have correlated equilibria that are *worse* than any NE.

If all players have low swap-regret, then empirical distribution of play is an approximate correlated equilibrium. We can see this by thinking of the empirical distribution of play like this. Suppose players have played for T time steps. Then:

- Correlator chooses random time $t \in \{1, \ldots, T\}$
- Tells each player to play action j at time t but does not reveal t

Thinking about it this way, we can see that $E[\text{incentive to deviate}] = \sum_j Pr(j)(\text{Regret } | j)$ =swap-regret of algorithm This suggests it is reasonable to see correlated equilibria in multi-agent systems where individuals optimize for themselves.

One more notion: *coarse-correlated equilibrium.*

Correlated equilibrium: No incentive to deviate even after seeing advice.

Coarse-correlated equilibrium: if choice is between seeing and following advice, or not to see at all, would prefer former (i.e. see and follow advice). Think of this as giving a financial advisor your money vs not getting their advice (as opposed to CE where you can decide what to do after seeing what the advice is).

Not hard to see: if there's low external regret, we have an approximate coarse-correlated equilibrium.

4 Summary

- exp^3 algorithm, which allows for using RWM in a multi-armed bandit problem (e.g. online pricing)
- Definitions of internal/swap regret and correlated equilibria
- Fun facts: Robert Aumann, who defined the notion of correlated equilibrium, received the Nobel prize in economics in 2005. Shapley received the Nobel prize in 2012. We'll see more econmics Nobels later.