Bandit algorithms, internal & swap regret, and correlated equilibria

Your guide: Avrim Blum

(Readings: Ch. 4.4-4.6 of AGT book)

RWM

\[ \begin{array}{c|c|c|c|c} \text{World – Life – Fate} \\ \\ c_1 \vdots c_n \end{array} \]

scaling so costs in [0,1]

Guarantee: \( E[\text{cost}] \leq \text{OPT} + 2(\text{OPT} \cdot \log n)^{1/2} \)

Since \( \text{OPT} \leq T \), this is at most \( \text{OPT} + 2(T \log n)^{1/2} \).

So, regret/time step \( \leq 2(T \log n)^{1/2}/T \to 0. \)

[ACFS02]: applying RWM to bandit setting

- What if only get your own cost/benefit as feedback?
- Use of RWM as subroutine to get algorithm with cumulative regret \( O((TN \log N)^{1/2}) \).
- [average regret \( O((N \log N)/T)^{1/2}) \].
- Will do a somewhat weaker version of their analysis (same algorithm but not as tight a bound).
- For fun, talk about it in the context of online pricing...

Online pricing

- Say you are selling lemonade (or a cool new software tool, or bottles of water at the world cup).
- For \( t=1,2,\ldots,T \)
  - Seller sets price \( p_t \)
  - Buyer arrives with valuation \( v_t \)
  - If \( v_t \geq p_t \), buyer purchases and pays \( p_t \), else doesn’t.
  - Repeat.
- Assume all valuations \( \leq h. \)
- Goal: do nearly as well as best fixed price in hindsight.
- If \( v_t \) revealed, run RWM. \( E[\text{gain}] \geq \text{OPT}(1-\epsilon) - O(\epsilon^{-1} h \log n) \).

Multi-armed bandit problem

\[ \begin{align*}
\text{Opt} & \\
\text{Expert} i & \sim q^i \\
\text{Distrib} p^i & \text{Gain vector} \hat{g}^i \\
& \hat{g}^i = (1-q^i) \cdot \text{unif} \\
& + q^i \cdot (0,0,\ldots,0,1) \\
& 1. \text{RWM believes gain is: } p^i - \hat{g}^i = p_i/g_i(q_i/q) = g_i^{\text{RWM}} \\
& 2. \sum g_i^{RWM} \geq \text{OPT}(1-\epsilon) \cdot O(\epsilon^{-1} h \log n) \\
& 3. \text{Actual gain is: } g_i^* = g_i^{RWM} (g_i/p_i) \geq g_i^{\text{RWM}} (1-\epsilon) \\
& 4. \text{E}[\text{Opt}] \geq \text{OPT}. \text{ Because } E[\hat{g}_i] = (1-q^i) \hat{g}_i/q_i + q^i g_i/q, \text{ so } E[\text{Opt}] = \text{OPT}. 
\end{align*} \]
Multi-armed bandit problem

Exponential Weights for Exploration and Exploitation ($\text{exp}^3$)

[$\text{Auer, Cesa-Bianchi, Freund, Schapire}$]

\[ \text{OPT} \]

\[ \text{Exp3} \]

\[ \text{RWM} \]

\[ n = \#\text{experts} \]

\[ \text{Gain vector } \hat{\beta} \]

\[ q^i = (1 - \varepsilon^2) p^i + \varepsilon \text{ unif} \]

\[ \hat{\beta} = (0, \ldots, 0, g^i / q^i, 0, \ldots, 0) \]

Conclusion ($\gamma = \varepsilon$):

\[ E[\text{Exp3}] \leq \text{OPT}(1 - \varepsilon^2) - O(\varepsilon^2 \log(n)) \]

Balancing would give $O((\text{OPT} \log(n))^{1/3})$ in bound because of $\varepsilon^2$. But can reduce to $\varepsilon^2$ and $O((\text{OPT} \log(n))^{1/2})$ more care in analysis.

Summary

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

- Application: play repeated game against adversary. Perform nearly as well as fixed strategy in hindsight.

Can apply even with very limited feedback.

- Application: which way to drive to work, with only feedback about your own paths; online pricing, even if only have buy/no buy feedback.

Internal/Swap Regret and Correlated Equilibria

What if all players minimize regret?

- In zero-sum games, empirical frequencies quickly approaches minimax optimal.

- In general-sum games, does behavior quickly (or at all) approach a Nash equilibrium?
  - After all, a Nash Eq is exactly a set of distributions that are no-regret wrt each other. So if they converge at all, they must converge to a Nash equil.
  - Well, unfortunately, no.

A bad example for general-sum games

- Augmented Shapley game from [Zinkevich04]:
  - First 3 rows/cols are Shapley game (rock / paper / scissors) but if both do same action then both lose.
  - 4th action "play foosball" has slight negative if other player is still doing r/p/s but positive if other player does 4th action too.

RWM will cycle among first 3 and have no regret, but do worse than only Nash Equilibrium of both playing foosball.

- We didn't really expect this to work given how hard NE can be to find...

A bad example for general-sum games

- [Balcan-Constantin-Mehta12]:
  - Failure to converge even in Rank-1 games (games where $R+C$ has rank 1).
  - Interesting because one can find equilibria efficiently in such games.
What can we say?
If algorithms minimize "internal" or "swap" regret, then empirical distribution of play approaches correlated equilibrium.
- Foster & Vohra, Hart & Mas-Colell,
- Though doesn't imply play is stabilizing.

What are internal/swap regret and correlated equilibria?

More general forms of regret
1. "best expert" or "external" regret:
   - Given n strategies. Compete with best of them in hindsight.
2. "sleeping expert" or "regret with time-intervals":
   - Given n strategies, k properties. Let S_i be set of days satisfying property i (might overlap). Want to simultaneously achieve low regret over each S_i.
3. "internal" or "swap" regret: like (2), except that S_i = set of days in which we chose strategy i.

Internal/swap-regret
- E.g., each day we pick one stock to buy shares in.
  - Don't want to have regret of the form "every time I bought IBM, I should have bought Microsoft instead".
- Formally, swap regret is wrt optimal function f:{1,...,n}→{1,...,n} such that every time you played action j, it plays f(j).

Weird... why care?
"Correlated equilibrium"
- Distribution over entries in matrix, such that if a trusted party chooses one at random and tells you your part, you have no incentive to deviate.
- E.g., Shapley game.

\[
\begin{matrix}
R & P & S \\
R & -1,-1 & -1,1 & 1,-1 \\
P & 1,-1 & -1,1 & -1,1 \\
S & -1,1 & 1,-1 & -1,1 \\
\end{matrix}
\]

In general-sum games, if all players have low swap-regret, then empirical distribution of play is apx correlated equilibrium.

Connection
- If all parties run a low swap regret algorithm, then empirical distribution of play is an apx correlated equilibrium.
  - Correlator chooses random time t ∈ {1,2,...,T}.
  - Tells each player to play the action j they played in time t (but does not reveal value of t).
  - Expected incentive to deviate: \(\Sigma_j Pr(j)(\text{Regret}|j)\) = swap-regret of algorithm
  - So, this suggests correlated equilibria may be natural things to see in multi-agent systems where individuals are optimizing for themselves

Correlated vs Coarse-correlated Eg
In both cases: a distribution over entries in the matrix. Think of a third party choosing from this distr and telling you your part as "advice".

"Correlated equilibrium"
- You have no incentive to deviate, even after seeing what the advice is.

"Coarse-Correlated equilibrium"
- If only choice is to see and follow, or not to see at all, would prefer the former.

Low external-regret ⇒ apx coarse correlated equilib.
Algorithms for achieving low regret of this form:
- Foster & Vohra, Hart & Mas-Colell, Fudenberg & Levine.
- Will present method of [BM05] showing how to convert any "best expert" algorithm into one achieving low swap regret.
- Unfortunately, #steps to achieve low swap regret is $O(n \log n)$ rather than $O(\log n)$.

Can convert any "best expert" algorithm $A$ into one achieving low swap regret. Idea:
- Instantiate one copy $A_j$ responsible for expected regret over times we play $j$.
- Allow us to view $p_j$ as prob we play $j$, or as prob we play alg $A_j$.
- $A_j$ guarantees \( \sum_t (p_j/c^t)q^t \leq \min_i \sum_t p_j/c^t + \text{[regret term]} \)
- Write as: \( \sum_t p_j(q^t/c^t) \leq \min_i \sum_t p_j/c^t + \text{[regret term]} \)

Variables are $p_{ij}$.
- Constraints for each row $i$.
  - For all $i'$, $\sum_j (p_{ij}/p_i) R_{ij} \geq \sum_j (p_{ij}/p_i) R_{i'j}$
  - Make linear by multiplying LHS,RHS by $p_i$.
- Constraints for each column $j$.
  - Similarly for column player.
- This is for 2-player games. In $m$-player games it's trickier but can use Ellipsoid alg.
- Or, just run a swap-regret-minimizing alg for each player to get an $\varepsilon$-CE.