10-806 Foundations of Machine Learning and Data Science

Lecturer: Avrim Blum 10/21/15

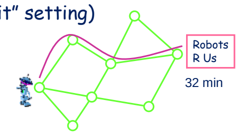
The Adversarial Multi-armed Bandit Problem

(2nd-half of lecture)

Plan for second-half of lecture

Online optimization / combining expert advice but:

- What if we only get feedback for the action we chose? (called the "multi-armed bandit" setting)



- Can we still achieve good regret bounds?
- But first, a quick discussion of [0,1] vs {0,1} costs for RWM algorithm

[0,1] costs vs {0,1} costs.

We analyzed Randomized Wtd Majority for case that all costs in {0,1} (and slightly hand-waved extension to [0,1])

Here is an alternative simple way to extend to [0,1].

- Given cost vector c, view c_i as bias of coin. Flip to create vector c' $\in \{0,1\}^n$, s.t. $E[c'_i] = c_i$. Feed c' to alg A.

world —c→ \$ —c'→ A

- For any sequence of vectors c' $\in \{0,1\}^n$, we have:
  - E_A[cost'(A)] $\leq$ $\min_i$ cost'(i) + [regret term] (Cost' = cost on c' vectors)
- So, $E_\$[E_A$[cost'(A)]] $\leq$ $E_\$[\min_i$ cost'(i)] + [regret term]
- LHS is E_A[cost(A)]. (since $E_\$[E_A[cost'(A)]] = E_\$[c' \cdot \vec{p}] = c \cdot \vec{p}$)
- RHS $\leq$ $\min_i E_\$$[cost'(i)] + [r.t.] = $\min_i$[cost(i)] + [r.t.]

In other words, costs between 0 and 1 just make the problem easier…

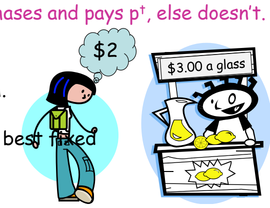
Experts → Bandit setting

- In the bandit setting, only get feedback for the action we choose. Still want to compete with best action in hindsight.
- [ACFS02] give algorithm with expected cumulative regret O((TN log N)$^{1/2}$).
  [average per-day 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 variety, will talk about it in the context of gains instead of losses.

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,…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.

View each possible price as a different action/expert



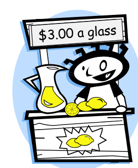
Online pricing

If v^t revealed, run RWM. E[gain] $\geq$ OPT(1-ϵ) - O(ϵ^{-1} h log n).

(algo scales gains to [0,1]; gets E[gain] $\geq OPT(1-\epsilon) - O(\epsilon^{-1} \log n)$ in the scaled world, which translates to above bound in the original world; i.e., by reduction)

- Buyer arrives with valuation v^t
- If $v^t \geq p^t$, buyer purchases and pays p^t, else doesn't.
- Repeat.

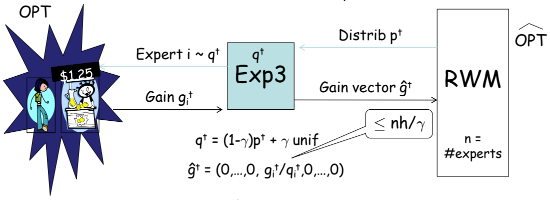
- Assume all valuations $\leq$ h.
- Goal: do nearly as well as best fixed price in hindsight.



Multi-armed bandit problem

Exponential Weights for Exploration and Exploitation (exp^3)

[Auer,Cesa-Bianchi,Freund,Schapire]

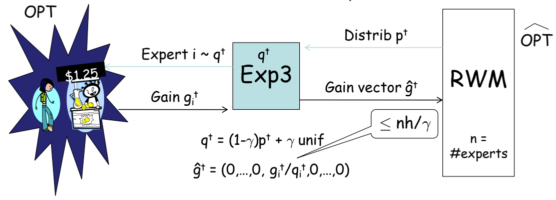


1. RWM believes gain is: $p^t \cdot \hat{g}^t = p_i^t(g_i^t/q_i^t) \equiv g^t_{RWM}$
2. $\sum_t g^t_{RWM} \geq \widehat{OPT}(1-\epsilon) - O(\epsilon^{-1}\, nh/\gamma \log n)$
3. Actual gain is: $g_i^t = g^t_{RWM}(q_i^t/p_i^t) \geq g^t_{RWM}(1-\gamma)$
4. $E[\widehat{OPT}] \geq OPT$. Because $E[\hat{g}_j^t] = (1-q_j^t)0 + q_j^t(g_j^t/q_j^t) = g_j^t$,
   so $E[\max_j[\sum_t \hat{g}_j^t]] \geq \max_j [E[\sum_t \hat{g}_j^t]] = OPT$.

Multi-armed bandit problem

Exponential Weights for Exploration and Exploitation (exp^3)

[Auer,Cesa-Bianchi,Freund,Schapire]



Conclusion ($\gamma = \epsilon$):

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

Balancing would give $O((OPT\ nh \log n)^{2/3})$ in bound because of ϵ^{-2}.
But can reduce to ϵ^{-1} and $O((OPT\ nh \log n)^{1/2})$ with better analysis.