How to judge policy performance?

Discussion on analysis methods

Our setting

- Multi-armed bandit problem
  - n “arms” (=actions, $A = \{a_1..a_n\}$)
  - At each time step t the agent chooses an action
    - Or a distribution over actions for step t, p_t
  - The chosen action yields some reward
    - Or expected reward $\sum_{i=1}^{n} p_t^i r_t^i$
    - (No significant difference between losses and rewards)
- How would you judge how well an agent is doing?

Example

- Very large action space (n actions)
- A single optimal action a^* with reward r
- A small subset of actions $\left|A_{suboptimal}\right| = m$, $m \ll n$, with reward $(1-\epsilon)r$, $0 < \epsilon \ll 1$.
- All the other actions yield a reward of 0.

- How should we judge our policy?

Example

- Very large action space (n actions)
- A single optimal action a^* with reward r
- A small subset of actions $\left|A_{suboptimal}\right| = m$, $m \ll n$, with reward $(1 - \epsilon)r$, $0 < \epsilon \ll 1$.
- All the other actions yield a reward of 0.

- How should we judge our policy?
  - Depends on what we want!
  - Optimal policy? Accumulated reward?
  - Asymptotic or bounded? Etc…

Non-stationary case

- What does optimality mean in the non-stationary case?
- What do we need to assume in order for our policy (or any policy) to be effective?

Non-stationary case

- What does optimality mean in the non-stationary case?
- What do we need to assume in order for our policy (or any policy) to be effective?

- Are we still making some hidden assumptions?

Non-stationary case

- What does optimality mean in the non-stationary case?
- What do we need to assume in order for our policy (or any policy) to be effective?

- Our we still making some hidden assumptions?
  - We assume the rewards for the actions are independent…
  - Does that assumption always hold?

Our setting, revisited

- It would be nice if we didn't need to assume *anything* about the reward distributions per action.
- Can we still get some concrete guarantees?

Adversarial model

- At each time step t, our agent chooses an action a_t.
- At that point, an adversary, which has full control over the environment, chooses how to assign the reward vector for all the actions.
  - Think of it as "*non-stationary with malice*"…
- The agent sees the reward it received for a_t.
- How can we judge performance now? Can we still simply consider accumulated reward?

Regret I

- Can't compare to the series of optimal actions (*why?*).
- Instead, let's compare ourselves to the *best single action* we could have stuck with the entire run of $t = 1..T$.
- Let our performance be $A = \sum_{t=1}^{T} \sum_{i=1}^{n} p_t^i r_t^i$
- Let $r_{1..T}^i = \sum_{t=1}^{T} r_t^i$, then:

$$r_{1..T}^{best} = \max_i \{ r_{1..T}^i \}$$

- We define:

$$regret = \min\left\{ r_{1..T}^{best} - A, 0 \right\}$$ (*why do need the "min"?*)

- This is called *external* regret.

Regret Example I

- 6 actions, 6 time steps:

Time	a_1	a_2	a_3	a_4	a_5	a_6
$t = 1$	0	0	0	0	0	1
$t = 2$	0	1	0	0	0	0
$t = 3$	0	0	1	0	0	0
$t = 4$	0	0	0	0	1	0
$t = 5$	0	0	0	1	0	0
$t = 6$	1	0	0	0	0	0

- Our series of actions:
  - $a_1 \rightarrow a_5 \rightarrow a_3 \rightarrow a_3 \rightarrow a_3 \rightarrow a_6$

Regret Example I

- 6 actions, 6 time steps:

Time	a_1	a_2	a_3	a_4	a_5	a_6
$t = 1$	0	0	0	0	0	1
$t = 2$	0	1	0	0	0	0
$t = 3$	0	0	1	0	0	0
$t = 4$	0	0	0	0	1	0
$t = 5$	0	0	0	1	0	0
$t = 6$	1	0	0	0	0	0

- Maximal possible reward – 6
- Regret? None. (why?)

Regret Example II

- 6 actions, 6 time steps:

Time	a_1	a_2	a_3	a_4	a_5	a_6
$t = 1$	0	0	0	0	0	1
$t = 2$	1	1	0	0	0	0
$t = 3$	1	0	1	0	0	0
$t = 4$	1	0	0	0	1	0
$t = 5$	0	0	0	1	0	0
$t = 6$	1	0	0	0	0	0

- Our series of actions:
  - $a_1 \rightarrow a_5 \rightarrow a_3 \rightarrow a_3 \rightarrow a_3 \rightarrow a_6$

Regret Example II

- 6 actions, 6 time steps:

Time	a_1	a_2	a_3	a_4	a_5	a_6
$t = 1$	0	0	0	0	0	1
$t = 2$	1	1	0	0	0	0
$t = 3$	1	0	1	0	0	0
$t = 4$	1	0	0	0	1	0
$t = 5$	0	0	0	1	0	0
$t = 6$	1	0	0	0	0	0

- Maximal possible reward – still 6
- Regret? Aplenty! (3, to be exact)

Regret II

- Another option: what if we compared ourselves to a small modification of our own policy?
- For instance, “every time you took action i, you should have actually taken action j”.
- This is the idea behind internal regret.
- Can be extended to “swap regret” (full mapping from actions to actions).
- Other notions exist (tracking regret, for instance, which reflects competitive analysis).

Summary and discussion

- How to compare performance in n-armed bandit settings?
- What are our assumptions?

- Stochastic vs. adversarial
- Regret

- Questions?
- Thank you!

References

- Class notes, "*Computational Game Theory*", taught by Prof. Yishay Mansour, TAU, Spring 2010.
- Class notes, "*Online Algorithms*", taught by Prof. Yossi Azar, TAU, Fall 2009.
- Nisan et al., "*Algorithmic Game Theory*", Cambridge Uni. Press 2007, chapter 4, "*Learning, Regret Minimization, and Equilibria*", A. Blum and Y. Mansour.
- Cesa-Bianchi and Lugosi, "*Prediction, Learning, and Games*", Cambridge Uni. Press 2006.