Dan Schwartz

thedanielschwartz

August 2019

1 Stochastic Bandits

1.1 Process

- Collection of distributions
- Learner and environment interact sequentially over n rounds
- Learner chooses action, environment samples reward and reveals to learner

1.2 Learning Objective

- Learner maximizes reward
- Cumulative reward is random quantity
- Learner doesn't know distributions

2 Stochastic Bandits with Finitely Many Arms

- Number of actions available is finite
- One action has no means on payoff of other arms
- Sequence of rewards associated with each action is I.I.D.

2.1 Explore-then-Commit Algorithm

- Explores by playing each arm a fixed number of times then exploits committing tto arm that appeared best during exploration

2.2 Upper Confidence Bound Algorithm

- Optimimism Principle
  - One should act as if the environment is as nice as plauisbly possible

3 Adversarial Bandits with Finitely Many Arms

- Adversarial bandit abandons all assumptions on how rewards are generated
- Adversary can examine algorithm and choose rewards accordingly

3.1 Exp3 Algorithm

3.1.1 Exponential-weight algorithm for Exploration and Exploitation

- k-armed adversarial bandit
- Exponential weighting
  - Large learning rate → concentrates arm with largest estimated reward and algorithm exploits aggressively
  - Small learning rate → explores more frequently

3.2 Exp3-IX Algorithm

3.2.1 Exponential-weight algorithm for Exploration and Exploitation Implicit Exploration

- Keep regret small and concentrated about its mean
- Since small losses correspond to large rewards, estimator is optimistically biased
- Exp3-IX explores more than standard Exp3
- Consequence of modifying loss estimates than directly altering P_t

4 Contextual and Linear Bandits