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

4.1 Contextual Bandits

4.1.1 One bandit per context

- Adversary secretly chooses rewards
- Adversary secretly chooses contexts
- Learner observes context
- Learner selects distribution
- Learner observes reward

4.1.2 Bandits with expert advice

- Use when context set is large and unstructured
- Measure similarity between pairs of contexts
- Adversary secretly chooses rewards
- Experts secretly choose predictions
- Learner observes predictions
- Learner selects distribution
- Action is sampled from distribution and reward

4.1.3 Exp4 (Exponential Weighting for Exploration and Exploitation with Experts

- Scores experts instead of actions (like in Exp3)

4.2 Stochastic Linear Bandits

4.2.1 Stochastic Linear Bandit

- Reward is assumed to have linear structure
- Allows learning to transfer from one context to another

4.2.2 Stochastic Contextual Bandits

- Same as adversarial contextual bandit, but reward function has 1-subgaussian noise

4.2.3 Stochastic Linear Bandits with Finitely Many Arms

- Choose each action in $a \in A$ $T_l(a)$ times
- Calculate empirical estimate
- Eliminate low rewarding arms

4.3 Stochastic Linear Bandits with Sparsity

- Similar to PCA

4.3.1 Sparse Linear Stochastic Bandits