Online Learning – EE 381V / CS 395T

Overview: Multi-armed Bandits.

Textbook: Bandit Algorithms, T. Lattimore and C. Czepesvári, Cambridge Univ. Press, (to be published), 2019.

① Problem setting: explore vs. exploit

→ Online advertising
→ Recommendation systems
→ Clinical trials
→ Online resource allocation
→ Online search (e.g. MCTS)

② Setup:

$A_1, X_1, A_2, X_2, \ldots$

A_t: action, X_t: reward

→ time.

Key assumptions:

1. Reward observed only corresponding to the action taken. (bandit feedback)
2. Action does not alter the environment
3. Taking an action does not restrict future actions

→ Distinguishes bandits from RL

Example: Two biased coins.

H

$P(H) = 0.6$

$P(T) = 0.4$

H

$P(H) = 0.75$

$P(T) = 0.25$