



DS 102: Data, Inference, and Decisions

Lecture 16

Michael Jordan

University of California, Berkeley

Decisions and Learning

- In our lectures on FDR, we talked about making sets of decisions
 - and in our lectures on online FDR, we talked about making sets of decisions over time
 - the sets of decisions can be unrelated
- We now wish to consider problems in which we make the same decision over and over again, and we want to get better at making that decision
 - let's call this a **learning** problem
 - learning involves both **exploration** and **exploitation**

Exploration and Exploitation

- Let's suppose that we obtain **rewards** based on our decisions
 - good decisions mean high rewards, and bad decisions mean low rewards
 - sometimes we refer to **losses** instead of rewards, converting between the two with negation
- We want to obtain a high rate of reward over time
 - in doing so we aim to **exploit** our knowledge as it accrues
- But no one is telling us which choice yields highest reward
 - we have to **explore** to figure that out

Exploration and Exploitation

- Too little exploration means that the learner may not discover which choice yields the highest reward
 - they stick with a suboptimal choice, and they will not be as happy or successful as they could be over the long run
 - but too much exploration means forgoing the opportunity to reap immediate rewards in the possibly vain hope of future rewards
- So, at any given moment there is a **tradeoff**---should I exploit what knowledge I currently have, or explore further to improve my knowledge?
- Can we devise a formal theory that makes that tradeoff explicit, and actionable?

The Multi-Armed Bandit

- We consider a decision-maker who is given K options to choose from
 - we refer to those options as “arms”
 - please Google “multi-armed bandit” to see a picture of a casino, from whence the language comes
- Associated with each arm is a probability distribution over rewards
- The decision-maker chooses an arm, and receives a reward sampled from the corresponding reward distribution
- This repeats

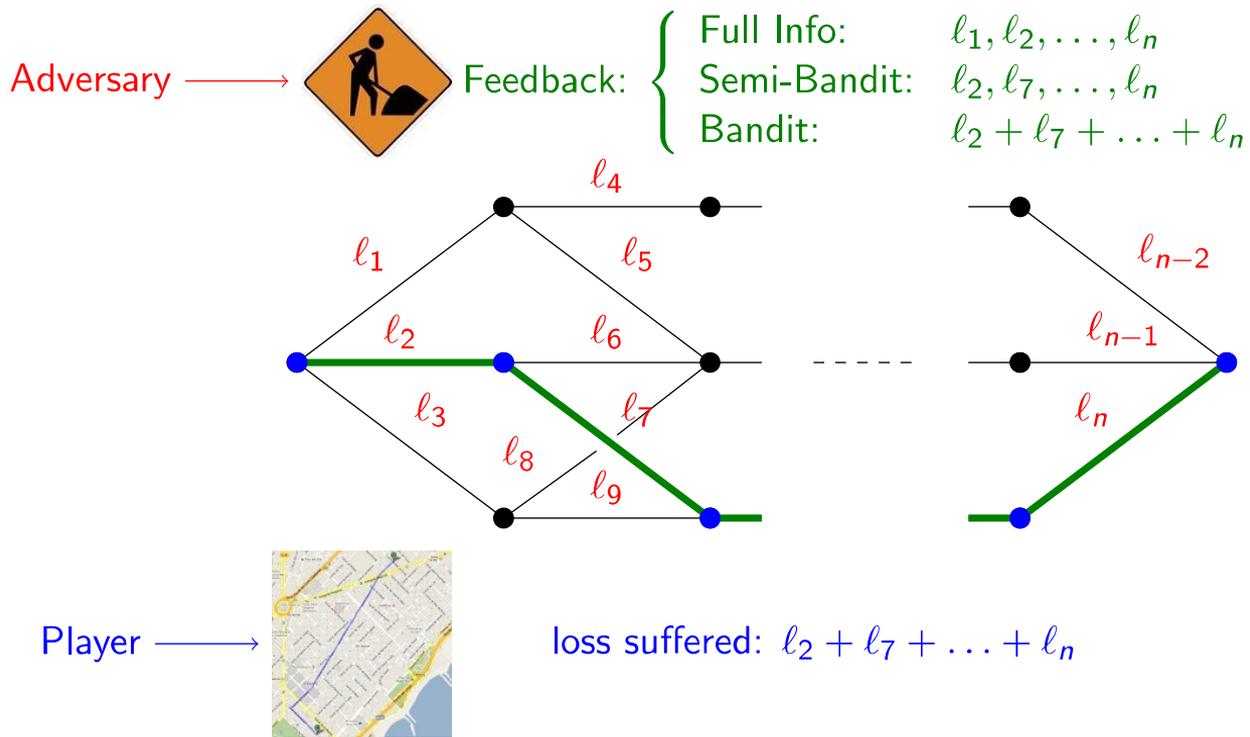
The Goal

- Generally the goal is to maximize the sum of the rewards over all time steps
- Sometimes the rewards are **discounted**, when rewards in the future are viewed as less valuable than more immediate rewards
- A key idea is **regret**, which is the difference between our total reward and the reward of some kind of **oracle** who knows more than we do
 - for example, an oracle who knows in advance which is the optimal arm

Variations on the basic multi-armed bandit

- The environment can be **stochastic** or **adversarial**
- **Contextual bandits** blend bandits with regression
- There are **structured** bandits, in which the actions and the rewards have mathematical structure that can be exploited
 - see the example of **path planning** on the following page
 - in such settings, the **feedback** can be more detailed than simply the reward associated with the selected arm

Example: path planning



Real-World Problems

- Drug discovery
- A/B testing
- Routing in networks
- Robot skills
- Human-computer interaction

To the Whiteboard and to Jupyter...

Comparing UCB to Explore-then-Commit

