EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 01 - 01/19/2021

Lecture 01: Introduction to Bandits and Reinforcement Learning

Lecturer: Jiantao Jiao *Scribe: Nima Hejazi, Wenxin Zhang*

In this lecture, we begin with a general overview of the two problem structures with which this course is concerned: (1) multi-armed bandits and (2) reinforcement learning. From this overview, we proceed to formalize the multi-armed bandits problem.

1 Problem Settings

1.1 Multi-armed Bandits

Note: in the bandits problem, there is no concept of state (S), instead, the learning algorithm observes only sequences of action (A) and reward (R) pairs. Bandits literature cover the following topics:

- **Types:** stochastic, contextual, adversarial
- "Best of both worlds": a phenomenon that the same algorithm is used for stochastic and adversarial bandits, which achieves the best performance when either assumption is true
- Tradeoff between exploration and exploitation: continue pulling known arm (exploitation) or attempt to learn about unknown arms (exploration)
- State-of-the-art algorithms: UCB (upper confidence bound), EXP4, Thompson sampling

1.2 Reinforcement Learning

Problem structure is a sequence of Markov chains — formally, Markov decision processes (MDPs) — and may be viewed as a generalization of the bandits setting. In particular, consider that the learning algorithm observes the following

$$(S_t \to A_t) \to (S_{t+1} \to A_{t+1}),$$

where t denotes the index of the *current* state (a summary of the current full information) and action while $t+1$ denotes the next state and action. In between the repeating series of (state, action) pairs, a reward based on the the state and action is observed $r(S_t, A_t)$. At each subsequent time, the selected action A_t is chosen based on the current state S_t (and possibly past states) and the most recently observed reward.

2 Formalizing the Bandits Problem

Consider a news site, where a new user arrives and an article to be displayed must be chosen; further, let the reward be binary, i.e., $R = 1$ if the user clicks on the article and $R = 0$ if not. Let there be K actions and T decision rounds — by what strategies may the exploration–exploitation tradeoff be optimized for our goal (e.g. maximizing the number of clicks)?

Next, we turn to discussing the bandits problem in terms of two distinct frameworks, based on the available information about actions and rewards, respectively.

2.1 Feedback Model

Several types of information feedback may be revealed when selecting an action at any given decision point of a bandits problem. In order of increasing information, these are

- Bandit feedback: only the reward corresponding to the selected (predicted) action is revealed.
- Partial feedback: some additional information about the rewards linked to unchosen arms is revealed. For example, in *dynamic pricing*, a sale at price X implies that a sale would have been made at all prices $\leq X$, i.e., a monotonicity of rewards corresponding to chosen actions is revealed.
- Full (causal) feedback: rewards linked to *all arms* that could have been chosen are revealed; this is analogous to observing all of the "potential outcomes" that would be realized from executing available actions. Note that in this setting, there is no missing information.

2.2 Reward Model

The reward observed when pulling a particular bandit arm (i.e., selecting an action) may be considered to be (un)structured in a variety of ways. Several common reward model variations include

- Stochastic reward: i.i.d. reward, such that arm a has a fixed corresponding reward distribution P_a (if arm a is pulled). For each pull of arm a, a reward r is drawn independently from distribution P_a, which does not change with time t. This may be *generalized* such that the reward r_t is not simply an i.i.d. draw but a realization of a stochastic process (*Generalized Stochastic Reward*).
- Adversarial reward: arbitrary reward, where the observed reward for a given arm pull is chosen by an adversary aiming to fool the learning algorithm.
- Constrained reward: similar to adversarial, tweaked to have constraints. Such constraints may enforce, e.g., that the reward of each arm not changing too much from one round to the next, or limit the total change via a fixed budget.
- Structured reward: a model for the reward holds, i.e., let $r(a) = \theta^T \phi(a)$. ϕ is a known function that maps the selected action to a d-dimensional vector in $\mathbb{R}^d$. $r(a) = f(\phi(a))$ is a (possibly linear) function of actions (a), with $f(\cdot)$ being convex, Lipschitz, or smooth, and unknown (to be learned) parameters θ^T. For example, returning to the familiar problem of displaying a news article to a website visitor, we may let $\phi(a) = [\text{topic}, \text{word}, \ldots]$ where a is a selected article.

A more complex, but nonetheless quite interesting, extension to the "usual" bandits problem setup is the *contextual* bandit settings. In the contextual bandits setting, a reward is not determined only by the chosen action a but also by the observed context c (or s). Building upon the aforementioned, let $P_{c,a}$ denote the distribution of (context, action) pairs. Here, the goal is to learn a policy mapping context to actions, i.e., $\pi : \mathcal{C} \to \mathcal{A}$ for $c \in \mathcal{C}$ and $a \in \mathcal{A}$. In this setting, the structured reward model may be extended under such that $r(c, a) = \theta^T \phi(c, a) \in \mathbb{R}^d$, where the action selection mechanism ϕ and reward are both functions of the context c as well as the action a. In the news-website setting, this may be thought of as choosing an article to display to a website visitor based on context available on the user, where such context may include relevant factors like the user's profile, location and demographics.

Sometimes the action space can be very complicated. We may need to take multiple actions at the same time, which is called *structured actions*. For example, the news website may want to display many articles to the reader simultaneously.

References

[1] A. Slivkins, "Chapter 1, introduction to multi-armed bandits," *arXiv preprint arXiv:1904.07272*, 2019.