Summary

Multi-armed bandits provide a versatile framework for learning in repeated decision-making scenarios. At each discrete timestep, a learner selects an action and observes only the loss associated with that action. The feedback for unchosen actions remains unobserved. The learner's performance is measured by regret, defined as the difference between the learner's cumulative loss and the loss of the best fixed action in hindsight. Bandit algorithms have seen widespread application, including in ad selection, hyperparameter tuning, and pathfinding.

This work explores bandits in four chapters. The first chapter introduces the classical bandit setting, outlining assumptions on actions, losses, and the feedback the learner observes. We discuss importance-weighted loss estimators and highlight the limitations of the Follow-the-Leader (FTL) algorithm, which simply chooses the best action in hindsight at each timestep. We then illustrate Follow-the-Regularized-Leader (FTRL), a well known improvement on FTL that achieves optimal regret bounds in many bandit settings.

In Chapter 2, we consider combinatorial bandits, where the learner plays multiple actions simultaneously. Depending on the feedback model, the learner either observes individual losses (semi-bandit feedback) or their sum (full-bandit feedback). We provide a regret analysis for FTRL in this setting by decomposing the regret to a stability and regularization term, offering a more gentle introduction to the advanced concepts developed later.

Chapter 3 introduces adversarial contextual combinatorial bandits. Contextual bandits extend the bandit setting by allowing the learner to observe a context each timestep. The regret is then augmented to be the difference between the loss of the algorithm and the best context to action mapping in hindsight, allowing for greater granularity. In contextual combinatorial bandits, the learner observes a context before choosing a combinatorial action with the incurred loss being linear in both the chosen action and the context. We introduce novel estimators for both the semi-bandit and full-bandit feedback settings, with the latter requiring the introduction of four-dimensional tensors, that we rigorously define. We prove the first (nearly optimal) regret bounds in this setting and validate our methods empirically on a synthetic dataset.

Chapter 4 addresses delayed feedback, a common challenge in real-world applications where losses are observed only after a delay. We present a new analysis technique that decomposes the stability term of the regret into the standard stability and regularization terms and a novel delay-dependent component, allowing us to isolate the cost of feedback delay. This leads to the first optimal (up to logarithmic factors) regret bounds for combinatorial semi-bandits and Markov Decision Processes (MDPs) with known transitions, and nearly optimal bounds for linear bandits. For MDPs with unknown transitions, our results match the best-known non-delayed regret bounds and achieve optimal delay-dependency. Empirical evaluations in the combinatorial and linear settings further support our theoretical findings.