Stat 6341.001: Numerical Linear Algebra and Statistical Computing, Fall 2026 **Topic 1**

The Multi-Armed Bandit Problem

Lecturer: Yun Wei *Scribe: Xi Tang Lin*

1 Introduction

The multi-armed bandit problem is a sequential decision-making formulation where an agent chooses between A distinct actions (arms) at each time step to maximize cumulative reward (or minimize cumulative regret):

- A available decisions/actions.
- Select an action at each round.
- Receive a stochastic, noisy reward.

2 Multi-Armed Bandit Protocol

For time steps $t = 1, 2, \ldots, T$:

1. Select a decision $\pi^t \in \Pi \coloneqq \{1, 2, \ldots, A\}$.
2. Observe a random reward $r^t \in \mathbb{R}$.

History The interaction history H_{t-1} contains all action-reward pairs observed up to time $t-1$:

$$H_{t-1} \coloneqq \{(a_1, r_1), (a_2, r_2), \ldots, (a_{t-1}, r_{t-1})\}$$

Assumption: Stochastic Reward The reward r_t is a random variable drawn from an underlying distribution conditioned on the interaction history:

$$r_t \sim M^*(\cdot \mid H_{t-1})$$

Reward Function The true expected reward function f^* for selecting action π is defined as:

$$f^*(\pi) \coloneqq \mathbb{E}[r \mid \pi]$$

Optimal Action The optimal decision π^* yields the highest expected reward:

$$\pi^* \coloneqq \arg\max_{\pi \in \Pi} f^*(\pi)$$

Regret The cumulative regret $\mathrm{Reg}(T)$ measures performance loss relative to the optimal action choice:

$$\mathrm{Reg}(T) \coloneqq \sum_{t=1}^{T} f^*(\pi^*) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^*(\pi^t)]$$

Objective The goal is to design an algorithm achieving sub-linear regret, meaning the average regret vanishes asymptotically:

$$\lim_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{T} = 0$$

3 Simple Example: Always Pure Exploitation

Suppose we always select the first decision, $\pi^t = 1$, for all $t = 1, 2, \ldots, T$. The total regret simplifies to:

$$\begin{aligned}\mathrm{Reg}(T) &\coloneqq \sum_{t=1}^{T} f^*(\pi^*) - \sum_{t=1}^{T} f^*(1) \\ &= T\big(f^*(\pi^*) - f^*(1)\big) = \mathcal{O}(T)\end{aligned}$$

This yields linear regret $\mathcal{O}(T)$, failing our sub-linear objective whenever action 1 is suboptimal.

4 Estimating f^* via Empirical Means

Since the true reward distribution f^* is unknown, we estimate it using the empirical average of past rewards:

$$\hat{f}^t(\pi) \coloneqq \frac{1}{n^t(\pi)} \sum_{s<t} r^s \mathbf{1}_{\{\pi^s=\pi\}}$$

where $n^t(\pi) \coloneqq \sum_{s<t} \mathbf{1}_{\{\pi^s=\pi\}}$ denotes the total number of times decision π was selected up to time t.

5 The Greedy Algorithm and Exploration

A naive approach is the **Greedy Algorithm**, which chooses the decision maximizing the empirical reward estimate at step t:

$$\hat{\pi}^t = \arg\max_{\pi\in\Pi} \hat{f}^t(\pi)$$

While greedy selection exploits current knowledge, it lacks a mechanism for exploration. Because feedback is partial (rewards are observed only for chosen actions), the algorithm can prematurely abandon optimal actions due to early noisy estimates. Once an optimal action's empirical reward drops below another arm's estimate, it is never selected again, leading to linear regret $\mathcal{O}(T)$.

6 Failure of the Greedy Algorithm

Consider a two-armed bandit setting with $\Pi = \{1, 2\}$:

- Decision 1 yields a deterministic reward: $f^*(1) = \frac{1}{2}$.
- Decision 2 yields a stochastic reward: $r \sim \mathrm{Ber}\left(\frac{3}{4}\right)$, so $f^*(2) = \frac{3}{4}$.

Consider the following trajectory:

1. At $t = 1$, choose Decision 1: receive $r^1 = \frac{1}{2}$. Thus, $\hat{f}^2(1) = \frac{1}{2}$.
2. At $t = 2$, choose Decision 2: receive $r^2 = 0$ (occurring with probability $\frac{1}{4}$). Thus, $\hat{f}^3(2) = 0$.
3. For all $t \geq 3$, the greedy strategy continuously chooses Decision 1 because $\hat{f}^t(1) = \frac{1}{2} > 0 = \hat{f}^t(2)$.

Under this event, the expected regret grows linearly:

$$\mathrm{Reg}(T) = \left(\frac{3}{4} - \frac{1}{2}\right) + \left(\frac{3}{4} - \frac{3}{4}\right) + (T-2)\left(\frac{3}{4} - \frac{1}{2}\right) = \frac{T-1}{4} = \mathcal{O}(T)$$

7 ϵ-Greedy Algorithm

To prevent premature convergence to suboptimal choices, the ϵ-greedy algorithm forces exploration:

Exploitation With probability $1-\epsilon$, exploit current knowledge by picking the empirically optimal action:

$$\hat{\pi}^t = \arg\max_{\pi \in \Pi} \hat{f}^t(\pi)$$

Exploration With probability ϵ, explore uniformly at random across all available actions:

$$\hat{\pi}^t \sim \text{Unif}(\{1, 2, \ldots, A\})$$

8 Sub-Gaussian Concentration Bounds

Definition 8.1 (σ-sub-Gaussian Variable)**.** A zero-mean random variable Z is σ-sub-Gaussian if its Moment Generating Function (MGF) satisfies:

$$\mathbb{E}\left[e^{\eta Z}\right] \leq \exp\left(\frac{\sigma^2\eta^2}{2}\right), \quad \forall \eta \in \mathbb{R}$$

Lemma 8.2 (Sub-Gaussian Hoeffding Bound)**.** *Let $Z_1, \ldots, Z_T$ be independent, zero-mean σ-sub-Gaussian random variables. Then for any $u > 0$:*

$$\mathbb{P}\left(\frac{1}{T}\sum_{i=1}^{T} Z_i \geq u\right) \leq \exp\left(-\frac{Tu^2}{2\sigma^2}\right)$$

References

[1] Foster, Dylan J., and Alexander Rakhlin. Foundations of Reinforcement Learning and Interactive Decision Making. arXiv, 2023, https://doi.org/10.48550/arXiv.2312.16730.