Stochastic Linear Bandits with Parameter Noise

Daniel Ezer[1], Dr. Alon Peled-Cohen[1,2] & Prof. Yishay Mansour[1,2]

[1] Tel Aviv University

[2] Google Research

Multi Armed Bandits

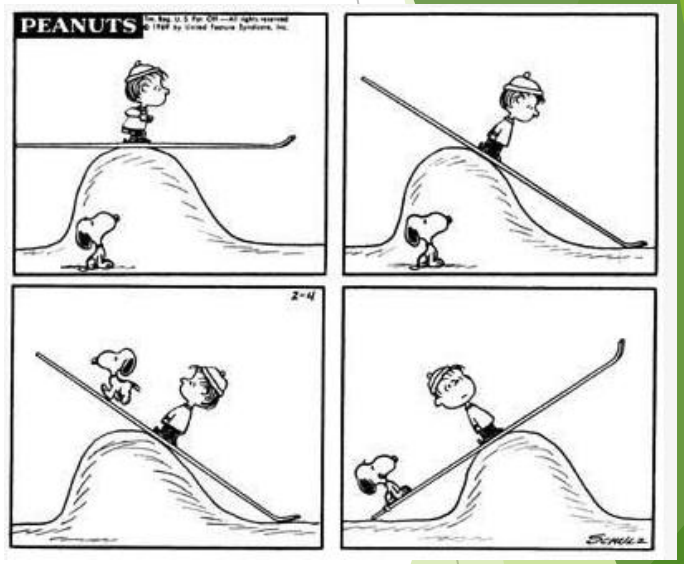
- For each $t = 1, \ldots, T$:
  - Learner chooses action $a_t \in A$
  - Learner observes reward $r_t(a_t)$
- Goal:
  - Minimize regret:

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^{T} r_t(a^*) - r_t(a_t)\right]$$

  - Where a^* is an optimal action

Stochastic Linear Bandits

- $A = \mathbb{R}^d$
- $\mathbb{E}[r_t(a)] = a^T\theta^*$ for unknown $\theta^* \in \mathbb{R}^d$
- Generalizes MAB with $A = \{e_1, \ldots, e_d\}$

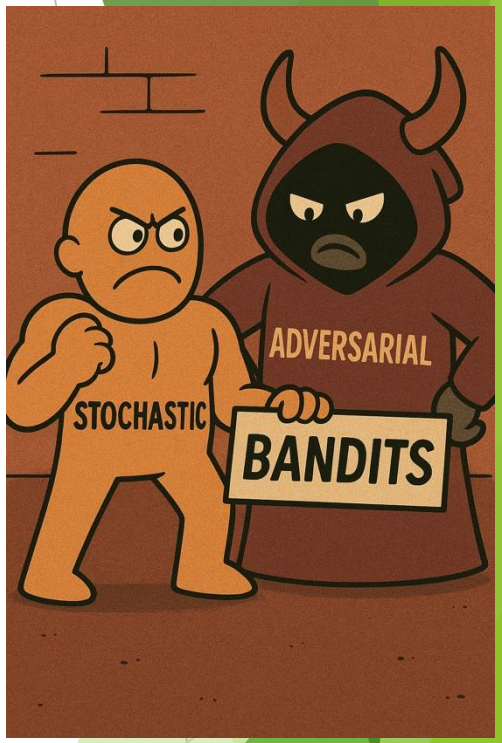


Adversarial Linear Bandits

- Reward vectors chosen by an adversary
- Formally:
  - $r_t(a_t) = a_t^\top \theta_t$
  - Where θ_t are arbitrary vectors
- This should be harder than stochastic

”Contradiction” between common models

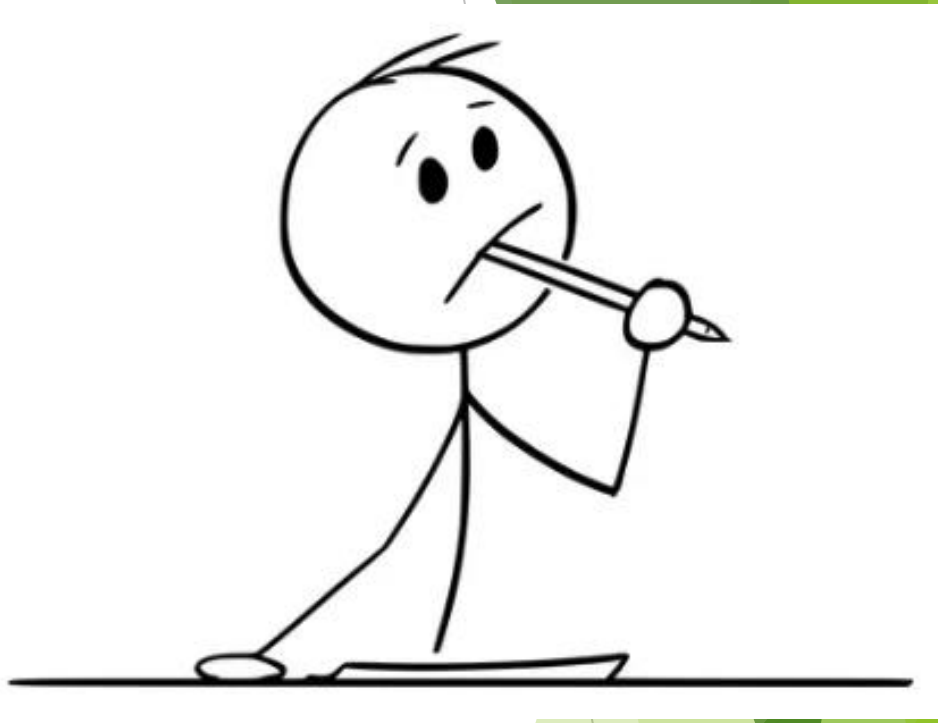
- Common stochastic model:
  - $r_t(a) = a^T\theta^* + \eta_t$
  - Where η_t is zero-mean and 1-Sub-Gaussian.
- Stochastic can be “harder” than adversarial!
- For $A = \ell_2$ unit ball:
  - Stochastic[1]: $R_T = \Omega\left(d\sqrt{T}\right)$
  - Adversarial[2]: $R_T = \tilde{O}\left(\sqrt{dT}\right)$



1. Lattimore, Tor, and Csaba Szepesvári. Bandit Algorithms. Cambridge: Cambridge University Press, 2020. Print.
2. Bubeck, Sébastien et al. “Towards Minimax Policies for Online Linear Optimization with Bandit Feedback.” Annual Conference Computational Learning Theory (2012).

Contradiction Intuition

- Stochastic model is not really linear!
- $r_t(a) = a^T \theta^* + \eta_t$
- Expectation is linear, realization is not!



Stochastic Linear Bandits with Parameter Noise

- In timestep t:
  - Sample $\theta_t \sim \nu$ with $\mathbb{E}[\nu] = \theta^*$ and covariance matrix Σ
  - Learner observes $r_t(a) = a^T \theta_t$
- Trivial reduction to adversarial

Main Results (Partial)

- For ℓ_p unit balls with $1 < p \leq 2$, the minimax regret is $\widetilde{\Theta}\left(\sqrt{dT\sigma_q^2}\right)$, where $\sigma_q^2 = \left(\sum_{i=1}^{d} \Sigma_{ii}^{\frac{q}{2}}\right)^{\frac{2}{q}}$
  - With a very simple explore-exploit algorithm
- Additional results for general finite action sets

Our Explore-Exploit Algorithm - ℓ_p Ball

- Explore:
  - Play each e_i for $\frac{1}{\varepsilon^2}$ times
  - Calculate estimator $\hat{\theta}$
- Exploit:
  - Play $\hat{a} = \frac{\hat{\theta}}{\|\hat{\theta}\|_q}$ for the rest of the timesteps

Thank You For Listening!