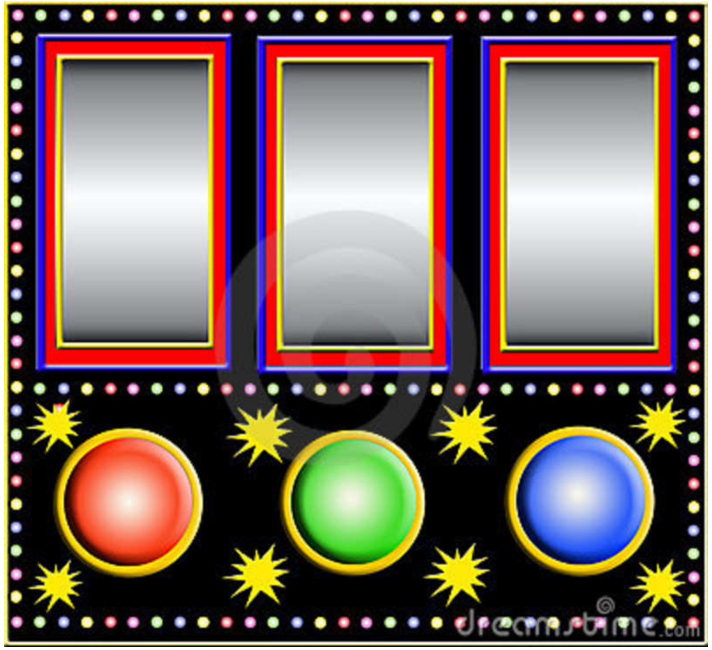
Introduction to Bandit Theory



Xiaoke Wang

Outline

- Introduction and Examples
- Stochastic Bandits
- Adversarial Bandits
- Extensions
- References

Introduction and Examples

- Ad placement
  - Arms: ads pool {1,2,…,K}
  - Action: decide which advertisement (X_n) to display to the next visitor
  - Bandit feedback: only the reward of displaying X_n is observed ($r_n(X_n)$)
  - Performance metric: regret (compare with an optimal strategy that consistently displays the most popular ad

$$R(T) = \underbrace{\max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{n=1}^{T} r_n(x)\right]}_{\text{oracle}} - \underbrace{\mathbb{E}\left[\sum_{n=1}^{T} r_n(x_n)\right]}_{\text{your algorithm}}.$$

Stochastic Bandits

- Setting
  - Reward $r_n(i)$ is drawn from an unknown distribution $V(i)$
  - Let $u(i)$ denote the expectation of $r_n(i)$
  - Let $u^*= \max u(i)$ and $i^* = \operatorname{argmax} u(i)$
- Optimal solution
  - always choose i^*
- Challenge
  - Reward is drawn from a distribution
  - Unaware of the value of $u(i)$

Stochastic Bandits: Upper Confidence Bound (UCB) Strategies

Assume rn(i) ϵ [0,1],

Algorithm:

At time t, select

$$x_t = \operatorname*{argmax}_{i=1,2,\ldots,K} [\hat{u}_{i,T_{i(t-1)}} + 2(\frac{\alpha \ln n}{T_{i(t-1)}})^2]$$

Intuition:

Based on Markov's inequality, with probability at least 1-δ,

$$\hat{u}_{i,s} + 2(\frac{1}{s} \ln \frac{1}{\delta})^2 > u_i$$

Regret:

$$O(K \ln T)$$

Stochastic Bandits: Thompson Sampling

Algorithm 1: Thompson Sampling for Bernoulli bandits

$S_i = 0, F_i = 0.$

foreach $t = 1, 2, \ldots,$ **do**

 For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ from the Beta$(S_i + 1, F_i + 1)$ distribution.

 Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward r_t.

 If $r = 1$, then $S_i = S_i + 1$, else $F_i = F_i + 1$.

end

Regret: $O(K^2 \ln T)$

Adversarial Bandits

- Setting
  - Reward $r_n(i)$ is bounded ($r_n(i) \in [0,1]$) and chosen by adversary
  - Oblivious adversary
- Optimal solution
  - always choose $i^* = \underset{i=1,2,\ldots,K}{\operatorname{argmax}} \sum_{t=1}^{n} r_t(i)$
- Regret

$$R_n = \max_{i=1,2,\ldots,K} \sum_{t=1}^{n} r_t(i) - \sum_{t=1}^{n} r_t(x_t)$$

- Challenge
  - No statistic information available
  - Unaware of the reward of other choice
  - Can we achieve sublinear bounds on the regret?

If $X_t=1$, then $r_t(2)=1$
If $X_t \neq 1$, then $r_t(1)=1$

$R_n >= n/2$

Adversarial Bandits: Exp3

Unbiased estimator for the loss of any other arm

$$\mathbb{E}_{I_t \sim p_t}[\widetilde{\ell}_{i,t}] = \sum_{j=1}^{K} p_{j,t} \frac{\ell_{i,t}}{p_{i,t}} \mathbb{1}_{j=i} = \ell_{i,t}.$$

Exp3 (Exponential weights for Exploration and Exploitation)

Parameter: a nonincreasing sequence of real numbers $(\eta_t)_{t \in \mathbb{N}}$.

Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

For each round $t = 1, 2, \ldots, n$

(1) Draw an arm I_t from the probability distribution p_t.
(2) For each arm $i = 1, \ldots, K$ compute the estimated loss $\widetilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} \mathbb{1}_{I_t = i}$ and update the estimated cumulative loss $\widetilde{L}_{i,t} = \widetilde{L}_{i,t-1} + \widetilde{\ell}_{i,s}$.
(3) Compute the new probability distribution over arms $p_{t+1} = (p_{1,t+1}, \ldots, p_{K,t+1})$, where

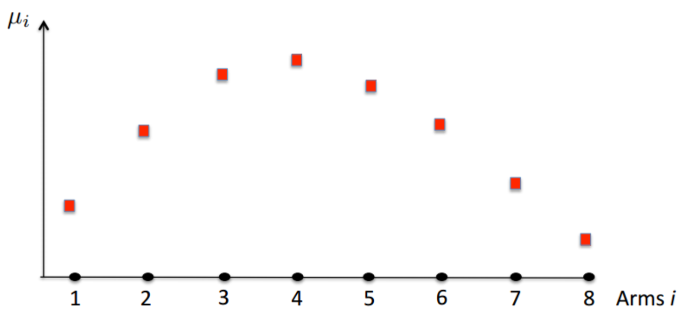
$$p_{i,t+1} = \frac{\exp(-\eta_t \widetilde{L}_{i,t})}{\sum_{k=1}^{K} \exp(-\eta_t \widetilde{L}_{k,t})}.$$

Exponential reweighting

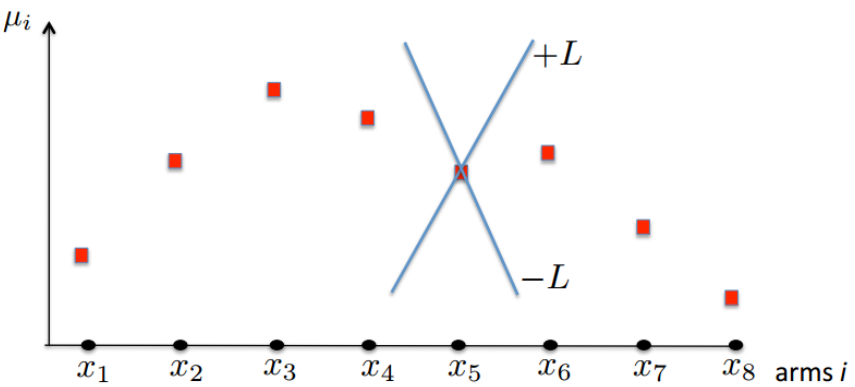
Regret: $O(\sqrt{TK \ln K})$

Extensions

- Structured Bandits



Classical unimodality, graph = line



Lipschitz

- Linear Bandits
  - Instead of i ϵ {1,2,…,K}, we have $i \in \mathbb{R}^d$

References

- Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems.
- http://www.princeton.edu/~sbubeck/tutorial.html