Mittagsseminar

Stochastic Bandits

Sebastian U. Stich

ETH Zürich

May 30, 2013
A multi-armed bandit problem is a sequential resource allocation problem defined by a set of actions. In every round, a unit resource is allocated to one action and some observable payoff is obtained. The goal of the player is to maximize her total payoff obtained in a sequence of rounds. In order to achieve this goal, the player must find the optimal trade-off between playing actions that did well in the past and exploring unknown actions that might give higher payoffs in the future.

In this talk, we will focus on the stochastic version of this problem where the actions are given by a set of probability distributions. Auer, Cesa-Bianchi, Fischer (2002) presented an elegant algorithm (UCB) that tackles the exploration/exploitation trade-off by estimating probabilistic upper bounds on the future performance of each action. These upper confidence bounds (UCB) follow from Chernoff’s inequality.

If time permits, we discuss some variations and applications of this problem.
1 Introduction
   - The Mensa Problem

2 Algorithms
   - Greedy Improved
   - UCB
   - Remarks

3 Optimization
   - Discrete
   - Continuous
Online Learning with Bandit Feedback

Adversary

SV Group
Passion for quality. Since 1914.

Polysnack

Player

Feedback

$\ell_1$, $\ell^*$, $\ell_J$, $\ell_K$

$\ell_J$ (bandit setting)
Stochastic Bandit Setting

Setting:
- Set of $K$ arms, defined by distributions $\nu_k$ (with support $[0, 1]$)
- The $\nu_k$ are not known to the player

For each round $t = 1, \ldots, n$:
1. The player chooses one arm $J_t \in [K]$
2. Receive reward $X_{J_t,t} \sim \nu_{J_t}$

Goal: Play the arm with the highest expected value ‘almost always’
The Regret

- Let $\mu_k$ be the expected value of arm $k$
- Let $\mu^* = \max_k \mu_k$ the best expected value

**Expected regret:**

\[
R_n = \sum_{t=1}^{n} \mu^* - \mathbb{E}[X_{J_t,t}] = \sum_{t=1}^{n} \mu^* - \mu_{J_t} = \sum_{k=1}^{K} \Delta_k \mathbb{E}[T_k(n)]
\]

where $\Delta_k = \mu^* - \mu_k$, and $T_k(s)$ the number of times arm $k$ has been played up to time $s$.

**Goal:** Find an arm selection policy to minimize $R_n$. 

---

The Mensa Problem

Introduction
Algorithms
Optimization

A. The Mensa Problem

B. The Regret

C. Expected regret:

\[
R_n = \sum_{t=1}^{n} \mu^* - \mathbb{E}[X_{J_t,t}] = \sum_{t=1}^{n} \mu^* - \mu_{J_t} = \sum_{k=1}^{K} \Delta_k \mathbb{E}[T_k(n)]
\]

where $\Delta_k = \mu^* - \mu_k$, and $T_k(s)$ the number of times arm $k$ has been played up to time $s$.

**Goal:** Find an arm selection policy to minimize $R_n$. 

---

S. Stich
Bandits
Example: $\epsilon$-greedy

$\epsilon$-greedy: at time $t$

- with probability $(1 - \epsilon)$ play the arm with the highest empirical mean
- with probability $\epsilon$ play a random arm

$$R_n = \sum_{k=1}^{K} \Delta_k E[T_k(n)] = \Theta(n)$$

This is ‘bad’. We are interested in a sublinear regret $R_n = o(n)$. 
Two algorithms with

\[ R_n = O(\log n) \]
Example: $\epsilon$-greedy

$\epsilon$-greedy: at time $t$

- with probability $(1 - \epsilon_t)$ play the arm with the highest empirical mean
- with probability $\epsilon_t$ play a random arm

**Theorem** [Auer, Cesa-Bianchi, Fischer ’02]

For $\Delta = \min_{k: \Delta_k > 0} \Delta_k$, $\exists c$ s.t. for $\epsilon_t = \frac{c}{t}$,

$$R_n = O(\log n / \Delta^2).$$
Proof (sketch)

- $\Pr[J_t = j] = O\left(\frac{1}{t}\right)$ for every suboptimal arm $j$.

$$\Pr[J_t = j] \leq \frac{\varepsilon_t}{K} \Theta\left(\frac{1}{t}\right)$$

$$\Pr[J_t = j] \leq \frac{\varepsilon_t}{K} + \left(1 - \frac{\varepsilon_t}{K}\right) \Pr[\bar{X}_{j,T_j(t-1)} \geq \bar{X}_{T^*(t-1)}]$$

- Hence $E[T_j(n)] = O(\log n)$.

At round $t$:

- With high probability every arm has been chosen $\Omega(\log t)$ times.

- Therefore $\Pr\left[\bar{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right] = o\left(\frac{1}{t}\right)$. 

□
Theorem [Chernoff-Hoeffding]

Let $X_1, \ldots, X_n$ be random variables with common range $[0, 1]$, and such that $\mathbb{E}[X_i \mid X_1, \ldots, X_{i-1}] = \mu$. $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

$$\Pr[\bar{X}_n \geq \mu + a] \leq e^{-2a^2n}$$

In other words...

For any $\epsilon > 0$, with probability at least $(1 - \epsilon)$, we have

$$\mu \leq \bar{X}_n + \sqrt{\frac{\log(\epsilon^{-1})}{2n}}.$$
UCB Algorithm

- For fixed $\epsilon$, we have upper confidence bounds for all arms.
- Set $\epsilon = 1/t^4$

**UCB1**: at time $t$
- play the arm $j$ that maximizes $\bar{X}_{j,T_j(t-1)} + \sqrt{\frac{2 \log t}{T_j(t-1)}}$

**Theorem** [Auer, Cesa-Bianchi, Fischer ’02]

$$R_n = O \left( \log n \cdot \sum_{k: \Delta_k > 0} \frac{1}{\Delta_k} \right)$$
Proof

- Suboptimal arm $j$ is played at most $\ell = \Theta \left( \frac{\log n}{\Delta^2_j} \right)$ times.

\[
T_j(n) \leq \ell + \sum_{t=1}^{n} \{ J_t = j, T_j(t-1) \geq \ell \}
\]

\[
\leq \ell + \sum_{t=1}^{n} \sum_{r=1}^{t-1} \sum_{s=\ell}^{t-1} \left\{ \bar{X}_r^* + \sqrt{\frac{2 \log t}{r}} \leq \bar{X}_j, s + \sqrt{\frac{2 \log t}{s}} \right\}
\]

- $\ell$ is such that at least one of the following must hold:

\[
\left\{ \bar{X}_r^* \leq \mu^* - \sqrt{2 \log t/r} \right\} \quad \text{or} \quad \left\{ \bar{X}_j, s \geq \mu_j + \sqrt{2 \log t/s} \right\}
\]

- But this happens with probability $\leq 1/t^4$!

\[
\mathbb{E}[T_j(n)] \leq \ell + \sum_{t=1}^{\infty} \sum_{r=1}^{t-1} \sum_{s=1}^{t-1} \frac{2}{t^4} \leq \ell + \text{const} \quad \square
\]
Remarks

- Thompson (1933)/Agrawal, Goyal (2012)
- Robbins (1952)
- Lai, Robbins (1985)
- Auer, Cesa-Bianchi, Fischer (2002)
- Audibert, Bubeck (2010)/Garivier, Capplle (2011)

**Distribution dependent bounds**

\[ R_n = \Omega \left( \log n \cdot \sum_{k: \Delta_k > 0} \frac{\Delta_k}{kl(v_k \parallel v^*)} \right) \quad R_n = O \left( \log n \cdot \sum_{k: \Delta_k > 0} \frac{1}{\Delta_k} \right) \]

**Distribution independent bounds**

\[ R_n = \Omega \left( \sqrt{Kn} \right) \quad R_n = O \left( \sqrt{Kn} \right) \]
Discrete Optimization with Bandit Feedback

Setting:
- Set of $K$ arms ($\nu_k, \mu_k$ as before)
- We want to identify the best arm $\arg\max \mu_i$

**PAC** (Probably Approximately Correct) [Bechhofer (1954)]

**Goal:** $n$ as small as possible

$$\Pr [\mu_J \leq \mu^* - \epsilon] \leq \delta$$

**Fixed Budget** $B$ [Bubeck, Munos, Stoltz (2009)]

**Goal:** Minimize the error $\mu^* - \mu_J$

$$n \leq B$$
Trivial strategy

- Sample each arm $\Theta\left(\frac{1}{\epsilon^2} \log \frac{K}{\delta}\right)$ times
- Output empirical best arm

Successive Elimination

[Even-Dar, Mannor, Mansour (2006)]

- Play in rounds, eliminate arms whose confidence intervals do not overlap with currently best arm
- $n = O\left(H \log \frac{K}{\delta \Delta^2}\right)$

\[ H = \sum_{k: \Delta_k > 0} \frac{1}{\Delta_k^2} \]
Trivial strategy

- Divide budget evenly among the arms
- For $P[\mu^* - \mu_J \geq \epsilon] \leq \delta$, one needs $n = \Theta \left( \frac{K}{\epsilon^2} \log \frac{K}{\delta} \right)$

Successive Rejection

[Audibert, Bubeck, Munos (2010)]

- Play in rounds, eliminate arm with worst performance after every round, distribute the budget cleverly...
- Success with probability at least $(1 - \delta)$ if $n = \Omega \left( H \log \frac{K}{\delta} \right)$
Continuous Optimization with Bandit Feedback

\[ f(x) = \underbrace{\mu(x)}_{\text{deterministic}} + \underbrace{\nu(x)}_{\text{noise}} \]

- \( \nu(x) \) distribution with mean 0, \( \mathbb{E}[f(x)] = \mu(x) \)
- \( \mu(x) \) continuous

**Strategy:** Compare function values/discretize the space

- convex: \( \checkmark \), \( K = \Theta(d) \)
- weaker assumptions: Lipschitz/Hölder: \( \checkmark \), but \( K = \Omega(2^d) \)
  - Hierarchical Optimistic Optimization (HOO)
  [Bubeck, Munos, Stoltz, Szepesvari (2011)]

We miss a natural assumption that would allow good scaling in \( d \).
See slides for cited papers and: