Foundations of Reinforcement Learning

Multi-arm bandits: adversarial bandits

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2023

Outline

Introduction and formulation

Algorithms: from full-information to bandits

Analysis

Introduction and formulation

Limitations of stochastic bandits

In stochastic bandits, the reward distributions of the stochastic bandits do not depend on past rewards or actions, and the draws are i.i.d. over time.

This can be too restrictive, and unrealistic in practice.

- Is there a reward distribution?
- Are the rewards i.i.d.?

Adversarial bandits

Idea: make minimal assumptions about reward generation, and compete with the best action in hindsight.

For an n-arm adversarial bandit,

- the rewards are in the interval $[0, 1]$
- in each round t, the reward vector $r_t = [r_{i,t}]_{1 \leq i \leq n}$ is an arbitrary sequence
- all rewards are determined before action is taken

Example

Let $n = 3$, consider the following arbitrary rewards ...

k	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$\dots$
1	0.1	0.7	0.4	0.3	
2	0.5	0.1	0.6	0.1	
3	0.8	0.4	0.4	0.8	

- If action = arm 1, $\sum_{t=1}^{4} r_{1,t} = 0.1 + 0.7 + 0.4 + 0.3 = 1.5;$
- If action = arm 2, $\sum_{t=1}^{4} r_{2,t} = 0.5 + 0.1 + 0.6 + 0.1 = 1.4;$
- If action = arm 3, $\sum_{t=1}^{4} r_{3,t} = 0.8 + 0.4 + 0.4 + 0.8 = 2.4;$
- The best arm is arm 3.
- The best arm might change with the rewards.

Can we compete with the best arm no matter what the rewards are?

Performance metric: regret

Definition 1 (Expected regret)

The expected regret over T rounds for an action selection rule is defined as

$$R_T(r) = \max_{1 \le i \le n} \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} \mathbb{E}\left[r_{i_t,t} \right],$$

where the expectation is over the randomness of the learner's actions, and $r := \{r_t\}_{t=1}^{T}$ is the reward. The worst-case regret over all rewards is

$$R_T^\star = \sup_{r \in [0,1]^{T \times n}} R_T(r).$$

- Minimizing the regret has a “min-max” flavor.

Goal: achieve sublinear regret $R_T^\star = o(T)$.

Algorithms: from full-information to bandits

The power of randomization

- Exploration vs exploitation remains an important issue.
- Exploitation appears to be even more dangerous.
  - adversary can "exploit our exploitation".

Example: Let $n = 3$, the selection of the learner is given in red.

k	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$\dots$
1	0.1	0.7	0	0	
2	0	0.1	0.6	0.1	
3	0.8	0	0.4	0.8	

- For learner, $\sum_{t=1}^{4} r'_{i_t,t} = 0 + 0 + 0 + 0 = 0$.
- $R_4(r') = \max_{i=1,2,3} \sum_{t=1}^{4} r'_{i,t} = 2$. High regret!

The power of randomization

Without randomization in actions, the regret can be linear.

A construction: for any deterministic sequence of actions $\{i_t\}_{t=1}^T$, we can construct a reward sequence such that

$$\forall t : \quad r_{i,t} = \begin{cases} 0 & i = i_t \\ 1 & \text{otherwise} \end{cases}$$

By construction,

$$\sum_{i=1}^{n} \sum_{t=1}^{T} r_{i,t} = (n-1)T \quad \Longrightarrow \max_{1 \le i \le n} \sum_{t=1}^{4} r_{i,t} \ge \frac{(n-1)T}{n}.$$

$$\Longrightarrow R_n(r) = \max_{1 \le i \le n} \sum_{t=1}^{T} r_{i,t} - \underbrace{\sum_{i=1}^{T} r_{i_t,t}}_{=0} \ge \frac{(n-1)T}{n}!$$

Detour: bandits with full information

Randomization: Let $p_t = [p_{i,t}]$ be the probability of choosing different arms in round t.

How to design/update p_t?

Detour: Let's visit the bandit problem with full information, where we observe the entire reward vector

$$r_t = [r_{1,t}, r_{2,t}, \dots, r_{n,t}].$$

This is also known as *online learning with expert advice*, often formulated with losses rather than rewards. A huge field!

Exponential weight algorithm

Recall the regret

$$R_T(r) = \max_{1 \le i \le n} \sum_{t=1}^{T} \left(r_{i,t} - \mathbb{E}_{i_t \sim p_t} \left[r_{i_t,t} \right] \right).$$

Exponential-weight algorithm: assign a higher probability to arms with better performance according to exponential weights [Vovk, 1990, Littlestone and Warmuth, 1994]:

$$p_{i,t+1} \propto \exp\left(\eta \sum_{\ell=1}^{t} r_{i,\ell} \right) = \frac{\exp\left(\eta \sum_{\ell=1}^{t} r_{i,\ell} \right)}{\sum_{i=1}^{n} \exp\left(\eta \sum_{\ell=1}^{t} r_{i,\ell} \right)}, \qquad 1 \le i \le n$$

where $\eta > 0$ is some parameter.

- The randomness ensures exploration: even arms with 0 cumulative rewards so far get chances
- Higher weights encourage exploitation: η controls the trade-off exploitation and exploration

A bit more discussions

Incremental update:

$$p_{i,t+1} \propto \exp\left(\eta \sum_{\ell=1}^{t} r_{i,\ell}\right) \propto p_{i,t} \cdot e^{\eta r_{i,t}}$$

- A multiplicative combination of history information and new information

This algorithm has many names, e.g. multiplicative weight updates (MWU) method, Hedge, and was rediscovered many times in different contexts; see [**?**] for more information.

Exponential weight algorithm with full information

1. **Initialization:** set p_1 as a uniform distribution over $[n]$; set $S_{i,0} = 0$ for $i \in [n]$; parameter $\eta > 0$.
2. For each round $t = 1, 2, \ldots, T$:
   - Observe the reward $r_{i,t}$ for $i \in [n]$, and update the cumulative reward
     $$S_{i,t} = S_{i,t-1} + r_{i,t}.$$
   - Update the sampling probability over arms
     $$p_{i,t+1} \propto \exp\Big(\eta S_{i,t}\Big).$$

What is the regret of exponential weight algorithm?

$$R_T(r) = \max_{1 \le i \le n} \sum_{t=1}^{T} \left(r_{i,t} - \mathbb{E}\left[r_{i_t,t}\right]\right) = \max_{1 \le i \le n} \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} r_{i,t}.$$

Regret of exponential weight algorithm

Theorem 2 (Regret under full information)

The exponential weights algorithm with parameter $\eta > 0$ incurs regret

$$R_T^\star = \sup_{r \in [0,1]^{T \times n}} R_T(r) \leq T\eta + \frac{\log n}{\eta}.$$

Choosing $\eta = \sqrt{\log n / T}$ gives

$$R_T^\star \leq 2\sqrt{T \log n}.$$

sublinear regret!

Can we translate this back to the bandit setting?

Back to bandits

In bandits, we only observe $r_{i_t,t}$ for the pulled arm i_t!

A general recipe:

- Step 1: estimate the entire reward vector
- Step 2: plug this into the full-information algorithm

Importance-sampling estimator:

$$\widetilde{r}_{i,t} = \frac{r_{i,t}}{p_{i,t}} \mathbb{I}_{i_t = i} = \begin{cases} \frac{r_{i_t,t}}{p_{i_t,t}} & i = i_t \\ 0 & \text{otherwise} \end{cases}$$

or equivalently

$$\widetilde{r}_t = \left[0, \ldots, \frac{r_{i_t,t}}{p_{i_t,t}}, \ldots, 0 \right].$$

This is an *unbiased* estimate of the reward vector: $\mathbb{E}[\widetilde{r}_{i,t}] = r_{i,t}$.

EXP3 for adversarial bandits

[Auer et al., 2002]: EXP3 $=$ Exponential-weight algorithm for Exploration and Exploitation

1. **Initialization:** set p_1 as a uniform distribution over $[n]$; set $\widetilde{S}_{i,0} = 0$ for $i \in [n]$; parameter $\eta > 0$.
2. For each round $t = 1, 2, \ldots, T$:
   - Draw an arm i_t from the distribution p_t;
   - For each arm $i \in [n]$, compute the estimated reward

     $$\widetilde{r}_{i,t} = \frac{r_{i,t}}{p_{i,t}} \mathbb{I}_{i_t = i}$$

     and update the cumulative reward $\widetilde{S}_{i,t} = \widetilde{S}_{i,t-1} + \widetilde{r}_{i,t}$.
   - Update the sampling probability over arms

     $$p_{i,t+1} \propto \exp\left(\eta \widetilde{S}_{i,t}\right).$$

Regret of EXP3

Theorem 3 (Regret of EXP3)

The EXP3 algorithm with parameter $\eta > 0$ incurs regret

$$R_T^\star \le nT\eta + \frac{\log n}{\eta}.$$

- The first term is worse by a factor of n compared with the full-information case;
- Choosing $\eta = \sqrt{\log n/(nT)}$ gives

$$R_T^\star \le 2\sqrt{nT\log n}.$$

- Adversarial bandits are not harder than stochastic bandits: matches the worst-case regret bound $\widetilde{O}(\sqrt{nT})$ of UCB for stochastic bandits.

Analysis

Roadmap

- We will first prove the regret bound for the full-information setting.
- We then adapt it to the bandit setting.

Proof of Theorem 2 (full information)

Step 1: introduce the “magic” measure of progress (log-sum-exp):

$$\Phi_t = \frac{1}{\eta} \log \left(\sum_{i=1}^{n} \exp(\eta S_{i,t}) \right)$$

Recall that $p_{i,t+1} = \frac{\exp\left(\eta S_{i,t}\right)}{\sum_{i=1}^{n} \exp\left(\eta S_{i,t}\right)}$.

Basic facts:

- $\Phi_0 = \frac{1}{\eta} \log n$.
- $\Phi_T = \frac{1}{\eta} \log \left(\sum_{i=1}^{n} \exp(\eta S_{i,T}) \right) \geq \frac{1}{\eta} \log \left(\exp(\eta S_{i,T}) \right) = S_{i,T}$, for $i \in [n]$; in other words, Φ_T upper bounds the performance of individual arms, the first term in the regret.

Proof of Theorem 2 (full information)

Step 2: understand the temporal difference of Φ_t.

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &= \frac{1}{\eta}\log\left(\sum_{i=1}^n e^{\eta S_{i,t}}\right) - \frac{1}{\eta}\log\left(\sum_{i=1}^n e^{\eta S_{i,t-1}}\right) \\
&= \frac{1}{\eta}\log\frac{\sum_{i=1}^n e^{\eta S_{i,t-1}} e^{\eta r_{i,t}}}{\sum_{i=1}^n e^{\eta S_{i,t-1}}} \\
&= \frac{1}{\eta}\log\sum_{i=1}^n \frac{e^{\eta S_{i,t-1}}}{\sum_{i=1}^n e^{\eta S_{i,t-1}}} e^{\eta r_{i,t}} \\
&= \frac{1}{\eta}\log\sum_{i=1}^n p_{i,t} e^{\eta r_{i,t}},
\end{aligned}$$

which is linked to the sampling probabilities.

Proof of Theorem 2 (full information)

Step 2 - continued: deconstructing the log-sum, using basic facts:

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &= \frac{1}{\eta} \log \sum_{i=1}^{n} p_{i,t} e^{\eta r_{i,t}} \\
&\leq \frac{1}{\eta} \log \sum_{i=1}^{n} p_{i,t} \left(1 + \eta r_{i,t} + \eta^2 r_{i,t}^2\right) \qquad (e^x \leq 1 + x + x^2) \\
&\leq \frac{1}{\eta} \log \left(1 + \eta \sum_{i=1}^{n} p_{i,t} r_{i,t} + \eta^2 \sum_{i=1}^{n} p_{i,t} r_{i,t}^2\right) \\
&\leq \sum_{i=1}^{n} p_{i,t} r_{i,t} + \eta \sum_{i=1}^{n} p_{i,t} r_{i,t}^2 \qquad (\log(1+x) \leq x)
\end{aligned}$$

- $\forall x : \quad \log(1+x) \leq x$
- For $x \leq 1 : \quad e^x \leq 1 + x + x^2$

Proof of Theorem 2 (full information)

Step 3: by telescoping

$$\Phi_T - \Phi_0 = \sum_{t=1}^{T} (\Phi_t - \Phi_{t-1}) \le \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} r_{i,t} + \eta \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} r_{i,t}^2 .$$

Therefore,

$$\begin{aligned} R_T(r) = \max_{1 \le i \le n} S_{i,T} - \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} r_{i,t} &\le \Phi_T - \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} r_{i,t} \qquad (S_{i,T} \le \Phi_T) \\ &\le \Phi_0 + \eta \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} r_{i,t}^2 \\ &\le \frac{1}{\eta} \log n + \eta \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} \quad \left(\Phi_0 \le \frac{1}{\eta} \log n\right) \\ &\le \frac{\log n}{\eta} + \eta T . \quad \left(\sum_i p_{i,t} = 1\right) \end{aligned}$$

Proof of Theorem 3 (bandit)

We highlight the differences from the full-information case.

Step 1: introduce a measure of progress (log-sum-exp)

$$\Phi_t = \frac{1}{\eta} \log \left(\sum_{i=1}^{n} \exp(\eta \widetilde{S}_{i,t}) \right)$$

Basic facts:

- $\Phi_0 = \frac{1}{\eta} \log n$.
- $\Phi_T = \frac{1}{\eta} \log \left(\sum_{i=1}^{n} \exp(\eta \widetilde{S}_{i,T}) \right) \geq \widetilde{S}_{i,T}, \qquad i \in [n]$
- $\mathbb{E}[\Phi_T] \geq \mathbb{E}[\widetilde{S}_{i,T}] = S_{i,T}$.

Proof of Theorem 3 (bandit)

Step 2: understand the temporal difference of Φ_t.

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &= \frac{1}{\eta}\log\left(\sum_{i=1}^{n} e^{\eta \widetilde{S}_{i,t}}\right) - \frac{1}{\eta}\log\left(\sum_{i=1}^{n} e^{\eta \widetilde{S}_{i,t-1}}\right) \\
&= \frac{1}{\eta}\log\frac{\sum_{i=1}^{n} e^{\eta \widetilde{S}_{i,t-1}} e^{\eta \widetilde{r}_{i,t}}}{\sum_{i=1}^{n} e^{\eta \widetilde{S}_{i,t-1}}} \\
&= \frac{1}{\eta}\log\sum_{i=1}^{n} \frac{e^{\eta \widetilde{S}_{i,t-1}}}{\sum_{i=1}^{n} e^{\eta \widetilde{S}_{i,t-1}}} e^{\eta \widetilde{r}_{i,t}} \\
&= \frac{1}{\eta}\log\sum_{i=1}^{n} p_{i,t} e^{\eta \widetilde{r}_{i,t}}
\end{aligned}$$

By same arguments, we obtain

$$\Phi_t - \Phi_{t-1} \le \sum_{i=1}^{n} p_{i,t}\widetilde{r}_{i,t} + \eta \sum_{i=1}^{n} p_{i,t}\widetilde{r}_{i,t}{}^2.$$

Proof of Theorem 3 (bandit)

Step 3: telescoping

$$\Phi_T - \Phi_0 \leq \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} \widetilde{r}_{i,t} + \eta \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} {\widetilde{r}_{i,t}}^2,$$

which leads to

$$\mathbb{E}[\Phi_T] - \Phi_0 \leq \sum_{t=1}^{T} \sum_{i=1}^{n} p_{i,t} r_{i,t} + \eta \sum_{t=1}^{T} \sum_{i=1}^{n} r_{i,t}^2$$

after taking expectations using

$$\mathbb{E}[\widetilde{r}_{i,t}] = r_{i,t} \qquad \mathbb{E}[{\widetilde{r}_{i,t}}^2] = \frac{r_{i,t}^2}{p_{i,t}}.$$

Proof of Theorem 3 (bandit)

Step 4: finishing up.

$$
\begin{aligned}
R_T(r) &= \max_{1\le i\le n} S_{i,T} - \sum_{t=1}^{T}\sum_{i=1}^{n} p_{i,t} r_{i,t} \\
&\le \mathbb{E}[\Phi_T] - \sum_{t=1}^{T}\sum_{i=1}^{n} p_{i,t} r_{i,t} \\
&\le \Phi_0 + \eta \sum_{t=1}^{T}\sum_{i=1}^{n} r_{i,t}^2 \\
&\le \frac{\log n}{\eta} + \eta n T.
\end{aligned}
$$

References I

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002).
The nonstochastic multiarmed bandit problem.
SIAM Journal on Computing, 32(1):48–77.

Littlestone, N. and Warmuth, M. K. (1994).
The weighted majority algorithm.
Information and Computation, 108(2):212–261.

Vovk, V. G. (1990).
Aggregating strategies.
In *Proceedings of the third annual workshop on Computational learning theory*, pages 371–386.