Exploration Analysis in Bernoulli Bandits Using Martingales

Clement Leung

Chinese University of Hong Kong, Shenzhen &

Shenzhen Research Institute of Big Data

clementleung@cuhk.edu.cn

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Clement LEUNG

- FULL PROFESSORSHIPS at
  - University of London, UK; National University of Singapore; Chinese University of Hong Kong, Shenzhen, China; Hong Kong Baptist University; Victoria University, Australia
- Two US patents, five books and over 150 research articles
- Program Chair, Keynote Speaker, Panel Expert of major International Conferences
- Editorial Board of ten International Journals
- Listed in Who's Who in the World and Great Minds of the 21st Century
- Fellow of the British Computer Society, Fellow of the Royal Society of Arts, Chartered Engineer

Bandit Machines

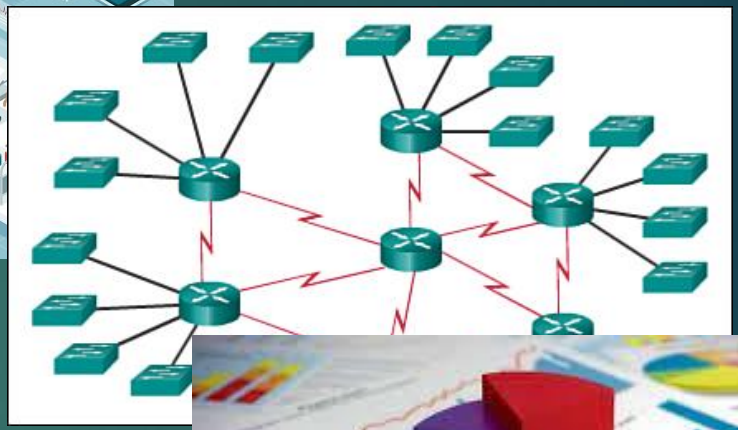
Which machine to play to maximize reward ?

Exploration and Exploitation

Should I always order my favorite dishes or try some exciting new ones?

Common Applications

- Patient management by examining the effect of different treatments
- Efficient network routing for improving performance
- Financial portfolio investment
- Dynamic allocation of resources to different projects



Bernoulli Bandits

Binary Outcome

After the lever is pulled, the machine provides a reward of +1 with probability p, and a zero reward with probability $q = 1- p$.

Let by U_i be the random variable of each such reward
$\Rightarrow$ the *gross positive reward* up to and including the t^{th} play is

$$L_t = U_1 + \ldots + U_t$$

Non-Bernoulli Bandits

- If the rewards are not of the Bernoulli type, then the rewards will not be just 0 or 1
- Since playing the game is not free, we would regard a zero reward to be actually a negative reward – one may consider that the player has lost this time, since he or she needs to pay in order to play the game in the first place
- Let the cost for each play (i.e. each pull of the bandit arm) be $c > 0$, and the payout for each positive reward be a positive random variable $V > 0$ with finite expectation
- Let R_i be the random variable denoting the reward outcome of the i^{th} play, which equals V with probability p and equals $-c$ with probability $q = 1-p$
- The *net positive reward* up to and including the t^{th} play is

$$M_t = R_1 + \ldots + R_t$$

- We denote the σ-field of R_t by F_t, i.e.

$$F_t = \sigma(R_1, \ldots, R_t)$$

and the σ-field of U_t by S_t, i.e.

$$S_t = \sigma(U_1, \ldots, U_t)$$

- The corresponding filtrations are denoted by $\{F_n\}$ and $\{G_n\}$ respectively

Predicting Future Reward Possibilities

Theorem I

With respect to the filtration $\{F_n\}$ the net positive reward process M_t is

- a supermartingale if $E(V) < c(1-p)/p$
- a martingale if $E(V) = c(1-p)/p$
- a submartingale if $E(V) > c(1-p)/p$

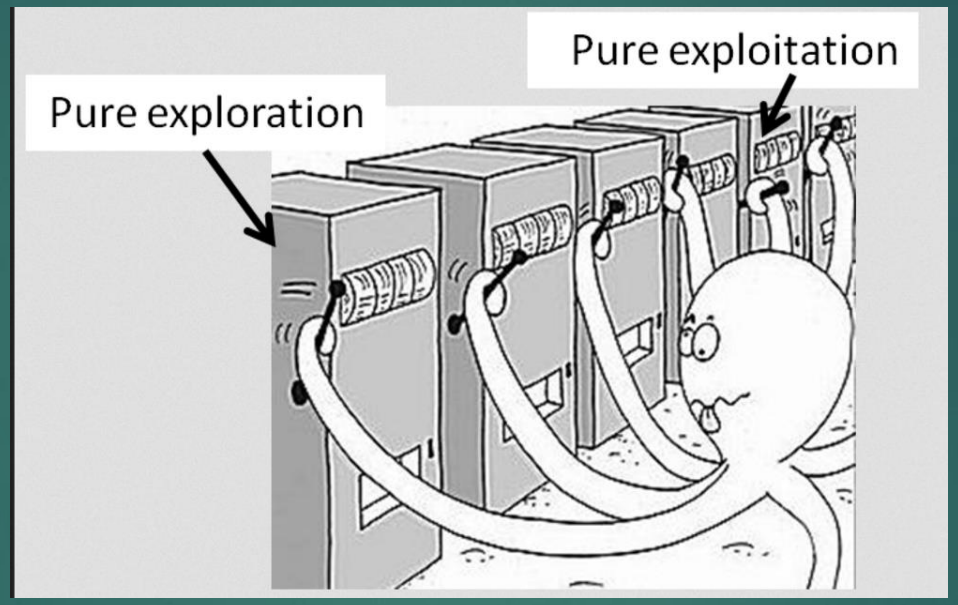
Greedy Algorithms

- Epsilon-greedy strategy
  - the most favorable arm is selected for a proportion epsilon of the time, and for the rest of the time, the arms are randomly and uniformly chosen
- Epsilon-first strategy
  - there is an exclusive exploration phase, during which an arm is randomly and uniformly chosen
  - After the completion of the exploration phase, there is an exclusive exploitation phase during which the most favorable arm is always chosen

Greedy Algorithms

- Epsilon-decreasing strategy
  - While there is a distinct transition point from the exploration phase to the exploitation phase in the epsilon-first strategy, the transition from exploration to exploitation occurs gradually over time through progressively reducing the value of epsilon in the epsilon-greedy strategy
- Adaptive epsilon-greedy strategies
  - Similar to the epsilon-decreasing strategy except that the epsilon value decreases either in accordance with the learning progress or some Bayesian update procedures

Exploring by hopping from Machine to Machine



Some exploration algorithms such as UCB or Thompson Sampling requires all the machines to be available for exploration, and the exploration procedures require hopping from machine to machine – this is often neither realistic nor possible because the machines may be occupied by other gamblers

More useful to systematically explore a single machine at a time

Exploration Episodes and Stopping Times

- Exploration can be viewed as a Binary Classification Problem
  - To classify bandits as exploitable and non-exploitable
  - Then one may, through further exploration, identify the best machine(s) to exploit
- Stopping Rule ST1: the exploration episode will stop immediately after *m* *consecutive* positive rewards have been received – this stopping time is denoted by T_1
- Stopping Rule ST2: the exploration episode will stop immediately after *m* *total* positive rewards have been received – this stopping time is denoted by T_2
- The termination of such episodes (i.e. ST1 and ST2) will result in the inclusion of the relevant bandit machines on the "White List", i.e. these are worthy of exploitation

Exploration Episodes and Stopping Times

- Stopping Rule ST3: the exploration episode will stop immediately after *r* consecutive negative rewards have been received – this stopping time is denoted by T_3
- Stopping Rule ST2: the exploration episode will stop immediately after *r* total negative rewards have been received – this stopping time is denoted by T_4
- The termination of the above episodes (i.e. ST3 and ST4) will result in the inclusion of the relevant bandit machines on the "Black List", i.e. these are unworthy of exploitation

Exploration Episodes and Stopping Times

Theorem II

- The random variables T_1, T_2, T_3, T_4 are stopping times of the sequence $\{U_i\}$ with respect to the filtration $\{G_n\}$

Corollary

- The random variables T_1, T_2 are stopping times of the sequence $\{R_i\}$ with respect to the filtration $\{F_n\}$

Unfair Game - Supermartingale

- The above stopping rules are quite realistic and often resemble the behavior and psychology of some actual gamblers
  - They tend to depart once a run of good luck or bad luck is experienced or when too many losses or wins are accumulated
- From the properties of martingales, we note that a supermartingale would result in an unfair game from the point of view of the player
  - Submartingales would be relatively rare unless an entrance fee is charged for playing the bandits
- We shall determine the cost of exploration for the above stopping times

Exploration Cost for ST1

Let b_n be the probability that m consecutive positive rewards occurs at trial n, with $n \geq m$, not necessarily for the first time, and we denote by $B(z)$ be the corresponding probability generating function. Then we have

$$B(z) = \frac{1 - z + qp^m z^{m+1}}{(1 - z)(1 - p^m z^m)} .$$

Exploration Cost for ST1

We denote by $A(z)$ the probability generating function for the event that the accumulation of m positive rewards occurs for the first time. Then the generating function $A(z)$ is related to $B(z)$ by

$$A(z) = \frac{B(z) - 1}{B(z)}$$

which gives

$$A(z) = \frac{p^m z^m}{1 - q^m \sum_{k=0}^{m-1} p^k z^k} .$$

Mean and Variance of the Exploration Duration for ST1

$$\mathrm{E}[T_1] = A'(1) = \frac{1-p^m}{qp^m},$$

$$\mathrm{Var}[T_1] = A''(1) + A'(1) - A'(1)^2$$
$$= \frac{1}{q^2p^{2m}} - \frac{2m+1}{qp^m} - \frac{p}{q^2}$$

Exploration Properties of ST1

Theorem III

The average gross positive reward under stopping rule ST1 is given by

$$\frac{1-p^m}{qp^{m-1}} - (1+p)m$$

and the associated average duration of the episode is given by

$$\frac{1-p^m}{qp^m}\,.$$

Exploration Cost for ST2

The probability generating function of $F(z)$ corresponding to T_2 is

$$F(z) = [\frac{pz}{(1-qz)}]^m$$

Exploration Properties for ST2

the mean and variance of T_2 can be obtained

$$E[T_2] = F'(1) = \frac{m}{p} ,$$

$$\mathrm{Var}[T_2] = F''(1) + F'(1) - F'(1)^2 = \frac{mq}{p^2} .$$

It is not hard to see that $T_1 \geq T_2$, and hence $E(T_1) \geq E(T_2)$. Hence we have:

Theorem IV

The average gross positive reward under stopping rule ST2 is m, and the average duration of the episode is m/p.

Exploration Properties of ST3 and ST4

The corresponding results for ST3 and ST4 may be obtained by making use of the reflection principle, through interchanging the role of *p* and *q*, and replacing *m* by *r*. Consequently, we have the following theorems.

Theorem V

The average gross negative reward under stopping rule ST3 is given by

$$\frac{1-q^r}{pq^{r-1}} - (1+q)r$$

and the average episode duration is given by

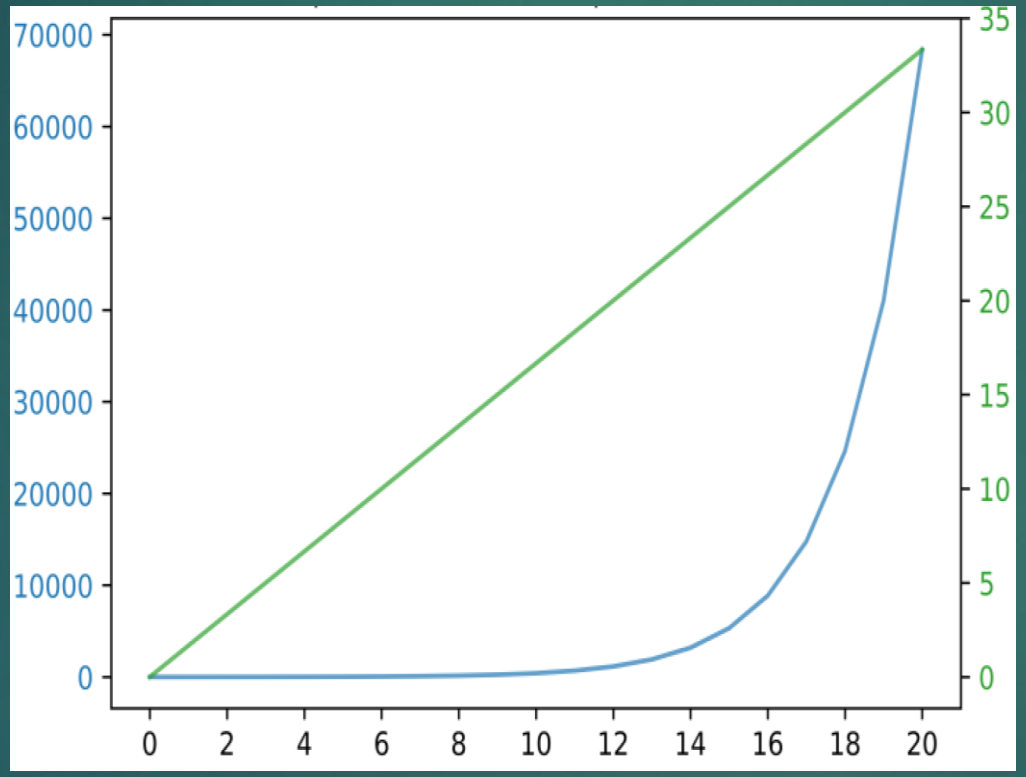
$$\frac{1-q^r}{pq^r}$$

Theorem VI

The average gross negative reward under stopping rule ST4 is *r*, and the average episode duration is *r*/*q*

Comparison of ST1 and ST2

No. of Rewards Required



The left vertical axis is used for $E(T_1)$ with an appropriate scale.

The right vertical axis is used for $E(T_2)$.

The performance of ST1 is manifested in a steep climb in the number of trials as m increases, as opposed to a relatively moderate increase in ST2.

Cost Comparison of Stopping Rules ST1 and ST2 ($p = 0.6$).

Non-Bernoulli Bandits for ST1

The average reward at the termination of the episode under rule ST1 is

$$[E(T_1) - m]pE(V) + mE(V) - cE(T_1)$$

The average reward under stopping rule ST1 for an episode

$$E(M_{T_1}) = p\left[\frac{1-p^m}{qp^m}\right]E(V) + (1-p)mE(V) - c\left[\frac{1-p^m}{qp^m}\right].$$

Non-Bernoulli Bandits for ST2

The average total positive rewards from stopping rule ST2, ignoring the cost of play, is $mE(V)$.

The average net reward under stopping rule ST2 is

$$E(M_{T_2}) = m[E(V) - \frac{c}{p}].$$

Regret Analysis

For a given suboptimal bandit machine with probability $p_j < p$, the average regret for a given bandit machine j under ST1 is

$$p\left[\frac{1-p^m}{qp^m}\right]E(V) + (1-p)mE(V) - c\left[\frac{1-p^m}{qp^m}\right]$$

$$- p_j\left[\frac{1-p_j^m}{q_j p_j^m}\right]E(V) + \left(1-p_j\right)mE(V) + c\left[\frac{1-p_j^m}{q_j p_j^m}\right].$$

Regret Analysis

The total average regret under ST1 is

$$E(\rho_1)$$

$$= pK\left[\frac{1-p^m}{qp^m}\right]E(V) + (1-p)mKE(V) - cK\left[\frac{1-p^m}{qp^m}\right]$$

$$-\sum_{j=1}^{K}\left\{\left[\frac{1-p_j^m}{q_j p_j^m}\right]E(V) + \left(1-p_j\right)mE(V) - c\left[\frac{1-p_j^m}{q_j p_j^m}\right]\right\}.$$

Regret Analysis

For a suboptimal bandit machine, the average regret for a particular sub-optimal machine *j* under ST2 is

$$E(\rho_2) = \frac{mc(p - p_j)}{pp_j}$$

giving the total average regret $E(\rho_2)$ incurred from the exploration of all *K* bandit machines under ST2 as

$$\sum_{j=1}^{K} \frac{mc(p - p_j)}{pp_j}.$$

Experiments

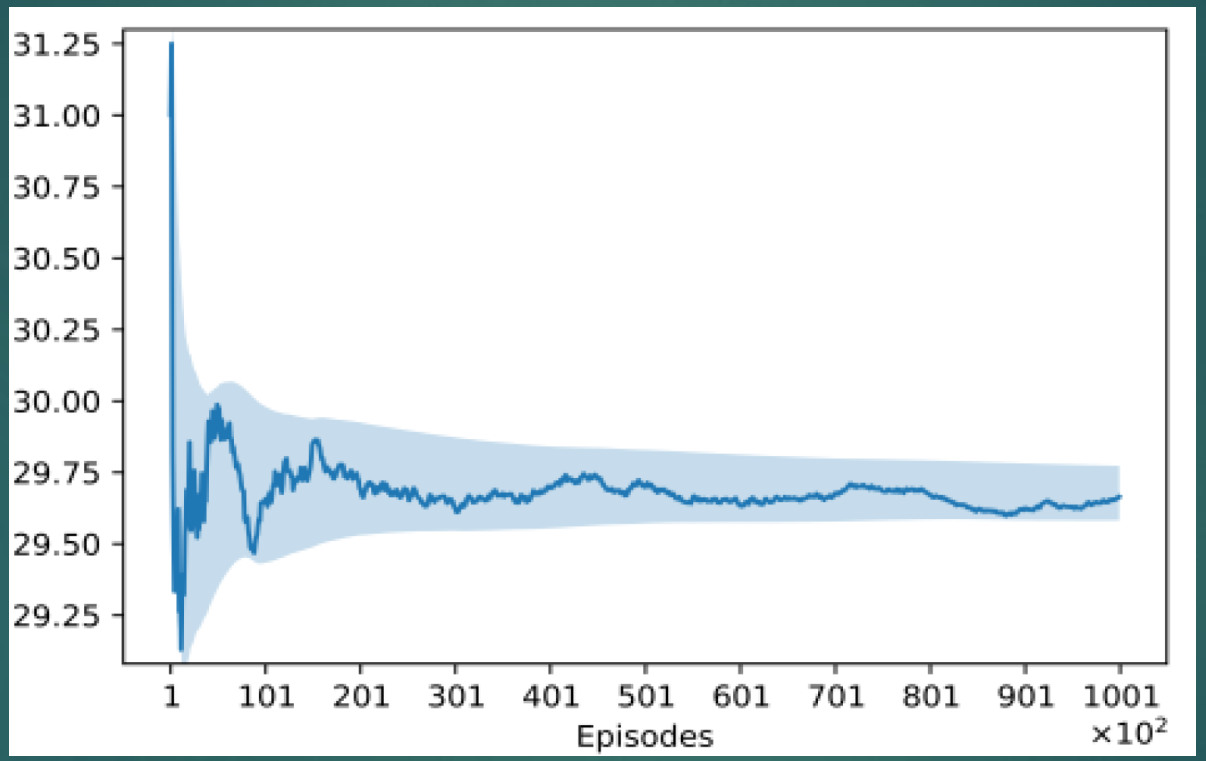
m	p	$E[T_1]$ (th)	$E[T_1]$ (expt)	Err (%)	$E[T_2]$ (th)	$E[T_2]$ (expt)	Err (%)	std. $[T_1]$ (th)	std. $[T_1]$ (expt)	Err (%)	std. $[T_2]$ (th)	std. $[T_2]$ (expt)	Err (%)
3	0.6	9.07	9.05	0.247	5.0	4.99	0.200	7.01	7.01	0.018	1.83	1.81	0.863
	0.75	5.48	5.47	0.018	4.0	4.00	0.000	3.40	3.38	0.307	1.15	1.15	0.006
	0.9	3.71	3.72	0.129	3.33	3.33	0.100	1.46	1.46	0.584	0.61	0.61	0.355
5	0.6	29.65	29.77	0.397	8.33	8.35	0.199	26.00	26.16	0.599	2.36	2.36	0.142
	0.75	12.86	12.84	0.121	6.67	6.66	0.100	9.31	9.31	0.035	1.49	1.49	0.142
	0.9	6.94	6.93	0.031	5.56	5.56	0.080	3.24	3.23	0.469	0.79	0.78	0.253
7	0.6	86.81	87.02	0.242	11.67	11.67	0.029	81.44	81.91	0.673	2.79	2.79	0.292
	0.75	25.97	25.85	0.461	9.33	9.34	0.071	20.89	20.84	0.226	1.76	1.77	0.355
	0.9	10.91	10.91	0.057	7.77	7.78	0.029	5.79	5.80	0.142	0.93	0.93	0.405
10	0.6	410.95	412.69	0.422	16.67	16.67	0.019	402.81	405.22	0.597	3.33	3.34	0.181
	075	67.03	67.05	0.030	13.33	13.33	0.025	59.51	59.37	0.244	2.11	2.10	0.334
	0.9	18.68	18.67	0.037	11.11	11.11	0.010	7.01	7.01	0.018	1.83	1.81	0.863

Comparison of theory and experiments

Since the error% all fall below 1%, we observe good agreement with the theoretical predictions

Experiments

$p = 0.6$
$m = 5$

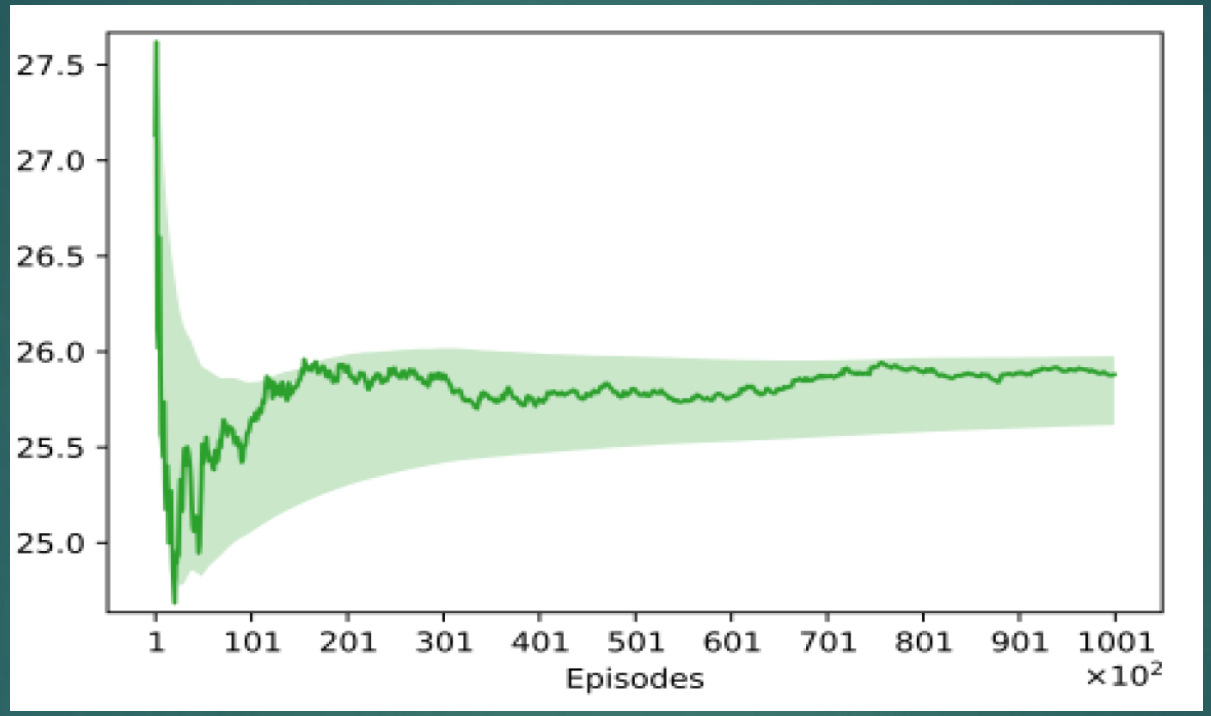


$E(T_1)$

Converges to the theoretical value

Experiments

$p = 0.6$
$m = 5$

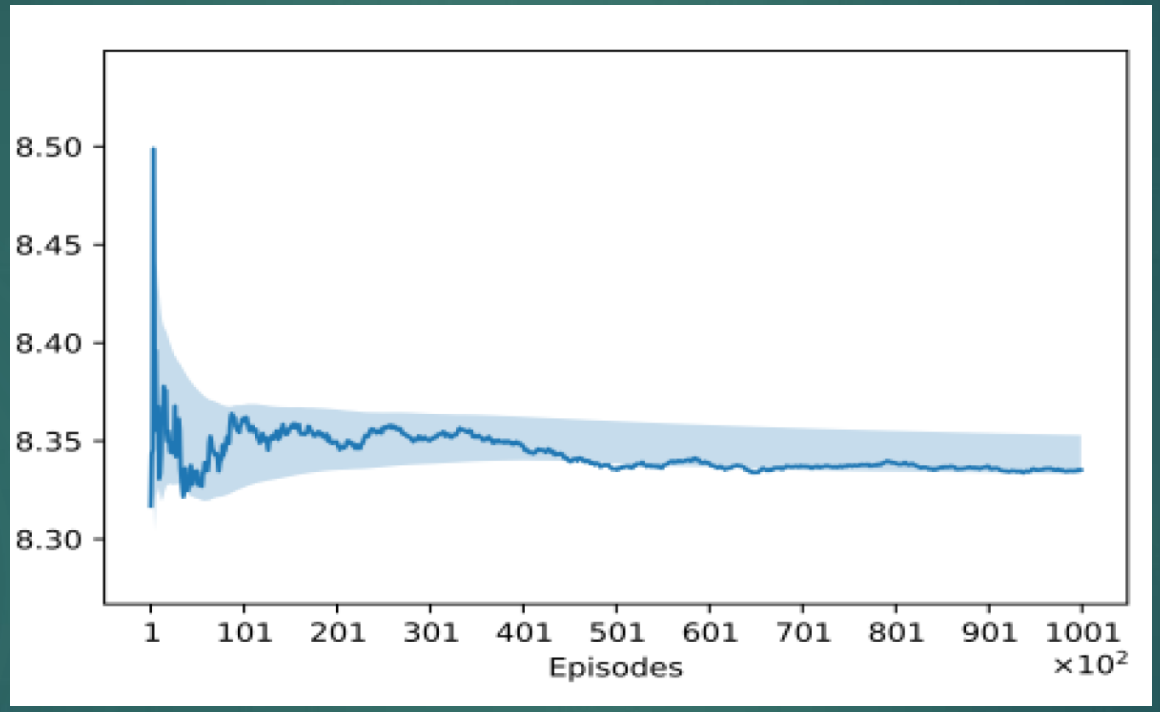


Std Dev(T_1)

Converges to the theoretical value

Experiments

$p = 0.6$
$m = 5$

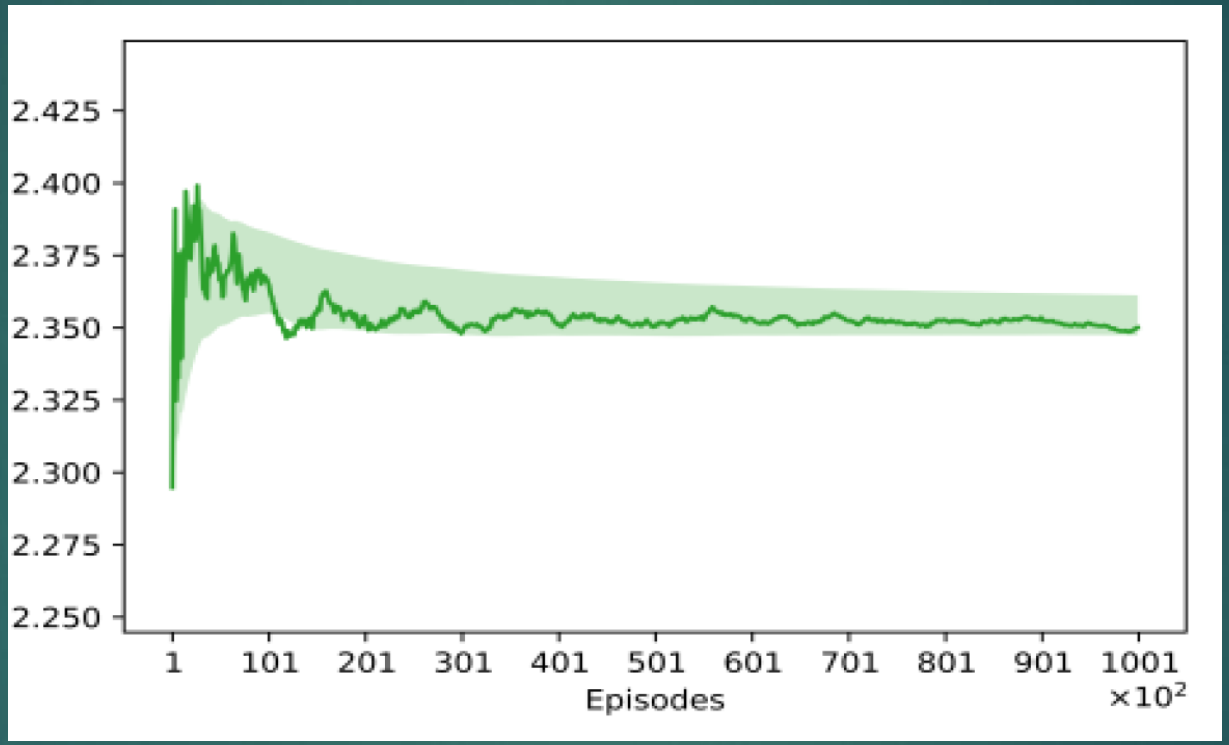


$E(T_2)$

Converges to the theoretical value

Experiments

$p = 0.6$
$m = 5$



Std Dev(T_2)

Converges to the theoretical value

Conclusion

- Uses martingales to study Bernoulli *K*-armed bandits as well as non-Bernoulli ones
- Analyzed the rewards for *K*-armed bandits, focusing on an episodic basis, where episodes are determined in terms of martingale stopping times
  - Martingales are particularly appropriate in the current situation of game playing
  - Typically supermartingales are prevalent
- The stopping times delimit an episode of play, and the rewards for each episode is determined under different stopping criteria, along with the duration of an episode
  - The characteristics of four stopping rules ST1, ST2, ST3, ST4 are analyzed
- Experimentations have been carried out, which show good agreements with the predicted findings
- Since bandit situations are ubiquitous and pervasive in many decision-making contexts, the same methodology and analysis may be usefully deployed in a broad variety of practical problems that have an equivalence with the *K*-armed bandit framework, supporting the making of sound operational decisions


Thank You