Lecture 6a: Multi-Armed Bandits
CS885 Reinforcement Learning

2025-01-23

Complementary readings: [SutBar] Sec. 2.1-2.7, [Sze] Sec. 4.2.1-4.2.2

Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

Outline

- Exploration/exploitation tradeoff
- Regret
- Multi-armed bandits
  - ϵ-greedy strategies
  - Upper confidence bounds

Exploration/Exploitation Tradeoff

- Fundamental problem of RL due to the active nature of the learning process
- Consider one-state RL problems known as bandits

Stochastic Bandits

- Formal definition:
  - Single state: $S = \{s\}$
  - A: set of actions (also known as arms)
  - Space of rewards (often re-scaled to be [0,1])
- No transition function to be learned since there is a single state
- We simply need to learn the **stochastic** reward function

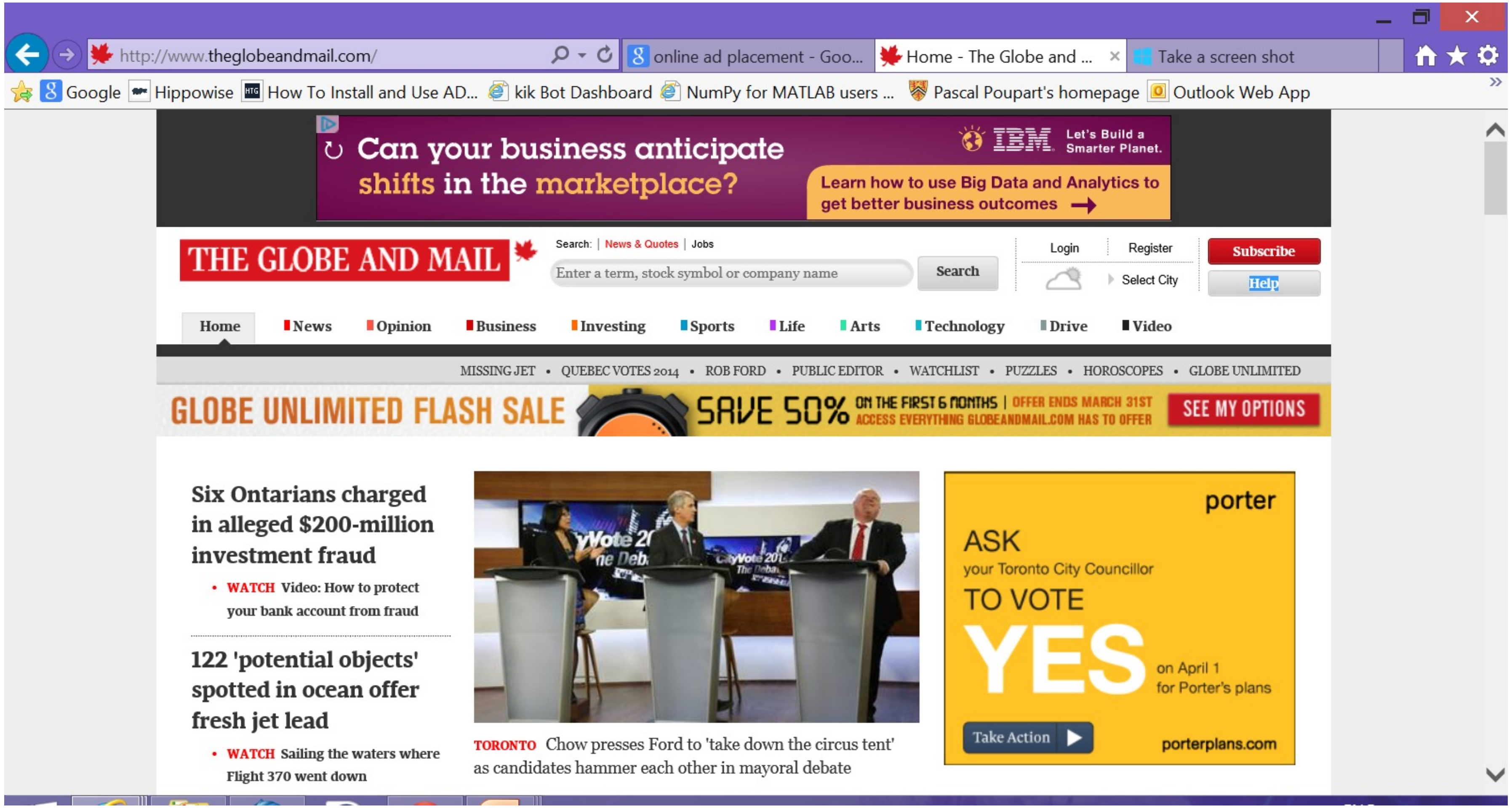
Origin and Applications

- "bandit" comes from gambling where slot machines can be thought as one-armed bandits.

Applications

- **Marketing** (ad placement, recommender systems)
- **Loyalty programs** (personalized offers)
- **Pricing** (airline seat pricing, cargo shipment pricing, food pricing)
- **Optimal design** (web design, interface personalization)
- **Networks** (routing)

Online Ad Placement



UNIVERSITY OF WATERLOO

Online Ad Optimization

- Problem: which ad should be presented?

- Answer: present ad with highest payoff

$$payoff = clickThroughRate \times payment$$

  - Click through rate: probability that user clicks on ad
  - Payment: $$ paid by advertiser
    - Amount determined by an auction

Simplified Problem

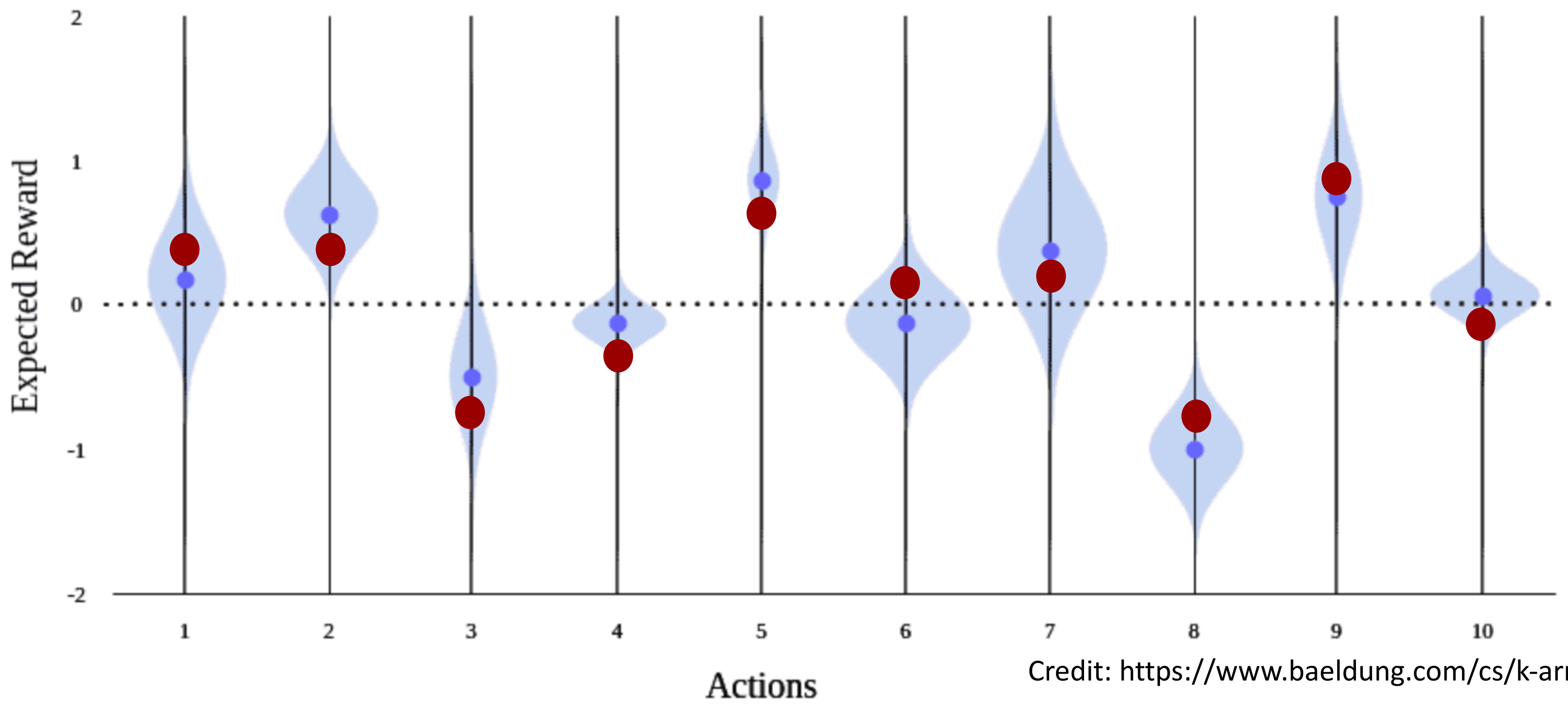
- Assume payment is 1 unit for all ads
- Need to estimate click through rate

- Formulate as a bandit problem:
  - Arms: the set of possible ads
  - Rewards: 0 (no click) or 1 (click)

- In what order should ads be presented to maximize revenue?
  - How should we balance exploitation and exploration?

Uncertainty Quantification

- Distribution of rewards: $\Pr(r|a)$
- Expected reward: $R(a) = E(r|a)$
- Empirical average reward: $\tilde{R}(a) = \frac{1}{n}\sum_t^n r_t$



Credit: https://www.baeldung.com/cs/k-armed-bandit-problem

Simple Heuristics

- Greedy strategy: select the arm with the highest average so far
  - May get stuck due to lack of exploration

- ϵ-greedy: select an arm at random with probability ϵ and otherwise do a greedy selection
  - Convergence rate depends on choice of ϵ

Regret

- Let $R(a)$ be the unknown average reward of a
- Let $r^* = \max_a R(a)$ and $a^* = argmax_a \, R(a)$

- Denote by $loss(a)$ the expected regret of a

$$loss(a) = r^* - R(a)$$

- Denote by $Loss_n$ the expected cumulative regret for n time steps

$$Loss_n = \sum_{t=1}^{n} loss(a_t)$$

Theoretical Guarantees

- When ϵ is constant, then
  - For large enough t: $\Pr(a_t \neq a^*) \approx \epsilon$
  - Expected cumulative regret: $Loss_n \approx \sum_{t=1}^{n} \epsilon = O(n)$
    - Linear regret

- When $\epsilon_\mathrm{t} \propto 1/t$
  - For large enough t: $\Pr(a_t \neq a^*) \approx \epsilon_t = O\left(\frac{1}{t}\right)$
  - Expected cumulative regret: $Loss_n \approx \sum_{t=1}^{n} \frac{1}{t} = O(\log\ n)$
    - Logarithmic regret

Empirical Mean

- Problem: how far is the empirical mean $\tilde{R}(a)$ from the true mean $R(a)$?

- If we knew that $\left|R(a) - \tilde{R}(a)\right| \leq bound$
  - Then we would know that $R(a) < \tilde{R}(a) + bound$
  - And we could select the arm with best $\tilde{R}(a) + bound$

- Overtime, additional data will allow us to refine $\tilde{R}(a)$ and compute a tighter $bound$.

Positivism in the Face of Uncertainty

- Suppose that we have an oracle that returns an upper bound $UB_n(a)$ on $R(a)$ for each arm based on n trials of arm a.

- Suppose the upper bound returned by this oracle converges to $R(a)$ in the limit:
  - i.e., $\lim_{n\to\infty} UB_n(a) = R(a)$

- Optimistic algorithm
  - At each step, select $argmax_a \; UB_n(a)$

Convergence

- Theorem: An optimistic strategy that always selects $\text{argmax}_{\text{a}} UB_n(a)$ will converge to a^*

- Proof by contradiction:
  - Suppose that we converge to suboptimal arm a after infinitely many trials.
  - Then $R(a) = UB_\infty(a) \geq UB_\infty(a') = R(a')\ \forall a'$
  - But $R(a) \geq R(a')\ \forall a'$ contradicts our assumption that a is suboptimal.

UNIVERSITY OF WATERLOO

Probabilistic Upper Bound

- Problem: We can't compute an upper bound with certainty since we are sampling
- However we can obtain measures f that are upper bounds most of the time
  - i.e., $\Pr\big(R(a) \leq f(a)\big) \geq 1 - \delta$
  - Example: Hoeffding's inequality

$$\boxed{\Pr\left(R(a) \leq \tilde{R}(a) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n_a}} \right) \geq 1 - \delta}$$

where n_a is the number of trials for arm a

Upper Confidence Bound (UCB)

- Set $\delta_n = 1/n^4$ in Hoeffding's bound
- Choose a with highest Hoeffding bound

UCB(h)

$V \leftarrow 0,\ \ n \leftarrow 0,\ \ n_a \leftarrow 0 \quad \forall a$

Repeat until $n = h$

Execute $\text{argmax}_a\ \tilde{R}(a) + \sqrt{\frac{2 \log n}{n_a}}$

Receive r

$V \leftarrow V + r$

$\tilde{R}(a) \leftarrow \frac{n_a \tilde{R}(a) + r}{n_a + 1}$

$n \leftarrow n + 1,\quad n_a \leftarrow n_a + 1$

Return V

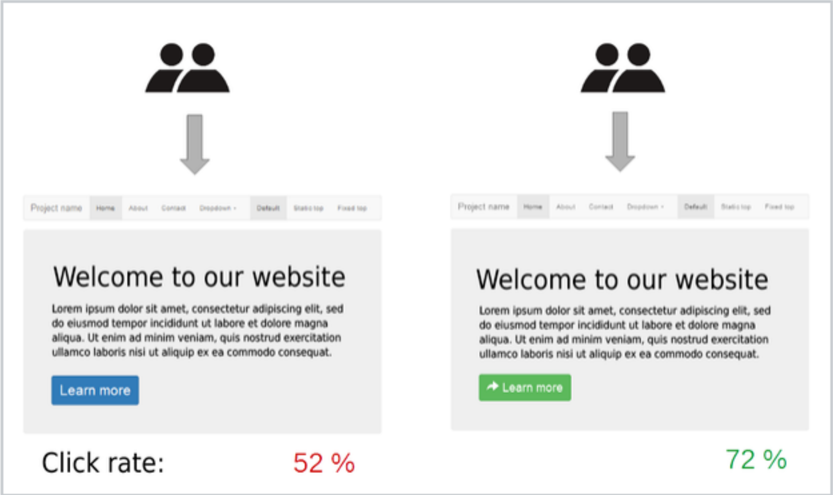
UCB Convergence

- **Theorem:** Although Hoeffding's bound is probabilistic, UCB converges.
- **Idea:** As n increases, the term $\sqrt{\frac{2\log n}{n_a}}$ increases, ensuring that all arms are tried infinitely often
- Expected cumulative regret: $Loss_n = O(\log n)$
  - Logarithmic regret

Extension of A/B Testing

- **A/B Testing:** randomized experiment with 2 variants
  - Select best variant after completion of experiment

Example: email marketing
- "Offer ends this Saturday! Use code A" (response rate: 5%)
- "Offer ends soon! Use code B" (response rate: 3%)



Example of A/B testing on a website. By randomly serving visitors two versions of a website that differ only in the design of a single button element, the relative efficacy of the two designs can be measured.

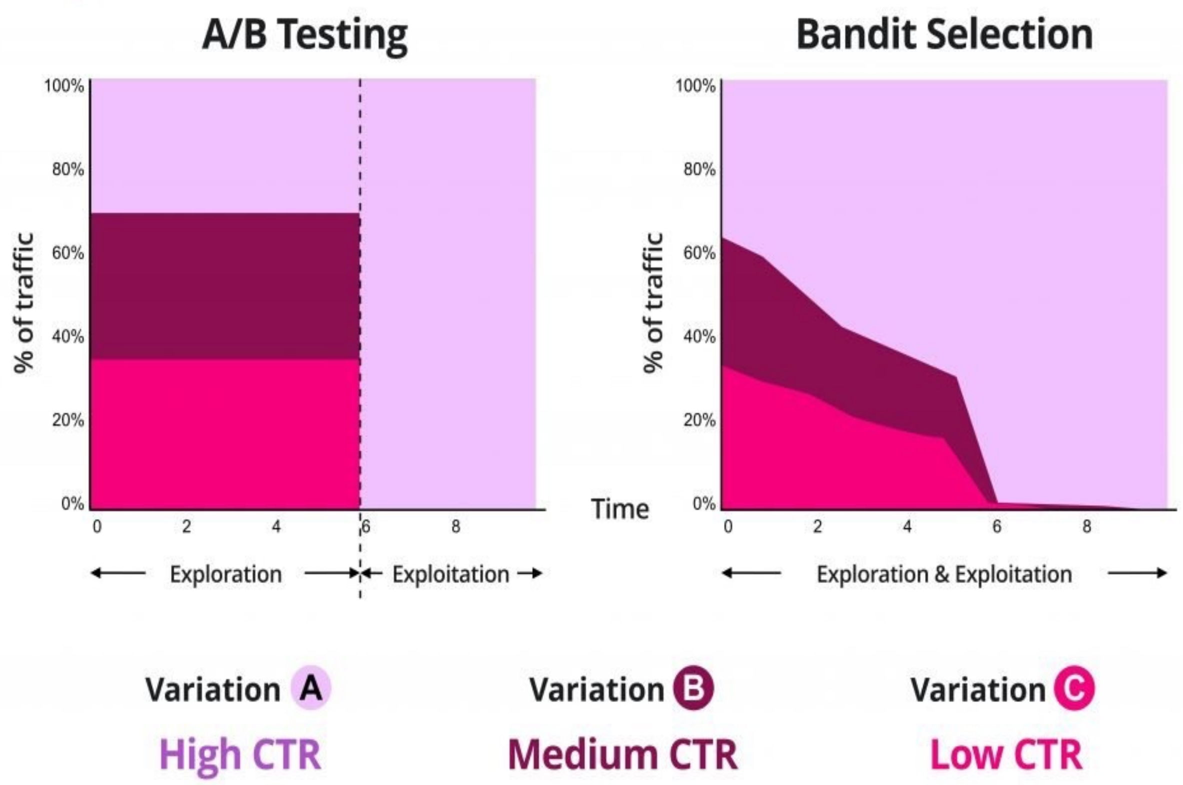
- **Multi-armed bandits:** form of continual A/B testing

UNIVERSITY OF WATERLOO

Multi-Armed Bandit

Components	Formal Def	Marketing
Actions (arms)	$a \in A$	{A, B, C}
Rewards	$r \in \mathbb{R}$	{0, 1}
Reward model	$\Pr(r\|a)$	unknown
Horizon	$h \in \mathbb{N}$ or ∞	$h = \infty$

Credit: Shubhankar Gupta (vwo.com)



Summary

- Stochastic bandits
  - Exploration/exploitation tradeoff
- ϵ-greedy and UCB
  - Theory: logarithmic expected cumulative regret

- In practice:
  - UCB often performs better than ϵ-greedy
  - Many variants of UCB improve performance