MAT8034: Machine Learning

Introduction to Multi-armed Bandits

Fang Kong

https://fangkongx.github.io/Teaching/MAT8034/Spring2025/index.html

What are bandits? [Lattimore and Szepesvári, 2020]

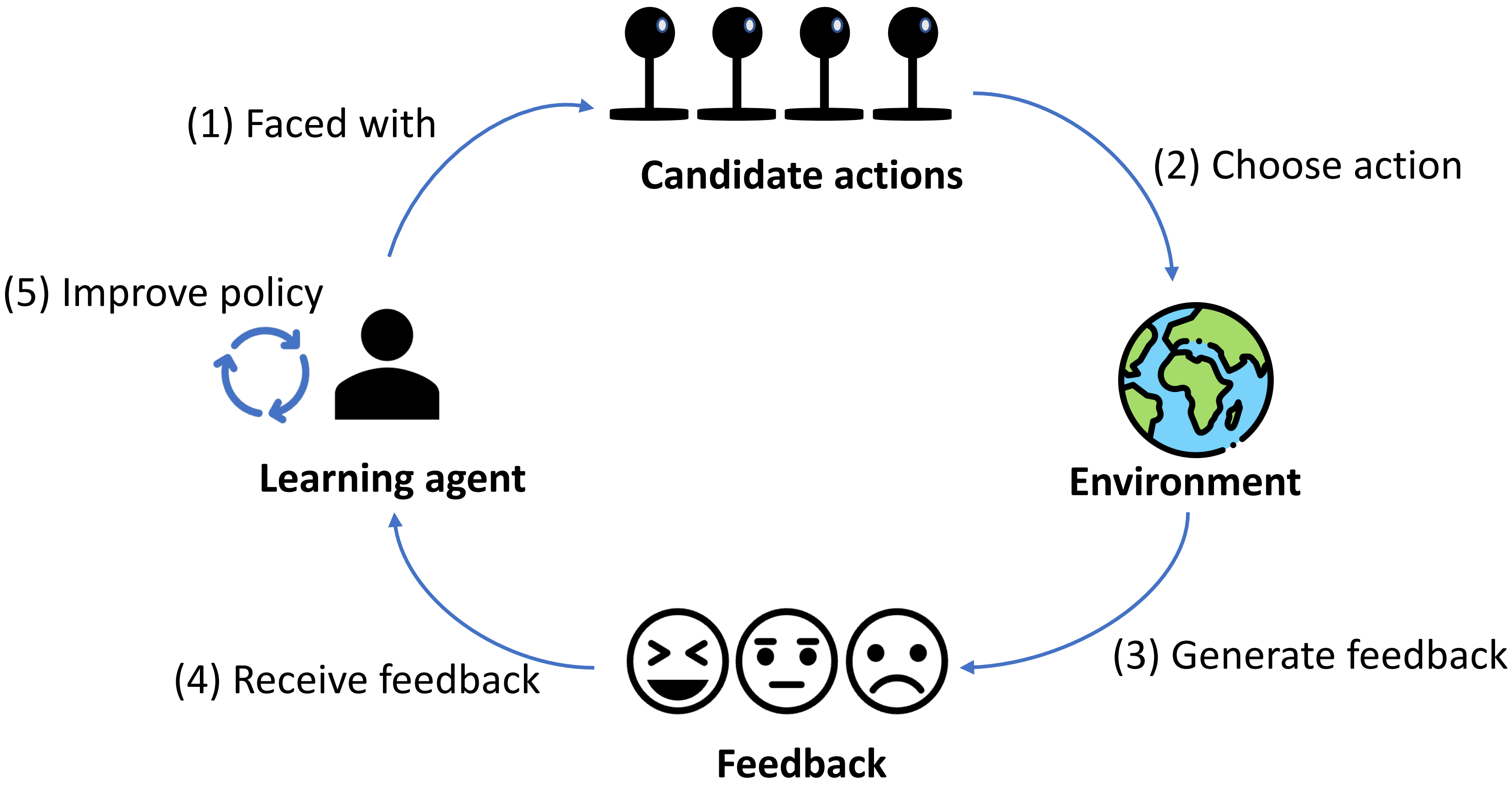
Time	1	2	3	4	5	6	7	8	9	10
Arm 1	\$1	\$0			\$1	\$1	\$0			
Arm 2			\$1	\$0						

To accumulate as many rewards, which arm would you choose next?

Select arms with higher rewards to accumulate more rewards ⟵ Exploitation V.S. Exploration ⟶ Select arms with less-observed times to learn the unknown knowledge

Over exploration leads to high costs
Insufficient exploration prevents from finding the optimal arm

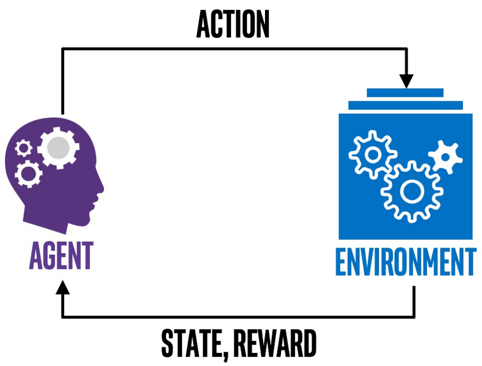
Interactive machine learning



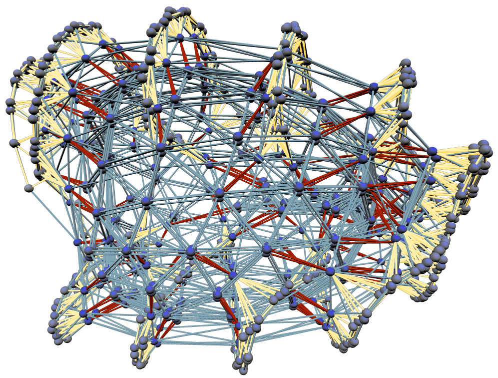
Applications

Recommendation systems
[Li et al., 2010]

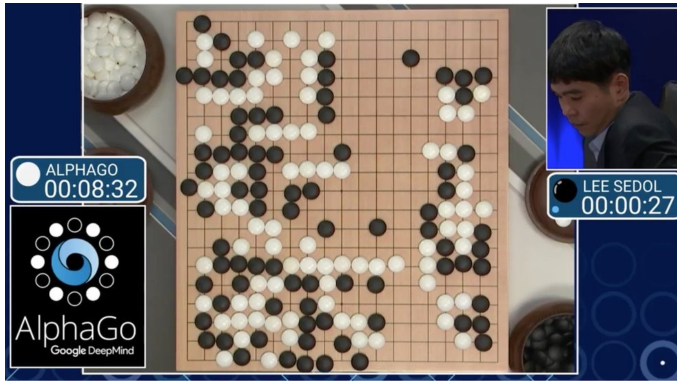
Advertisement placement
[Yu et al., 2016]



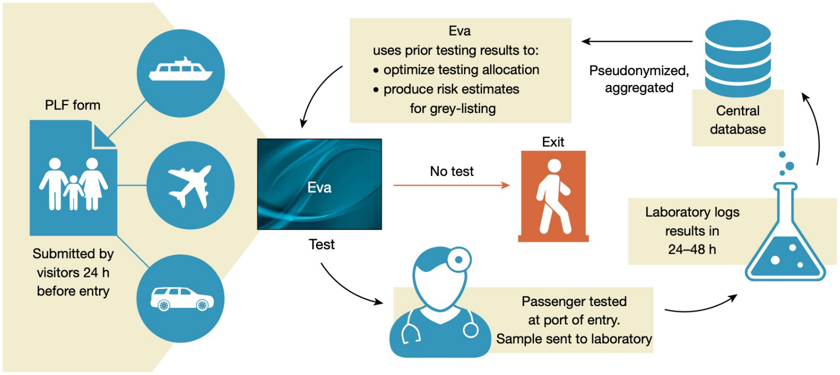
Key part of reinforcement learning
[Hu et al., 2018]



SAT solvers
[Liang et al., 2016]

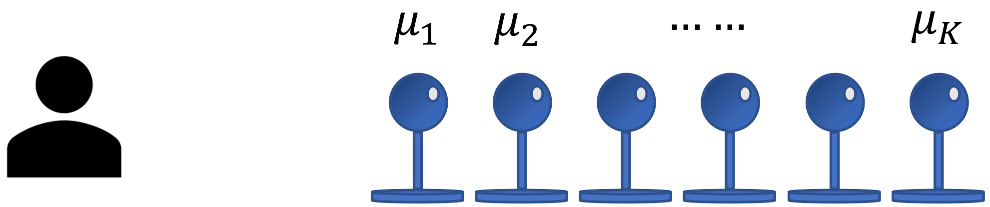


Monte-carlo Tree Search (MCTS) in AlphaGo
[Kocsis and Szepesvári, 2006; Silver et al., 2016]



Public health: COVID-19 border testing in Greece
[Bastani et al., 2021]

Multi-armed bandits (MAB)



- A player and K arms

  Items, products, movies, companies, …
- Each arm a_j has an unknown reward distribution P_j with unknown mean μ_j

  CTR, preference value, …

- In each round $t = 1,2,\ldots$:
  - The agent selects an arm $A_t \in \{1,2,\ldots,K\}$
  - Observes reward $X_t \sim P_{A_t}$

    Click information, satisfaction, …

Assume P_j is supported on [0,1]

Objective

- Maximize the expected cumulative reward in T rounds

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right]$$

- Minimize the regret in T rounds
  - Denote $j^* \in \mathrm{argmax}_j\ \mu_j$ as the best arm

$$Reg(T) = T \cdot \mu_{j^*} - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right]$$

Explore-then-commit (ETC) [Garivier et al., 2016]

- There are $K = 2$ arms (choices/plans/...)
- Suppose
  - $\mu_1 > \mu_2$
  - $\Delta = \mu_1 - \mu_2$

A/B testing

- Explore-then-commit (ETC) algorithm
  - Select each arm h times
  - Find the empirically best arm A
  - Choose $A_t = A$ for all remaining rounds

h rounds for a_1

h rounds for a_2

$T - 2h$ rounds for the better performed one

Explore-then-commit (cont.)

h rounds for a_1 | h rounds for a_2 | $T - 2h$ rounds for the better performed one

- Regret analysis:

$$Reg(T) = T \cdot \mu_1 - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right]$$
$$= h\Delta + (T - 2h) \cdot \Delta \cdot \mathbb{P}(\hat{\mu}_1 < \hat{\mu}_2)$$

Sample mean

$$= h\Delta + (T - 2h) \cdot \Delta \cdot \mathbb{P}\left((\hat{\mu}_2 - \mu_2) - (\hat{\mu}_1 - \mu_1) > \Delta\right)$$
$$\leq \underbrace{h\Delta}_{\text{Exploration}} + \underbrace{T \cdot \Delta \cdot \exp\left(-\frac{h\Delta^2}{4}\right)}_{\text{Exploitation}}$$

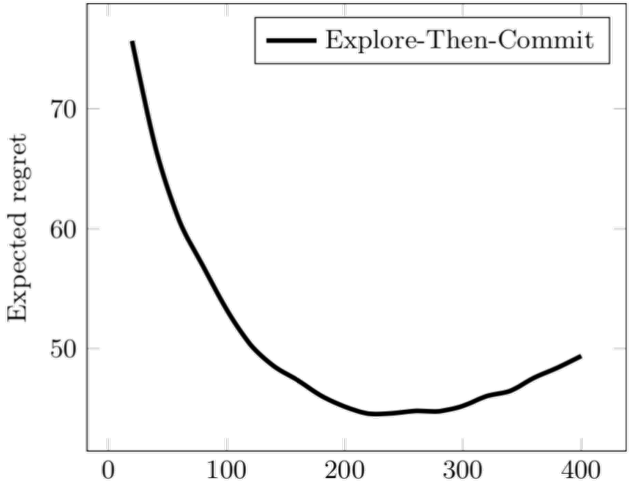
Hoeffding's inequality

$$\leq O\left(\frac{\log T}{\Delta}\right)$$

Choose $h = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{T\Delta^2}{4}\right) \right\rceil$

require the knowledge of Δ

- $Reg(T) = \Omega(T\Delta)$ if $h = 100$
- $Reg(T) = \Omega(T\Delta)$ if $h = T/10$



Only with the best choice of h the regret would be smallest

A soft version: ε-greedy

- For each round t
  - $\varepsilon_t \in (0,1)$
  - With probability ε_t, exploration (uniformly random select arms)
  - With probability $1 - \varepsilon_t$, exploitation (select the best performed arm so far)
- When $\varepsilon_t = \min\left\{1, \frac{c}{t\Delta^2}\right\}$, $Reg(T) = O\left(\frac{\log T}{\Delta}\right)$

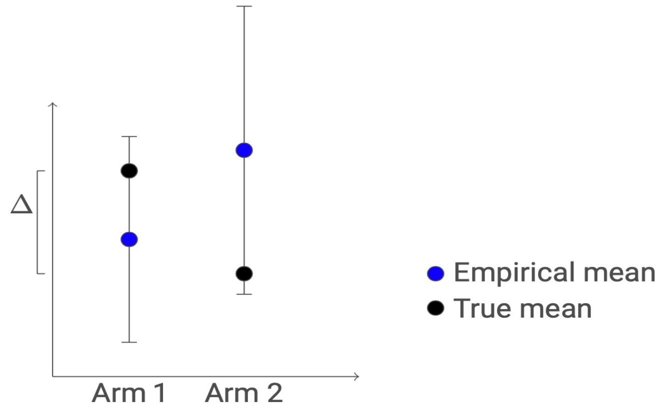
Demo: https://cse442-17f.github.io/LinUCB/

Upper confidence bound (UCB) [Auer et al., 2002]

- With high probability $\geq 1-\delta$ By Hoeffding's inequality

$$\mu_j \in \left[\hat{\mu}_j - \sqrt{\frac{\log 1/\delta}{T_j}}, \hat{\mu}_j + \sqrt{\frac{\log 1/\delta}{T_j}}\right]$$

Sample mean Number of selections of a_j



- Optimism: Believe arms have higher rewards, encourage exploration
  - The UCB value represents the reward estimates
- For each round t, select the arm

$$A(t) \in \operatorname{argmax}_{j\in[K]} \left\{ \underbrace{\hat{\mu}_j}_{\text{Exploitation}} + \overbrace{\underbrace{\sqrt{\frac{\log 1/\delta}{T_j(t)}}}_{\text{Exploration}}}^{\text{Upper confidence bound (UCB)}} \right\}$$

Upper confidence bound (UCB) (cont.)

- Assume arm a_1 is the best arm
- If sub-optimal arm a_j is selected
  - w/ high probability

    $$\mu_1 \leq \mathrm{UCB}_1 \leq \mathrm{UCB}_j \leq \mu_j + 2\sqrt{\frac{\log 1/\delta}{T_j(t)}}$$

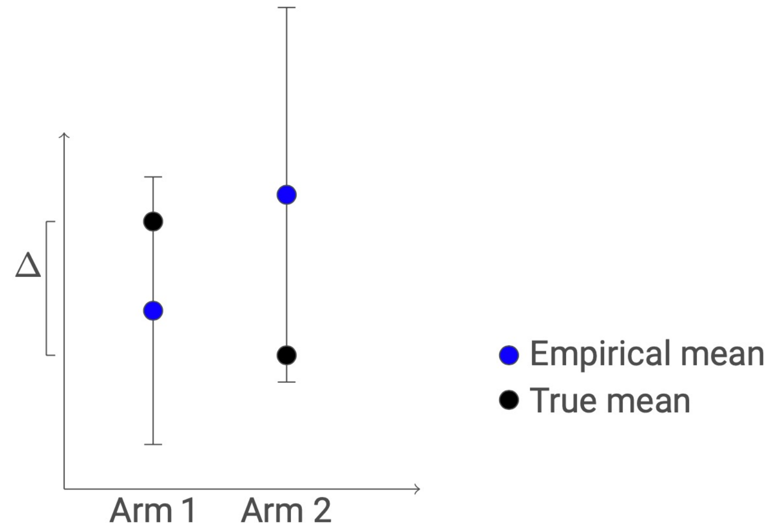
  - $\Longrightarrow 2\sqrt{\frac{\log 1/\delta}{T_j(t)}} \geq \Delta_j := \mu_1 - \mu_j$
  - $\Longrightarrow T_j(t) \leq O\left(\frac{\log 1/\delta}{\Delta_j^2}\right)$

Can choose δ adaptive to time t

- By choosing $\delta = 1/T$, cumulative regret:

$$O\left(\sum_{j\neq 1}\frac{\log T}{\Delta_j^2}\cdot\Delta_j\right) = O(K\log T/\Delta)$$

$\Delta := \min_{j\neq 1}\Delta_j$
Without knowing Δ



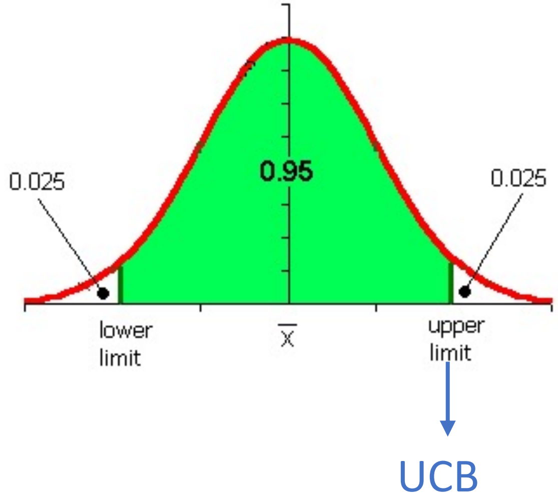
Thompson sampling (TS) [Agrawal and Goyal, 2013]

- Assume each arm has prior $\mathrm{Gaussian}(0,1)$
- Sample an estimate $\tilde{\mu}_j$ from the posterior distribution

$$\tilde{\mu}_j \sim \mathrm{Gaussian}\left(\hat{\mu}_j, \frac{1}{1+T_j(t)}\right)$$

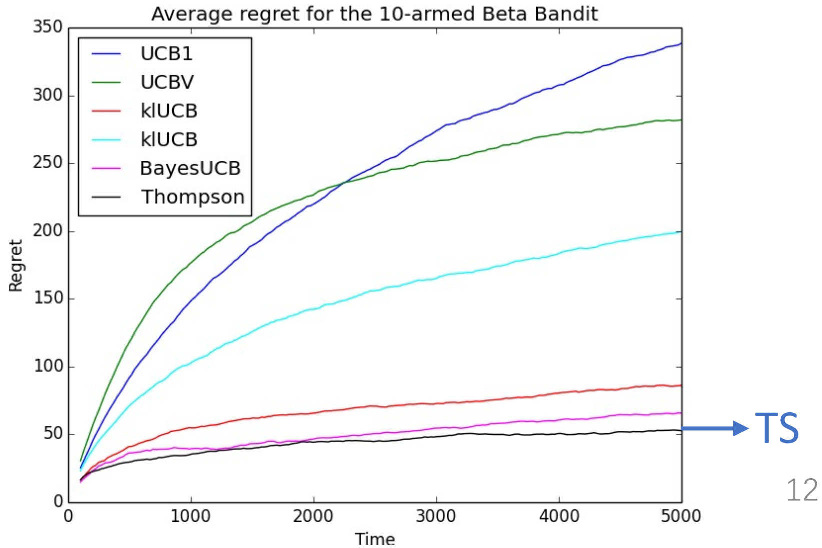
Exploitation

Exploration



- Select the arm $A(t) \in \mathrm{argmax}_{j\in[K]}\tilde{\mu}_j$

- Also have $O(K \log T/\Delta)$ regret
- Usually outperforms UCB



https://learnforeverlearn.com/bandits/

References I

- Lattimore, Tor, and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.
- Li, Lihong, et al. "A contextual-bandit approach to personalized news article recommendation." International conference on World wide web. 2010.
- Kocsis, Levente, and Csaba Szepesvári. "Bandit based monte-carlo planning." European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
- Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." nature 529.7587 (2016): 484-489.
- Yu, Baosheng, Meng Fang, and Dacheng Tao. "Linear submodular bandits with a knapsack constraint." Proceedings of the AAAI Conference on Artificial Intelligence. 2016.
- Liang, Jia Hui, et al. "Learning rate based branching heuristic for SAT solvers." Theory and Applications of Satisfiability Testing–SAT 2016: 19th International Conference, Bordeaux, France, July 5-8, 2016, Proceedings 19. Springer International Publishing, 2016.
- Bastani, Hamsa, et al. "Efficient and targeted COVID-19 border testing via reinforcement learning." Nature 599.7883 (2021): 108-113.

References II

- Hu, Yujing, et al. "Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application." Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018.
- Garivier, Aurélien, Tor Lattimore, and Emilie Kaufmann. "On explore-then-commit strategies." Advances in Neural Information Processing Systems 29 (2016).
- Audibert, Jean-Yves, and Sébastien Bubeck. "Best arm identification in multi-armed bandits." COLT-23th Conference on learning theory-2010. 2010.
- Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem." Machine learning 47 (2002): 235-256.
- Agrawal, Shipra, and Navin Goyal. "Further Optimal Regret Bounds For Thompson Sampling." Sixteenth International Conference on Artificial Intelligence and Statistics. 2013.
- Lai, Tze Leung, and Herbert Robbins. "Asymptotically efficient adaptive allocation rules." Advances in applied mathematics 6.1 (1985): 4-22.