

# Multi-armed Bandits

Shuai Li

John Hopcroft Center, Shanghai Jiao Tong University

<https://shuaili8.github.io>

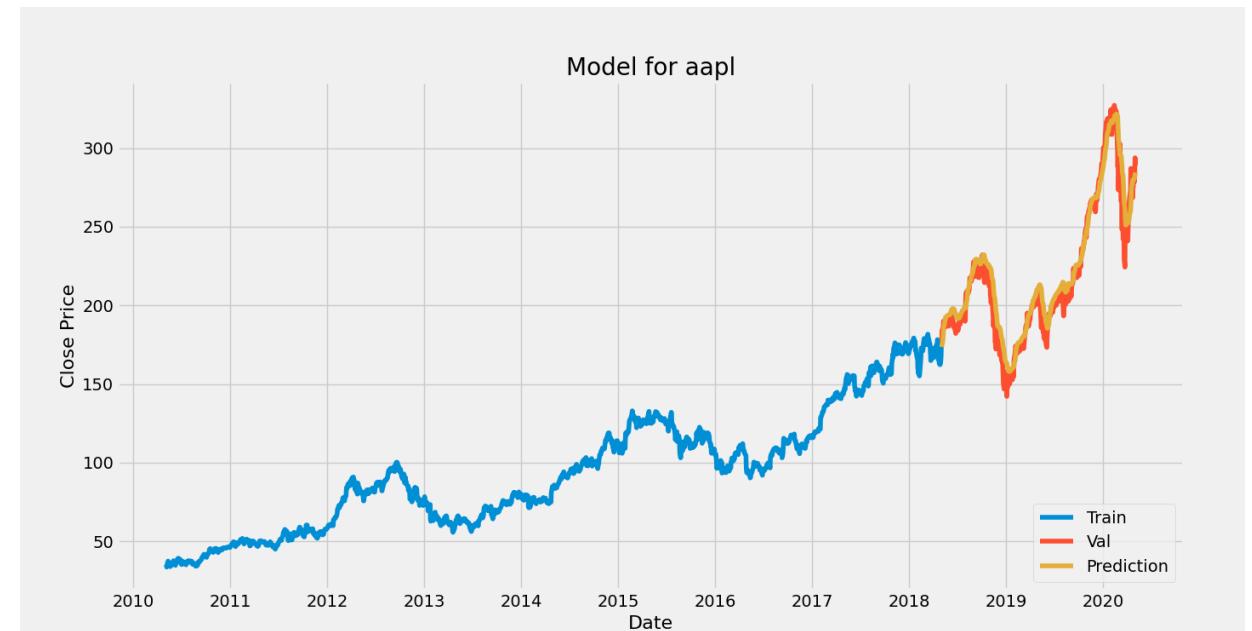
<https://shuaili8.github.io>

# Self introduction

- Position
  - Tenure-track associate professor 2023-now
  - Assistant professor at John Hopcroft Center Aug 2019 - 2022
- Education
  - PhD in Computer Science from the Chinese University of Hong Kong
  - Master in Math from the Chinese Academy of Sciences
  - Bachelor in Math from Chu Kochen Honors College, Zhejiang University
- Research interests
  - Bandit algorithms
  - Reinforcement learning algorithms
  - Machine learning theory
  - Machine learning applications

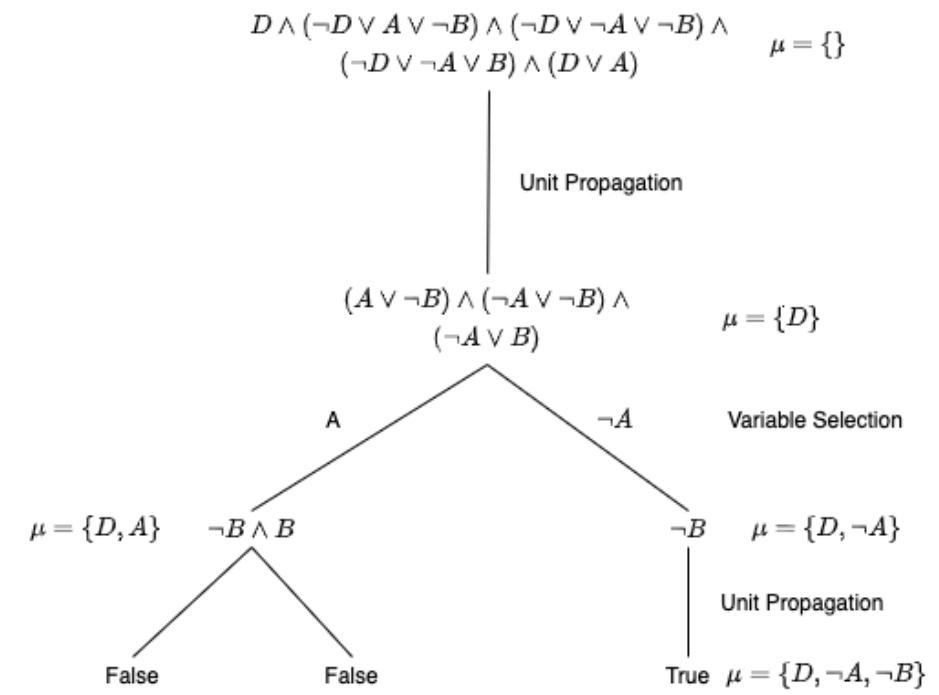
# Online Learning w/ Full Information

- Can observe feedback of every action



# Online Learning w/ Bandit Feedback

- Can only observe feedback for the selected action



# Bandits



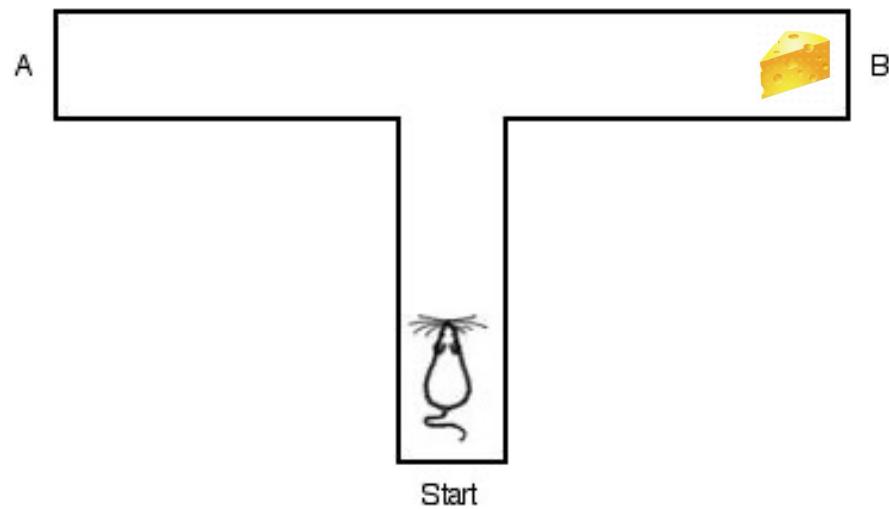
<i>Time</i>	1	2	3	4	5	6	7	8	9	10	11	12
<i>Left arm</i>	\$1	\$0			\$1	\$1	\$0					
<i>Right arm</i>				\$1	\$0							

- Five rounds to go. Which arm would you choose next?

What are bandits, and why  
should you care

# What's in the name?

- First bandit algorithm proposed by [Thompson \(1933\)](#)



- [Bush and Mosteller \(1953\)](#) were interested in how mice behaved in a T-maze

# Why care about bandits?

- Many applications
- They isolate an important component of reinforcement learning: exploration-vs-exploitation
- Theoretically guaranteed algorithms
- Rich and beautiful mathematics

# Applications: Recommendation systems

- Yahoo news [Li et al. (2010)]

The screenshot shows the Yahoo News homepage with a navigation bar at the top labeled "Featured", "Entertainment", "Sports", and "Life". The main feature is a large story titled "McNair's final hours revealed STORY" with a subtext about police releasing 50 text messages. Below this, there are four smaller featured stories: F1 (Steve McNair's final hours revealed), F2 (Cindy Crawford stays fierce in a black mini), F3 (Watch for dozens of 'shooting stars' tonight), and F4 (At team's big moment, star player isn't around). At the bottom right, there is a link "» More: Featured | Buzz".

**Featured** Entertainment | Sports | Life

**McNair's final hours revealed STORY**

Police release 50 text messages that depict the late NFL player's alleged killer as losing control. [» Details](#)

- UConn murder victim mourned

[Find Steve McNair murder case](#)

**F1** Steve McNair's final hours revealed

**F2** Cindy Crawford stays fierce in a black mini

**F3** Watch for dozens of 'shooting stars' tonight

**F4** At team's big moment, star player isn't around

» More: [Featured](#) | [Buzz](#)

# Applications: A part of RL

- A way of isolating an interesting part of reinforcement learning
  - Recommending items  
[\[Hu et al. \(2018\)\]](#)
  - Achieved more than 30% growth in GMV

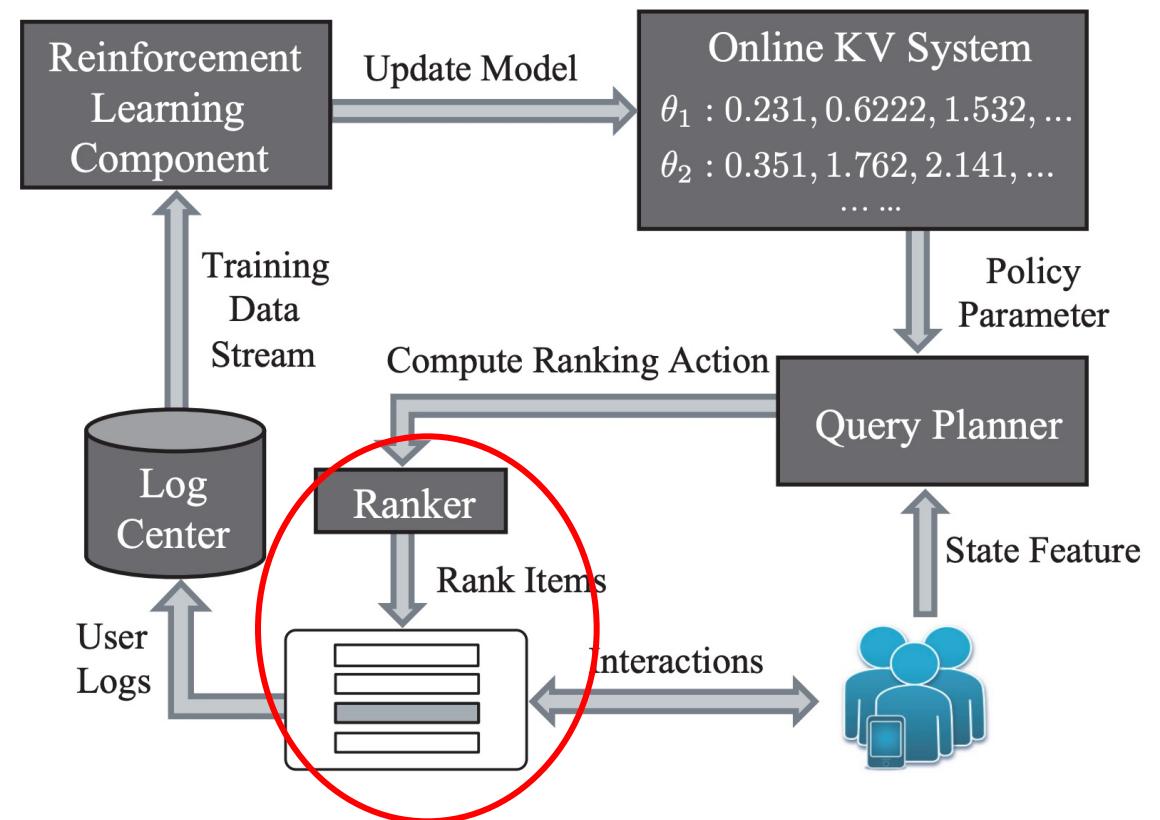
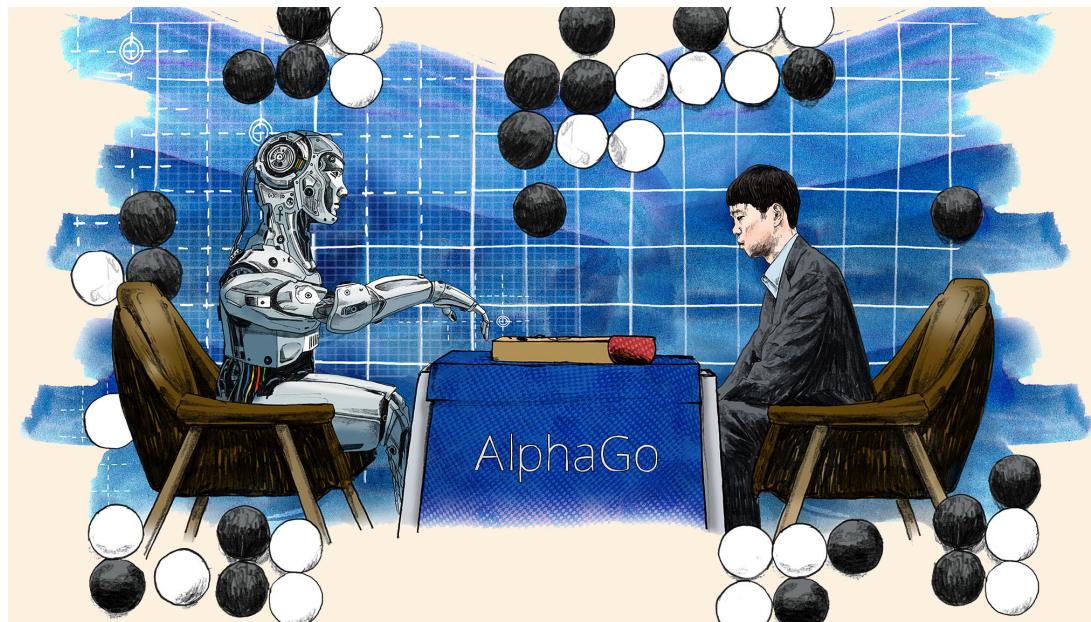


Figure 6: RL ranking system of *TaoBao* search engine

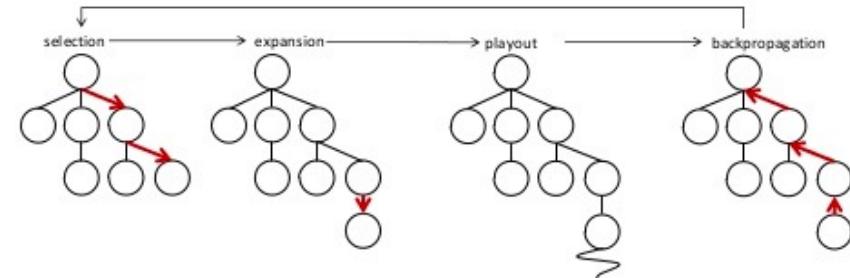
# Applications: A component of game-playing

- A component of game-playing algorithms
  - Monte-Carlo tree search (MCTS) – AlphaGo [[Silver et al. \(2016\)](#)]
  - UCT algorithm [[Kocsis and Szepesvari \(2006\)](#)]
  - Drives its search uses a bandit algorithm at each node



## UCT

- Repeat Selection → Expansion → Playout → Backpropagation until
  - Reaching the predefined maximum time-length or the maximum number of playouts
- Use **UCB1 value** in Selection
- Finally select the action associated with the adjacent child node, of the root node, having **maximum number of visits**



# Applications: Select policies

- Select the best ranking policy [Yue et al. (2012)]
  - Online interleaving



Showing 1–50 of 1,299 results for all: bandit

bandit

Show abstracts  Hide abstracts

50 results per page. Sort results by Announcement date (newest first) Go

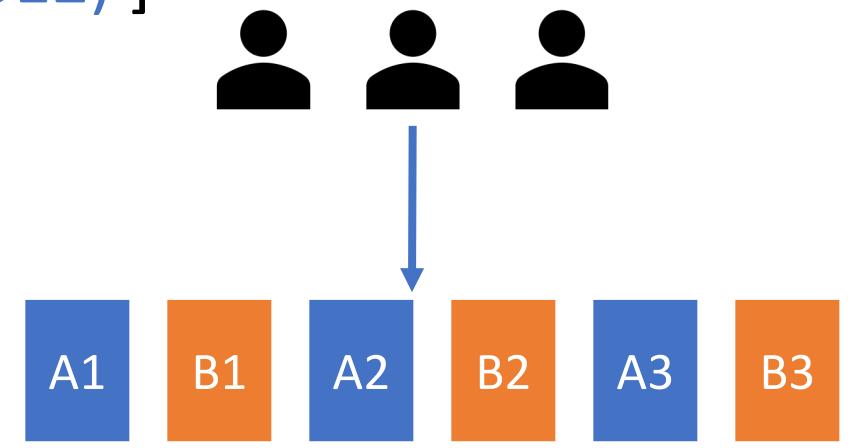
1 2 3 4 5 ...

1. arXiv:1908.06256 [pdf, other] cs.LG stat.ML  
A Batched Multi-Armed Bandit Approach to News Headline Testing  
Authors: Yizhi Mao, Miao Chen, Abhinav Wagle, Junwei Pan, Michael Natkovich, Don Matheson  
Submitted 17 August, 2019; originally announced August 2019.  
Comments: IEEE BigData, 2018

2. arXiv:1908.06158 [pdf, other] cs.IR cs.LG  
Accelerated learning from recommender systems using multi-armed bandit  
Authors: Meisam Hejazinia, Kyler Eastman, Shuqin Ye, Abbas Amirabadi, Ravi Divvela  
Submitted 16 August, 2019; originally announced August 2019.

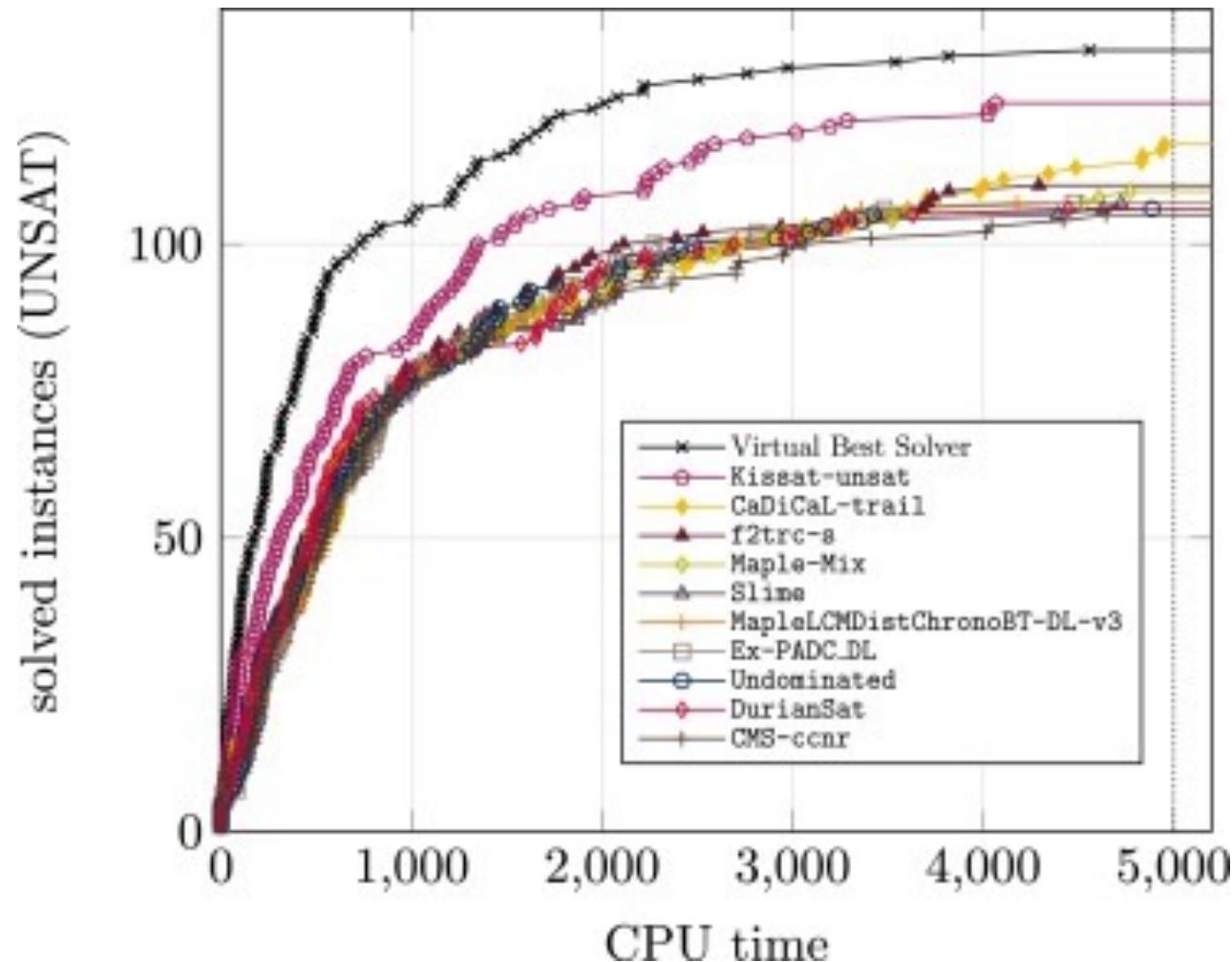
3. arXiv:1908.05814 [pdf, other] cs.LG stat.ML  
Linear Stochastic Bandits Under Safety Constraints  
Authors: Sanae Amani, Mahnoosh Alizadeh, Christos Thrampoulidis  
Submitted 15 August, 2019; originally announced August 2019.  
Comments: 23 pages, 7 figures

4. arXiv:1908.05531 [pdf, other] math.ST  
Exponential two-armed bandit problem  
Authors: Alexander Kolnogorov, Denis Grunew  
Submitted 15 August, 2019; originally announced August 2019



Combined Ranker A&B

# Applications: SAT solvers



# Other applications

- Clinical trials [[Villar et al. \(2015\)](#)]
- Network routing [[Le et al. \(2014\)](#)]
- Experimental design [[Rafferty et al. \(2018\)](#)]
- Hyperparameter tuning [[Li et al. \(2017\)](#)]
- A/B testing [many]
- Ad placement [[Yu et al. \(2016\)](#)]
- Dynamic pricing (eg.,for Amazon products) [[Babaioff et al. \(2015\)](#)]
- Ranking (eg.,for search) [[Radlinski et al. \(2008\)](#)]
- Waiting problems (when to auto-logout your computer) [[Lattimore et al. \(2014\)](#)]
- Resource allocation [[Larrnaaga et al. \(2016\)](#)]

# Finite-armed stochastic bandits

# Setting: Finite-armed stochastic bandits

items/products/movies/news/...

- There are  $L$  arms
  - Each arm  $a$  has an unknown reward distribution  $v_a$  with unknown mean  $\alpha(a)$
  - The best arm is  $a^* = \operatorname{argmax}_a \alpha(a)$



CTR/profit/...

- At each time  $t$ 
  - The learning agent selects an arm  $a_t$
  - Observes the reward  $X_{a_t,t} \sim v_{a_t}$

bandit feedback

# Objective

- Maximize the expected cumulative reward in  $T$  rounds

$$\mathbb{E} \left[ \sum_{t=1}^T \alpha(a_t) \right]$$

- Minimize the **regret** in  $T$  rounds

$$R(T) = T \cdot \alpha(a^*) - \mathbb{E} \left[ \sum_{t=1}^T \alpha(a_t) \right]$$

- Balance the trade-off between **exploration** and **exploitation**
  - Exploitation: Select arms that yield good results so far
  - Exploration: Select arms that have not been tried much before
- Smaller order of  $T$  in  $R(T)$  is better

# A/B testing

- There are  $L = 2$  arms (choices/plans/...)
- Suppose

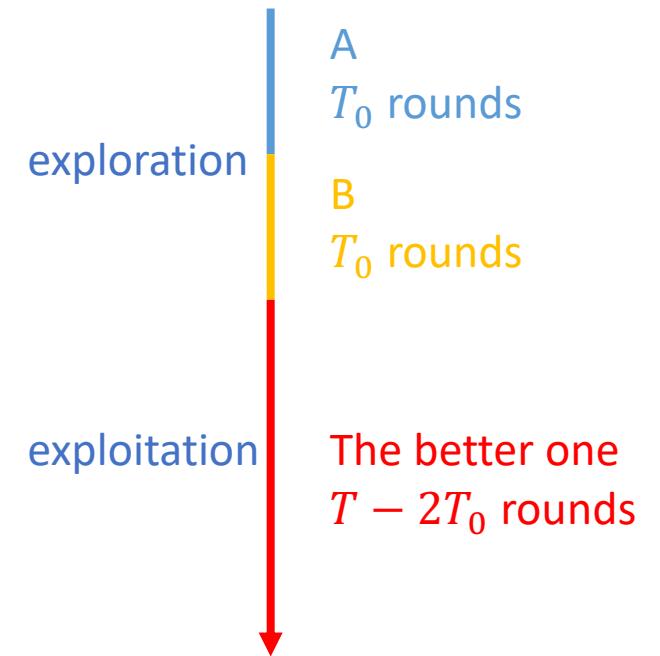
$$v_A = \text{Gaussian}(\alpha_A, 1)$$

$$v_B = \text{Gaussian}(\alpha_B, 1)$$

$$\alpha_A > \alpha_B,$$

$$\Delta = \alpha_A - \alpha_B$$

- Explore-then-commit algorithm
  - Select each of A and B for  $T_0$  rounds and then select the one with larger sample mean for the remaining  $T - 2T_0$  rounds



# A/B testing (continued)

- Regret

$$R(T)$$

$$= T_0 \cdot (\alpha_A - \alpha_B) + \mathbb{P}[\hat{\alpha}_A < \hat{\alpha}_B](T - 2T_0)(\alpha_A - \alpha_B)$$

$$< T_0 \Delta + T \Delta \cdot \exp\left(-\frac{T_0 \Delta^2}{4}\right)$$

$$= O\left(\frac{1}{\Delta} \log T\right)$$

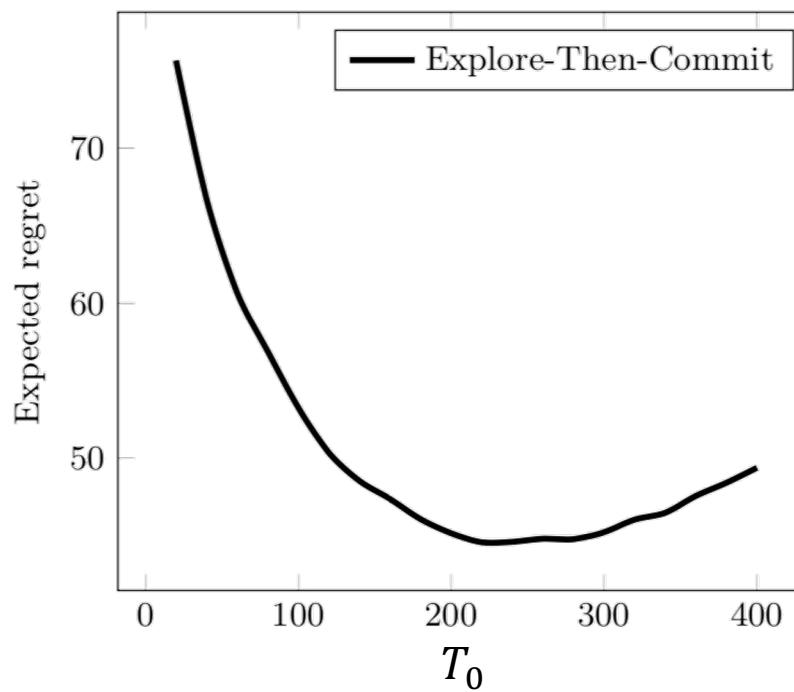
$$T_0 = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{T \Delta^2}{4}\right) \right\rceil$$

need the knowledge of  $\Delta$

- $R(T) = \Omega(T\Delta)$  if  $T_0 = \frac{1}{5}T$
- $R(T) = \Omega(T\Delta)$  if  $T_0 = 1000$



# A/B testing (continued)



**Figure 6.2** Expected regret for Explore-Then-Commit over  $10^5$  trials on a Gaussian bandit with means  $\mu_1 = 0, \mu_2 = -1/10$

- Lattimore and Szepesvári (2018)

# Epsilon-greedy algorithm

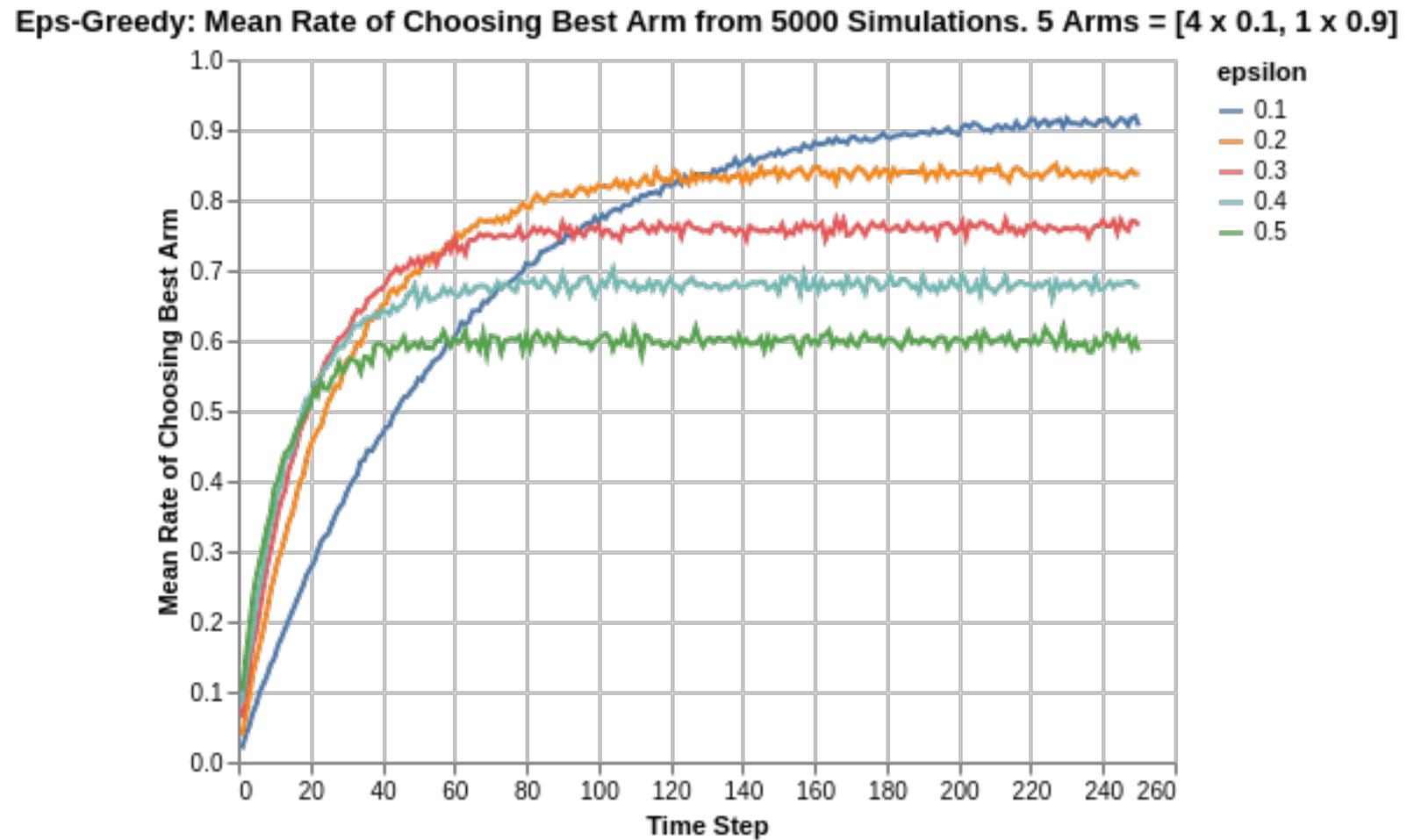
- For each time  $t$ 
  - $\epsilon_t \in (0,1)$
  - With probability  $\epsilon_t$ , randomly choose an arm
  - With probability  $1 - \epsilon_t$ , choose the one with highest sample mean
- When  $\epsilon_t = \min \left\{ 1, \frac{c}{t\Delta^2} \right\}$ , regret  $R(T) = O \left( \frac{L}{\Delta} \log T \right)$

exploration

exploitation

need the knowledge of  $\Delta$

# Epsilon-greedy algorithm 2



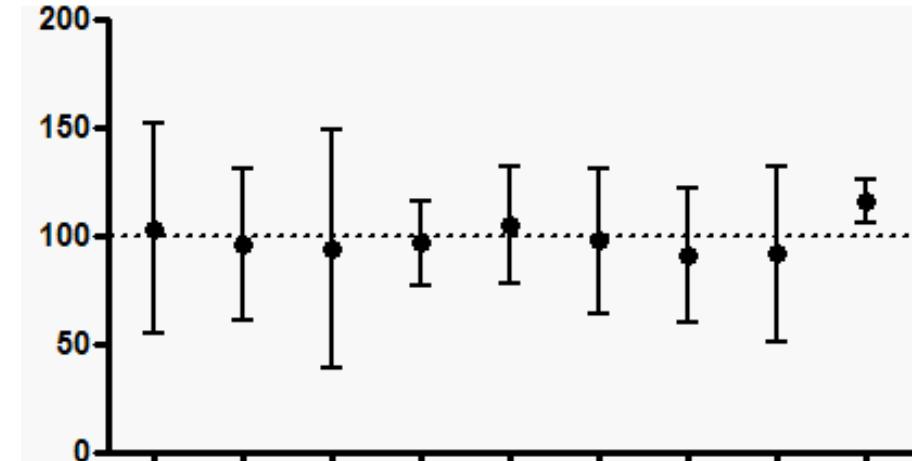
# UCB – Upper confidence bound [Auer et al.(2002)]

- With high probability

$$\alpha_a \in \left[ \hat{\alpha}_a(t) - \sqrt{\frac{2 \log t}{T_a(t)}}, \hat{\alpha}_a(t) + \sqrt{\frac{2 \log t}{T_a(t)}} \right]$$

sample mean  
round t

Hoeffding's inequality



selection times of arm a till round t

- Principle: optimism in face of uncertainty
- UCB policy:

$$a_t = \operatorname{argmax}_a \hat{\alpha}_a + \sqrt{\frac{2 \log t}{T_a(t)}}$$

exploitation

exploration

# UCB – Upper confidence bound 2

- Regret

$$R(T) = O\left(\frac{L}{\Delta} \log T\right)$$

- Proof sketch

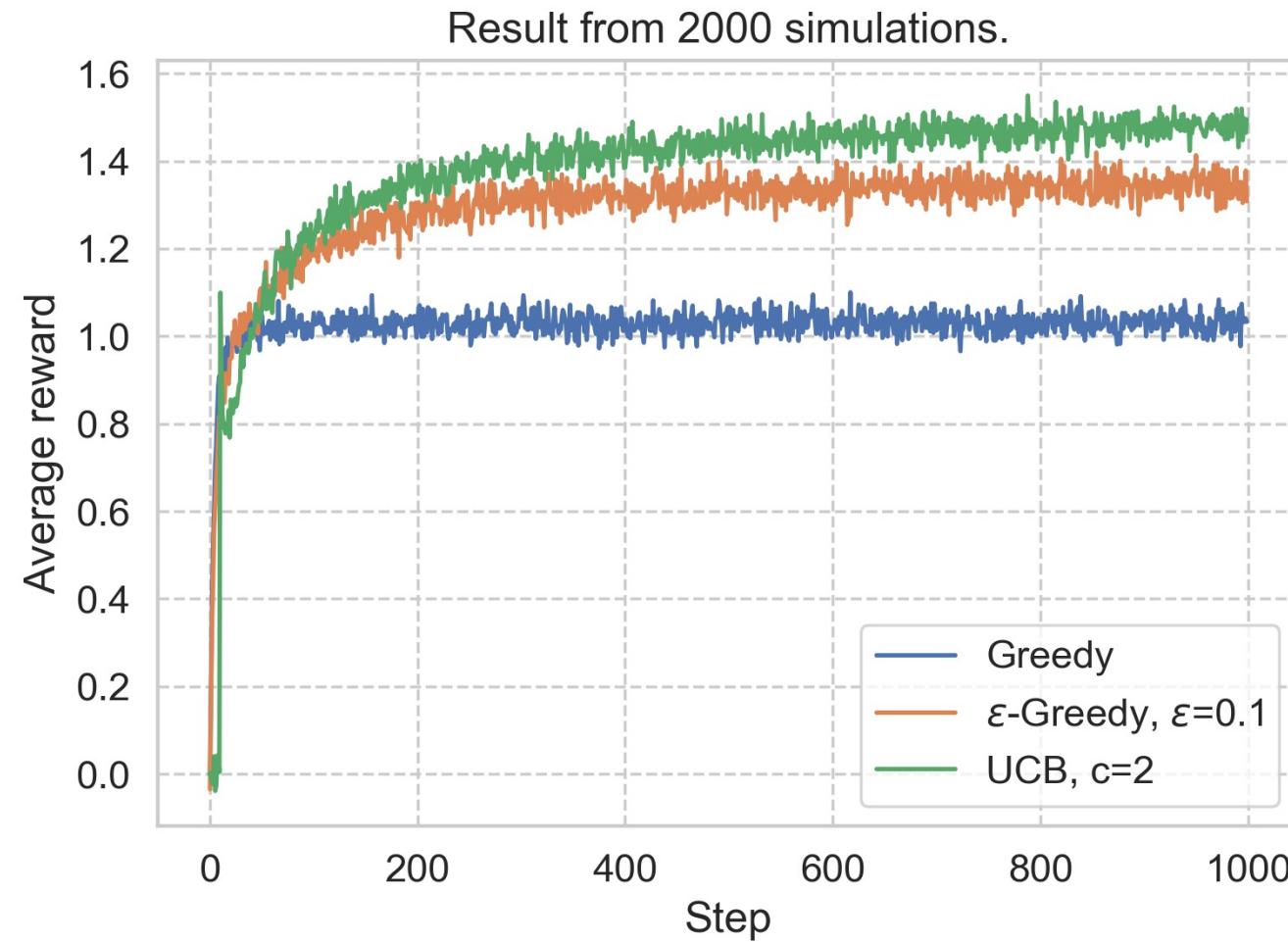
- Under good event (w/ high probability)
  - If arm  $a$  is pulled, then

$$\alpha(a^*) \leq \text{UCB}_{a^*} \leq \text{UCB}_a \leq \alpha(a) + 2 \text{ radius}_a$$

$$\bullet \Rightarrow \sqrt{\frac{2 \log t}{T_a(t)}} = \text{radius}_a \geq \frac{\alpha(a^*) - \alpha(a)}{2}$$

$$\bullet \Rightarrow T_a(t) \leq \frac{8 \log t}{\Delta_a^2}$$

# UCB – Upper confidence bound 3



# Thompson sampling [Agrawal and Goyal (2013)]

- Assume each arm has prior  $\text{Gaussian}(0,1)$
- Then posterior distribution for  $\alpha_a$  is

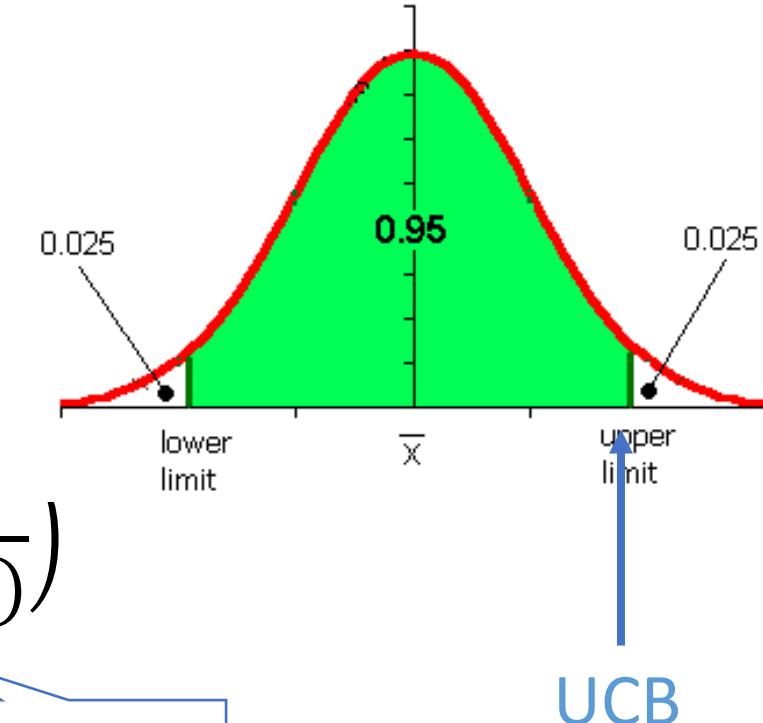
$$\text{Gaussian}\left(\hat{\alpha}_a(t), \frac{1}{1 + T_a(t)}\right)$$

- Sample

$$\tilde{\alpha}_a(t) \sim \text{Gaussian}\left(\hat{\alpha}_a(t), \frac{1}{1 + T_a(t)}\right)$$

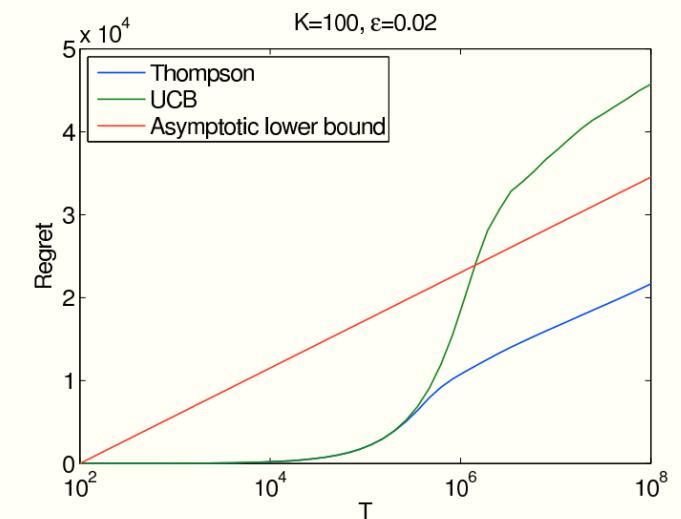
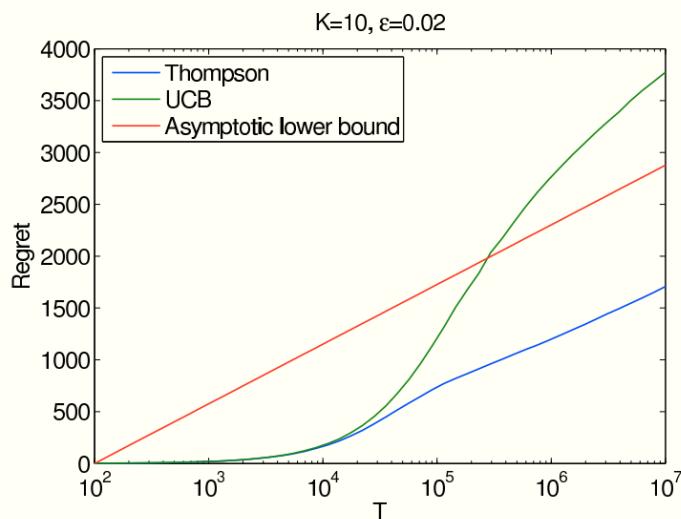
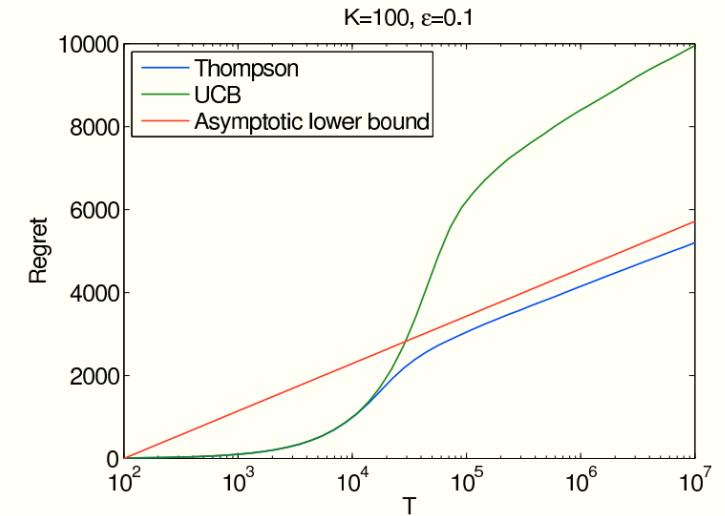
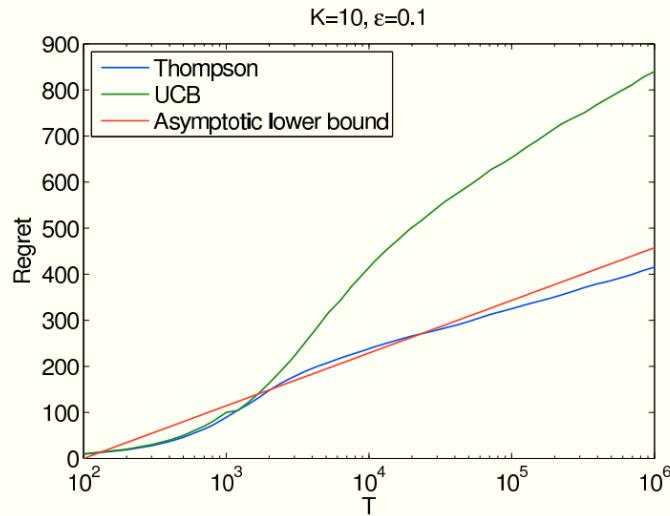
and select

$$a_t = \operatorname{argmax}_a \tilde{\alpha}_a(t)$$

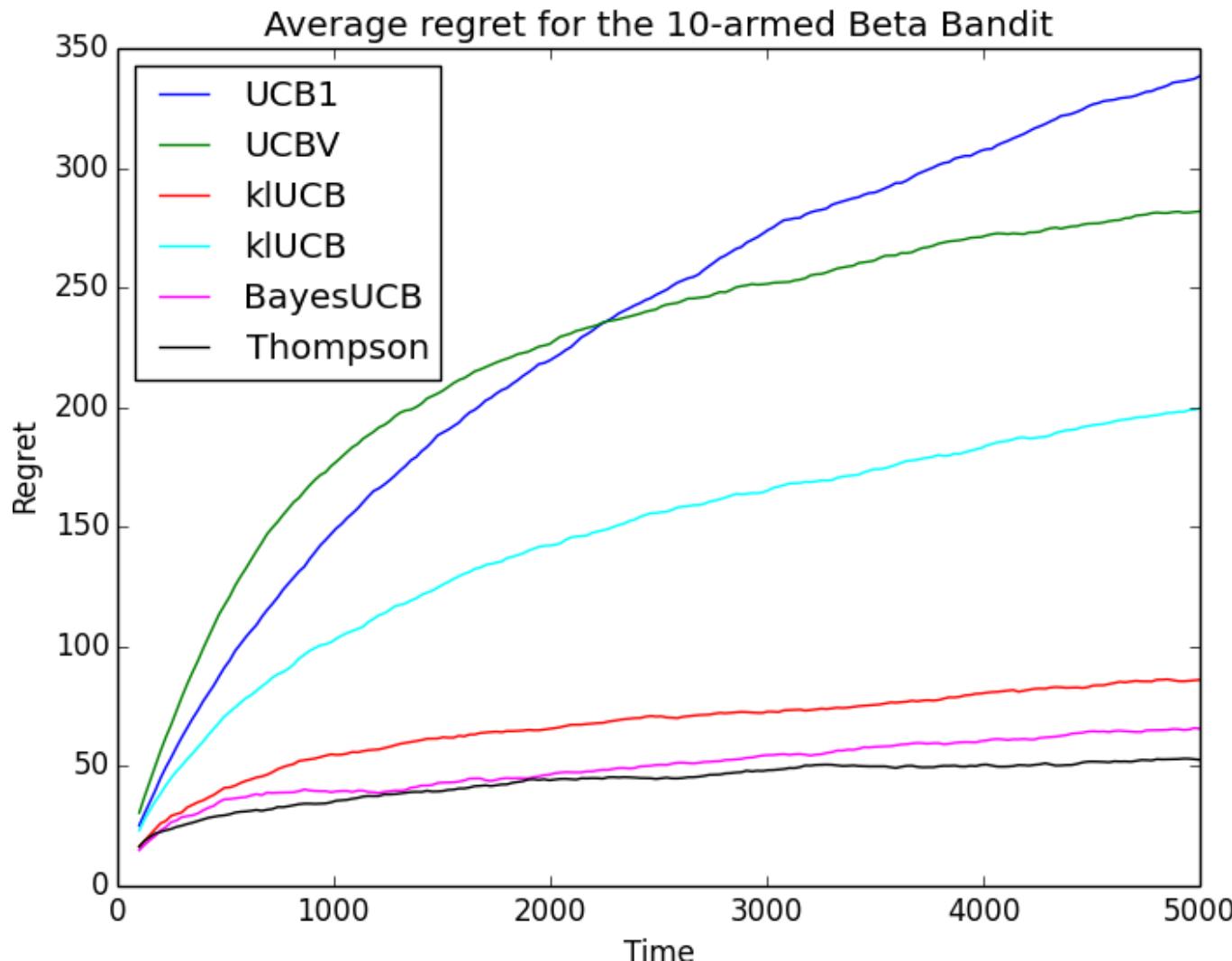


# Thompson sampling 2

- Also has optimal regret bound
- Outperform UCB  
[Chapelle and Li (2011)]



# Thompson sampling 3



# Finite-armed adversarial bandits

# Setting: Adversarial MAB

- There are  $L$  arms
  - An adversary secretly preselects all **loss** vectors  $\{l_{t,a}\}_{t,a}$  from  $[0,1]$
  - The best arm is  $a^* = \operatorname{argmin} \sum_{t=1}^T l_{t,a}$



# Setting: Adversarial MAB 2

- At each time  $t$ 
  - The learning agent selects one arm  $a_t$
  - Observe the loss  $l_{t,a_t}$
- Objective:
  - Minimize the expected cumulative loss in  $T$  rounds  $\mathbb{E}\left[\sum_{t=1}^T l_{t,a_t}\right]$
  - Minimize the **regret** in  $T$  rounds

$$R(T) = \mathbb{E}\left[\sum_{t=1}^T l_{t,a_t}\right] - \min_a \sum_{t=1}^T l_{t,a}$$

- Balance the trade-off between exploration and exploitation
  - Exploitation: Select arms that yield good results so far
  - Exploration: Select arms that have not been tried much before

# Exp3: Exponential Weight Algorithm for Exploration and Exploitation

- Importance-weight estimator

$$\hat{l}_{t,a} = \frac{\mathbb{I}\{a_t = a\} \cdot l_{t,a}}{\mathbb{P}(a_t = a)}$$

- For each time  $t$

- Calculate the sampling distribution

$$\mathbb{P}(a_t = a) = \frac{\exp(-\eta \hat{L}_{t-1,a})}{\sum_{b=1}^n \exp(-\eta \hat{L}_{t-1,b})}$$

Learning rate

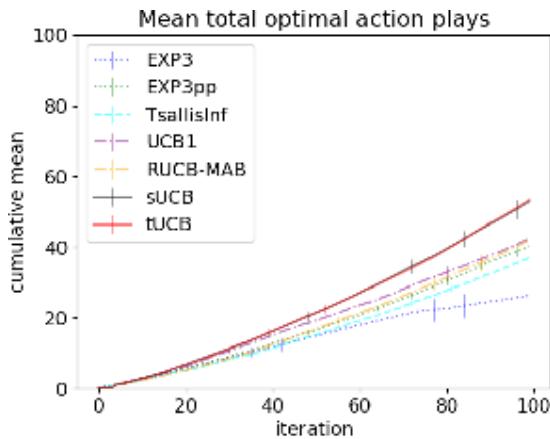
Exponential weighting

- Sample  $a_t \sim \mathbb{P}(a_t = a)$  and observe  $l_{t,a}$
    - Calculate  $\hat{L}_{t,a} = \sum_{s=1}^t \hat{l}_{t,a}$
  - Regret bound  $O(\sqrt{LT \log L})$

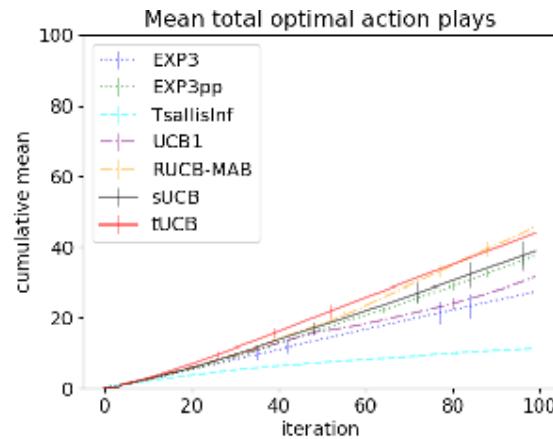
# Comparison between Stochastic and Adversarial Environments

- Stochastic
- Reward fixed distribution on  $[0,1]$  with fixed mean
- Best arm  $a^* = \operatorname{argmax} \alpha(a)$
- Regret bound  $O(\log T)$
- Runs in adversarial setting
  - Regret may not even converge
- Adversarial
- Loss arbitrary on  $[0,1]$
- Best arm  $a^* = \operatorname{argmin} \sum_{t=1}^T l_{t,a}$
- Regret bound  $O(\sqrt{T})$
- Runs in stochastic setting
  - Regret bound  $O(\sqrt{T})$

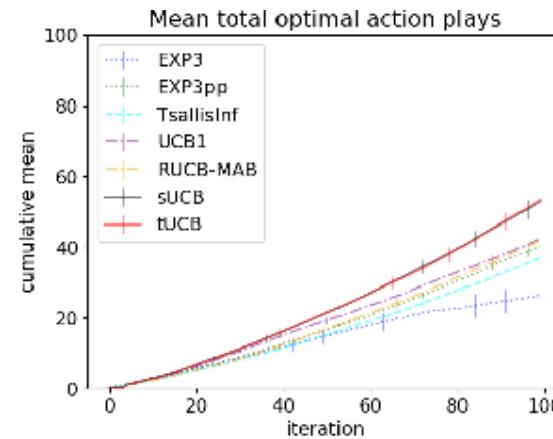
# Performance comparisons



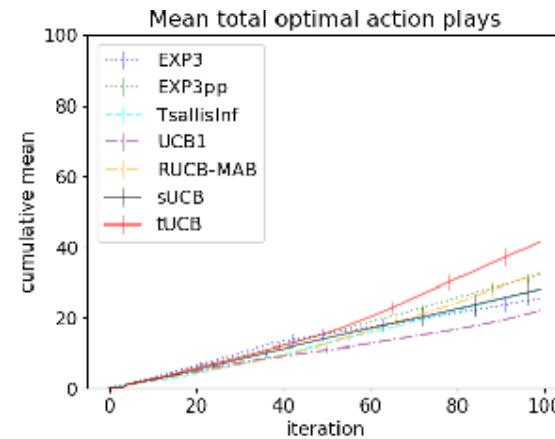
(a) No adversary



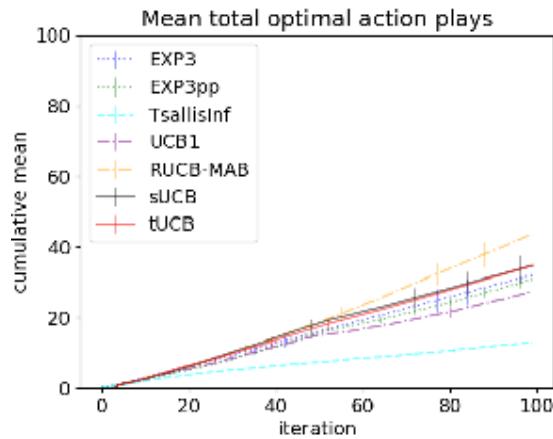
(b)  $\epsilon = 0.05$



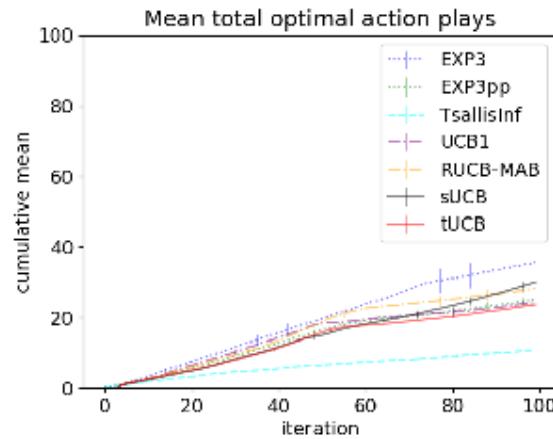
(a) No adversary



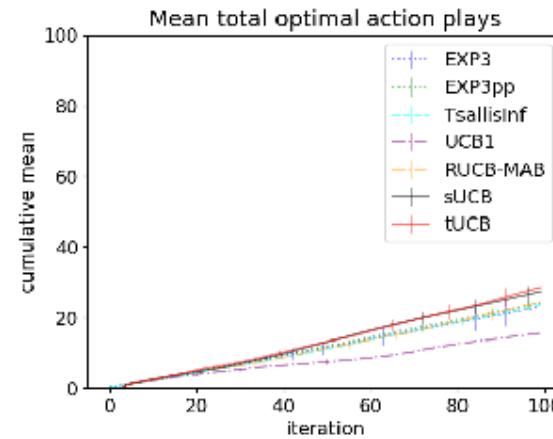
(b)  $\epsilon = 0.05$



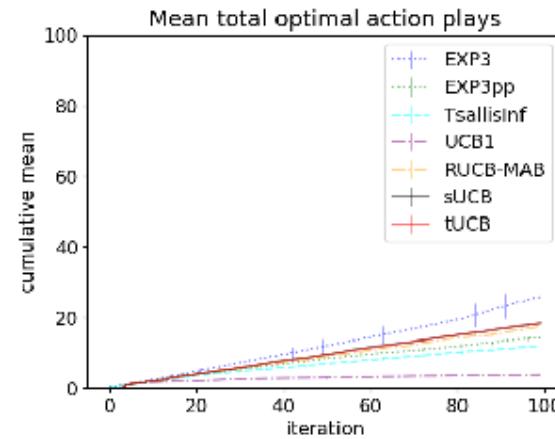
(c)  $\epsilon = 0.1$



(d)  $\epsilon = 0.3$



(c)  $\epsilon = 0.1$



(d)  $\epsilon = 0.3$

# Linear bandits

# Setting

- At each time  $t$

- The learning agent receives  $D_t \subset \mathbb{R}^d$
- Selects an arm  $a_t \in D_t \subset \mathbb{R}^d$
- Receives a random reward  $X_{a_t,t} \sim v_{a_t}$  with mean  
 $\alpha(a_t) = \theta^\top a_t$

for some fixed but unknown weight vector  $\theta$

a set of  
items/videos/news/...

bandit feedback w/  
linear structure

- Allow for large-scale applications

# LinUCB [Li et al. (2010)]

- The observed feedback is

$$\{(a_1, X_{a_1,1}), (a_2, X_{a_2,2}), \dots, (a_t, X_{a_t,t}), \dots\}$$

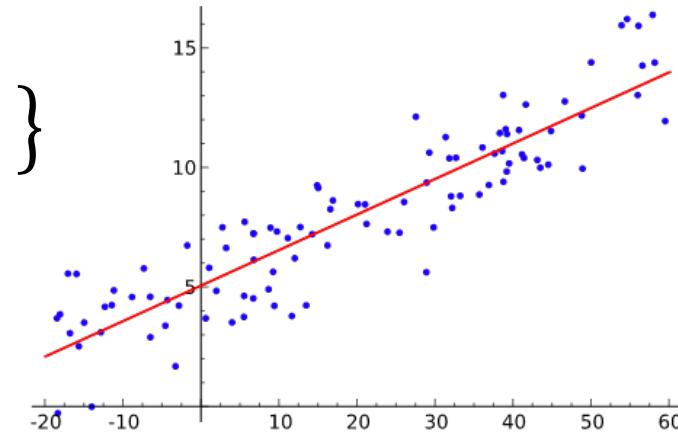
- Let

$$V_t = I + \sum_{s=1}^t a_s a_s^T, b_t = \sum_{s=1}^t X_{a_s,s} a_s$$

linear regression estimator

exploitation

$$\hat{\theta}_t = V_t^{-1} b_t$$



- With high probability

$$\|\theta - \hat{\theta}_t\|_{V_t} \leq C \sqrt{\log t}$$

# How to understand $\|\theta - \hat{\theta}_t\|_{V_t} \leq C\sqrt{\log t}$

- $\|x\|_{V_t} = \sqrt{x^\top V_t x}$
- If  $V_t = \begin{bmatrix} T_1 & \\ & T_2 \end{bmatrix}$ , then  $\|x\|_{V_t} = \sqrt{T_1|x_1|^2 + T_2|x_2|^2}$ 
  - When  $a_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, a_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

implies

$$V_t = I + \sum_{s=1}^t a_s a_s^T = \begin{bmatrix} 1 + T_{a_1}(t) & \\ & 1 + T_{a_2}(t) \end{bmatrix}$$

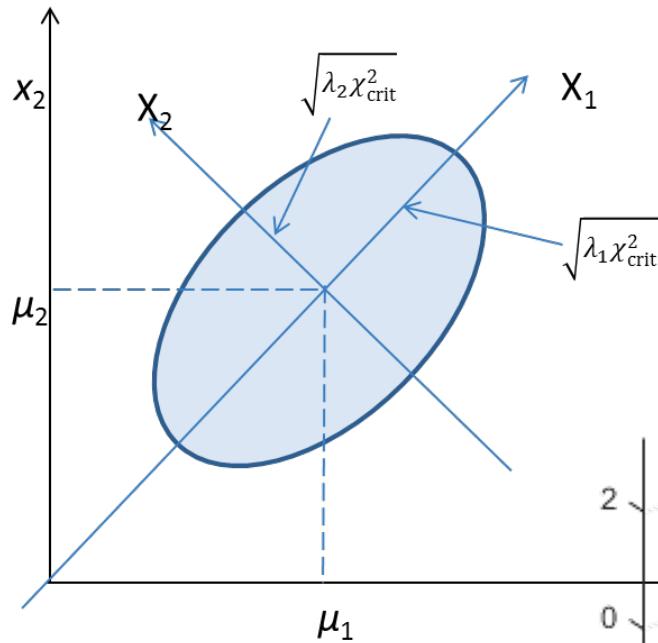
$$|\alpha(a_1) - \hat{\alpha}(a_1)| = |\theta_1 - \hat{\theta}_1| \leq C \sqrt{\frac{\log t}{1 + T_{a_1}(t)}}$$

$$|\alpha(a_2) - \hat{\alpha}(a_2)| = |\theta_2 - \hat{\theta}_2| \leq C \sqrt{\frac{\log t}{1 + T_{a_2}(t)}}$$

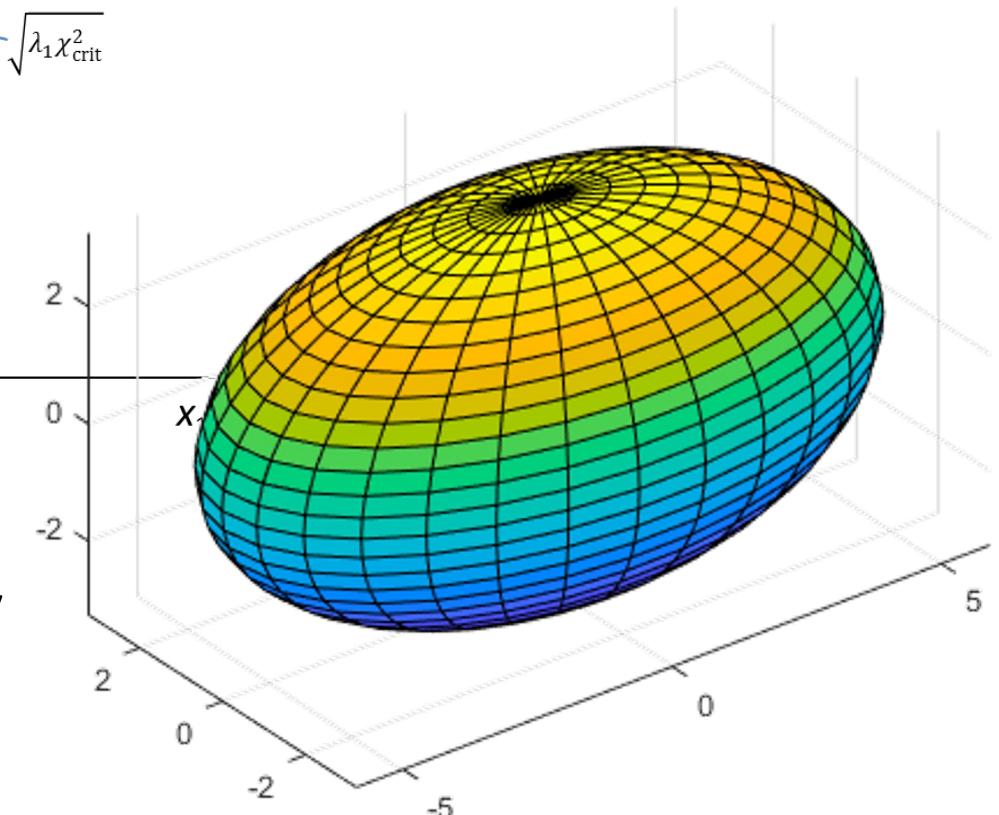


# How to understand $\|\theta - \hat{\theta}_t\|_{V_t} \leq C\sqrt{\log t}$

- For general  $V_t$
- Confidence ellipse



- Confidence ellipsoid
  - narrower direction means less uncertainty



# How to use it

- With high probability

$$|\alpha(a) - \hat{\theta}_t^\top a| \leq C\sqrt{\log t} \|a\|_{V_t^{-1}}$$

- Select

$$a_{t+1} = \operatorname{argmax}_a \hat{\theta}_t^\top a + C\sqrt{\log t} \|a\|_{V_t^{-1}}$$

The equation  $a_{t+1} = \operatorname{argmax}_a \hat{\theta}_t^\top a + C\sqrt{\log t} \|a\|_{V_t^{-1}}$  is shown with two blue arrows pointing from the terms  $\hat{\theta}_t^\top a$  and  $C\sqrt{\log t} \|a\|_{V_t^{-1}}$  to two separate blue boxes. The left box contains the word "exploitation" and the right box contains the word "exploration".

- Regret

$$R(T) = O(d\sqrt{T} \log T)$$

## LinTS [Agrawal and Goyal (2013b)]

- Suppose  $\theta$  has prior  $\text{Gaussian}(0, I)$

- Then the posterior distribution for  $\theta$  is

$$\text{Gaussian}(\hat{\theta}, V_t^{-1})$$

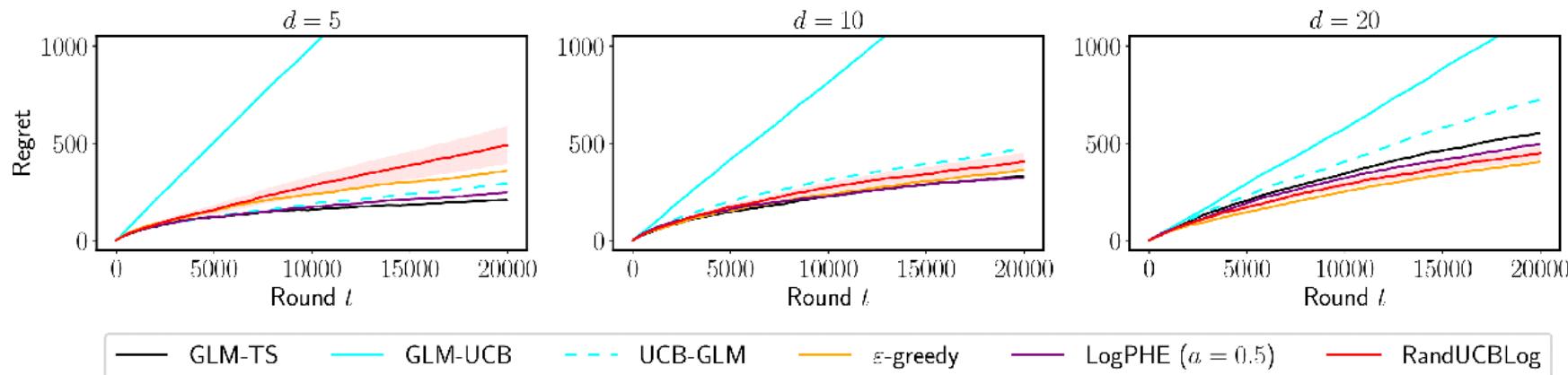
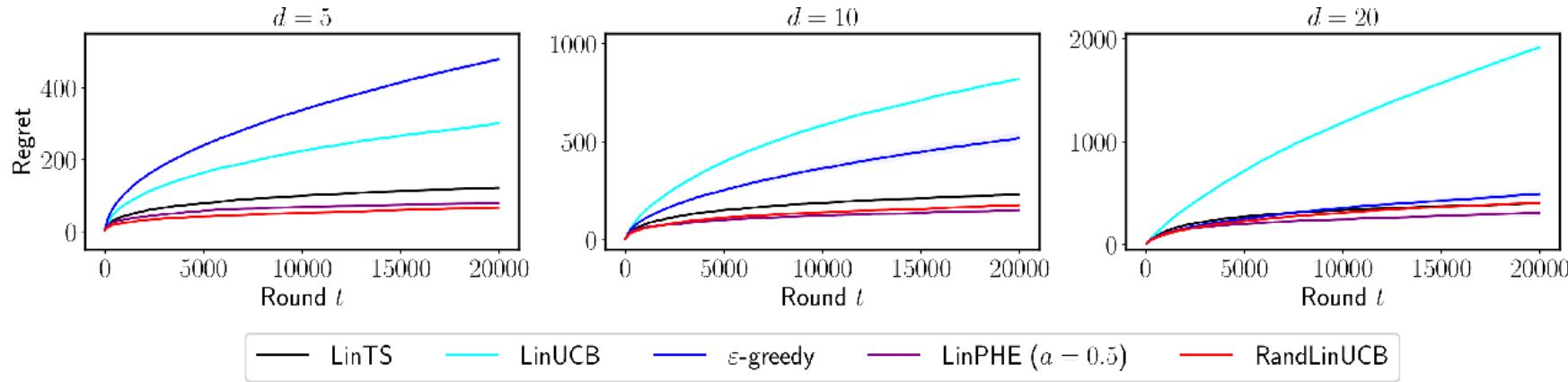
- Draw a random sample

$$\tilde{\theta} \sim \text{Gaussian}(\hat{\theta}, V_t^{-1})$$

and select

$$a_{t+1} = \operatorname{argmax}_a \tilde{\theta}^\top a$$

# Comparisons



# Summary

- What are bandits, and why should you care
  - Many applications
- Finite-armed bandits
  - Explore-then-commit
  - epsilon-greedy
  - UCB:  $a_t = \operatorname{argmax}_a \hat{\alpha}_a + \sqrt{\frac{2 \log t}{T_a(t)}}$
  - Thompson sampling:  $\tilde{\alpha}_a(t) \sim \text{Gaussian} \left( \hat{\alpha}_a(t), \frac{1}{1+T_a(t)} \right)$
  - EXP3
- Linear bandits
  - LinUCB, LinTS

Shuai Li

<https://shuaili8.github.io>

Questions?

# References

- Agrawal, S., & Goyal, N. (2013). Further Optimal Regret Bounds for Thompson Sampling. *international conference on artificial intelligence and statistics*.
- Agrawal, S., & Goyal, N. (2013b). Thompson Sampling for Contextual Bandits with Linear Payoffs. *international conference on machine learning*.
- Auer, P., Cesabianchi, N., & Fischer, P. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2), 235-256.
- Babaioff, M., Dughmi, S., Kleinberg, R., & Slivkins, A. (2015). Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation (TEAC)*, 3(1), 4.
- Bush, R. R. and Mosteller, F. (1953). A stochastic model with applications to learning. *The Annals of Mathematical Statistics*, pages 559–585.
- Chapelle, O., & Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24.
- Hu, Y., Da, Q., Zeng, A., Yu, Y., & Xu, Y. (2018). Reinforcement Learning to Rank in E-Commerce Search Engine: Formalization, Analysis, and Application. *knowledge discovery and data mining*.
- Kocsis, L., & Szepesvari, C. (2006). Bandit based monte-carlo planning. *european conference on machine learning*.
- Larrnaaga, M., Ayesta, U., & Verloop, I. M. (2016). Dynamic control of birth-and-death restless bandits: Application to resource-allocation problems. *IEEE/ACM Transactions on Networking*, 24(6), 3812-3825.

# References II

- Lattimore, T., György, A., & Szepesvári, C. (2014). On learning the optimal waiting time. In International Conference on Algorithmic Learning Theory (pp. 200-214). Springer, Cham.
- Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.
- Le, T., Szepesvari, C., & Zheng, R. (2014). Sequential Learning for Multi-Channel Wireless Network Monitoring With Channel Switching Costs. *IEEE Transactions on Signal Processing*, 62(22), 5919-5929.
- Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. *the web conference*, 661-670.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: a novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1), 6765-6816.
- Radlinski, F., Kleinberg, R., & Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. In Proceedings of the 25th international conference on Machine learning (pp. 784-791). ACM.
- Rafferty, Anna & Ying, Huiji & Jay Williams, Joseph. (2018). Bandit Assignment for Educational Experiments: Benefits to Students Versus Statistical Power. [10.1007/978-3-319-93846-2\\_53](https://doi.org/10.1007/978-3-319-93846-2_53).
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Den Driessche, G. V., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.
- Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

# References III

- Villar, S. S., Bowden, J., & Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2), 199-215.
- Yu, B., Fang, M., & Tao, D. (2016). Linear submodular bandits with a knapsack constraint. In Thirtieth AAAI Conference on Artificial Intelligence.
- Yue, Y., Broder, J. M., Kleinberg, R., & Joachims, T. (2012). The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5), 1538-1556.