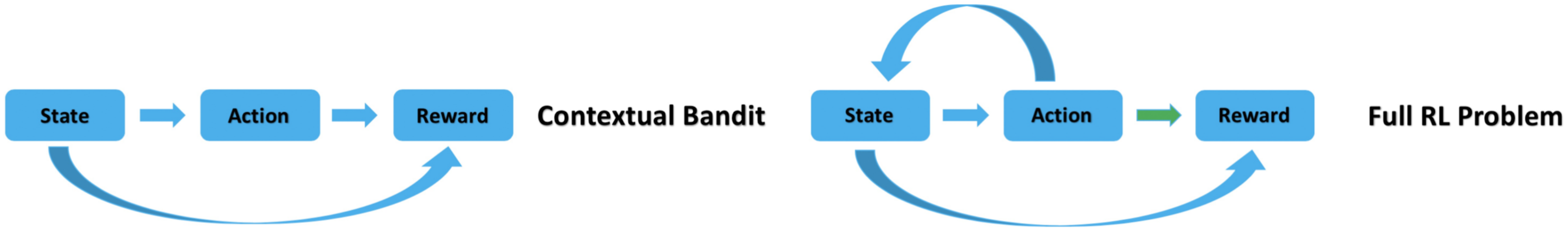
Adaptive Exploration in Linear Contextual Bandits

Botao Hao
Joint work with Tor Lattimore (Deepmind) and Csaba Szepesvari (Deepmind)

Exploration in Contextual Bandit

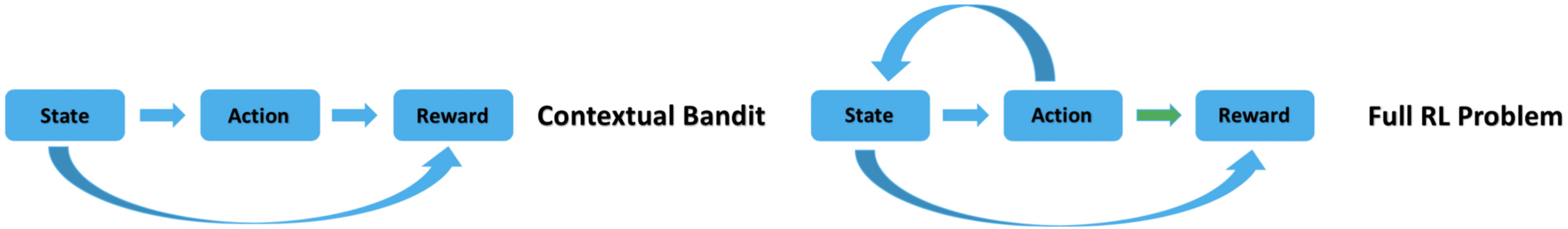
Exploration in Contextual Bandit

- “Simple” reinforcement learning model.



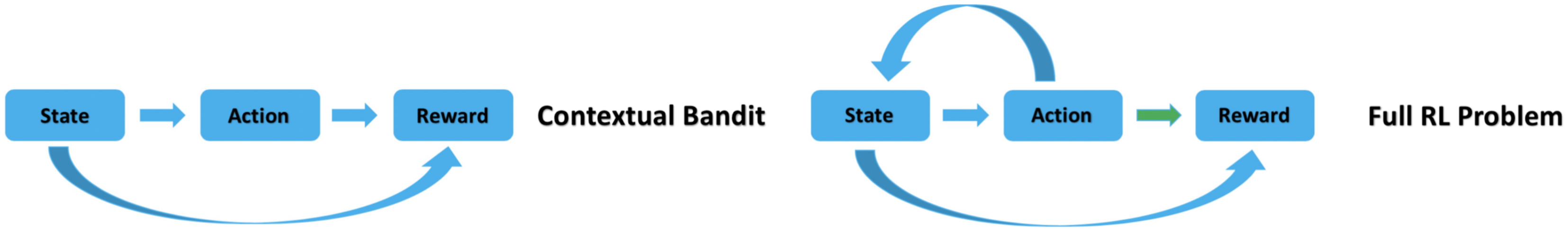
Exploration in Contextual Bandit

- “Simple” reinforcement learning model.
  - Provide **better principles** to design exploration strategies in RL.



Exploration in Contextual Bandit

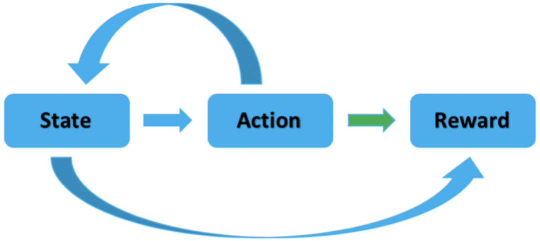
- "Simple" reinforcement learning model.
  - Provide **better principles** to design exploration strategies in RL.
  - Synthetic check for sophisticated methods in RL.



Exploration in Contextual Bandit

- "Simple" reinforcement learning model.
    - Provide **better principles** to design exploration strategies in RL.
    - Synthetic check for sophisticated methods in RL.

Contextual Bandit



Full RL Problem

- Popular model for recommender systems and online advertising.

Motivation

- **Optimism principle** (UCB or Thompson sampling) can be arbitrarily bad!
  - Why? Do not exploit the context structure properly.
  - Do not optimize the trade-off between information and regret.

    Regret: difference between rewards collected by the optimal policy and proposed policy

Motivation

- **Optimism principle** (UCB or Thompson sampling) can be arbitrarily bad!
    - Why? Do not exploit the context structure properly.
    - Do not optimize the trade-off between information and regret.

      Regret: difference between rewards collected by the optimal policy and proposed policy

- Some foundational questions have not been answered yet.
    - How hard is the problem? Dependence of regret on problem structures?
    - Lower bound…

Motivation

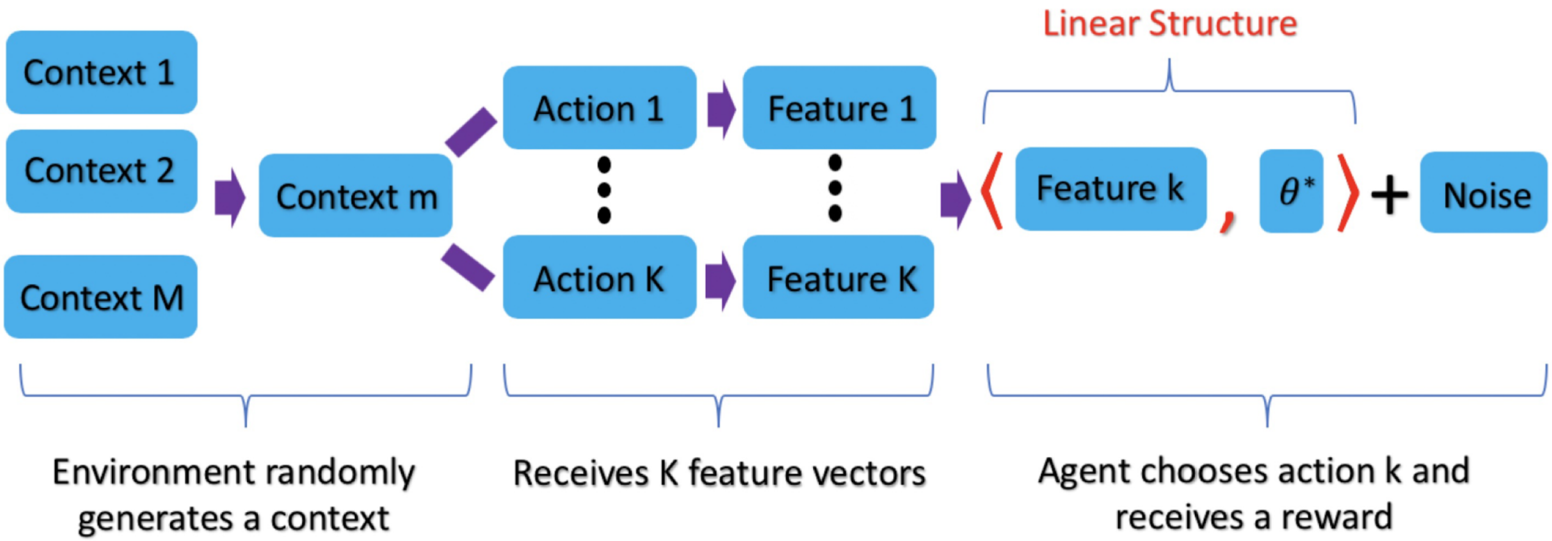
- **Optimism principle** (UCB or Thompson sampling) can be arbitrarily bad!
  - Why? Do not exploit the context structure properly.
  - Do not optimize the trade-off between information and regret.

    Regret: difference between rewards collected by the optimal policy and proposed policy

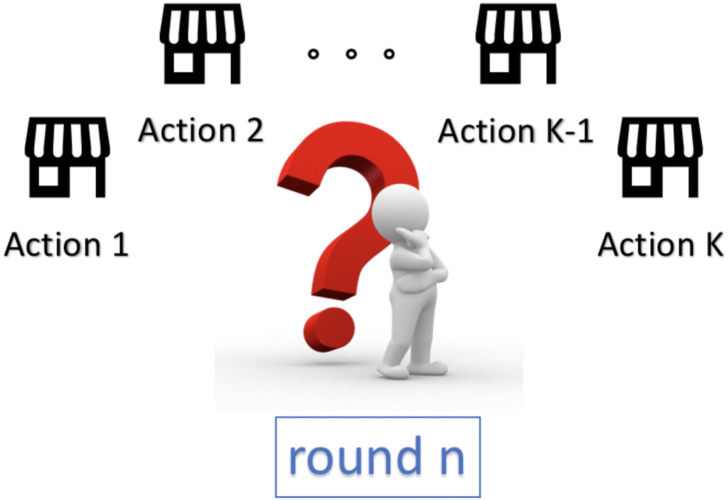
- Some foundational questions have not been answered yet.
  - How hard is the problem? Dependence of regret on problem structures?
  - Lower bound...

Can we design better algorithms for contextual bandits?

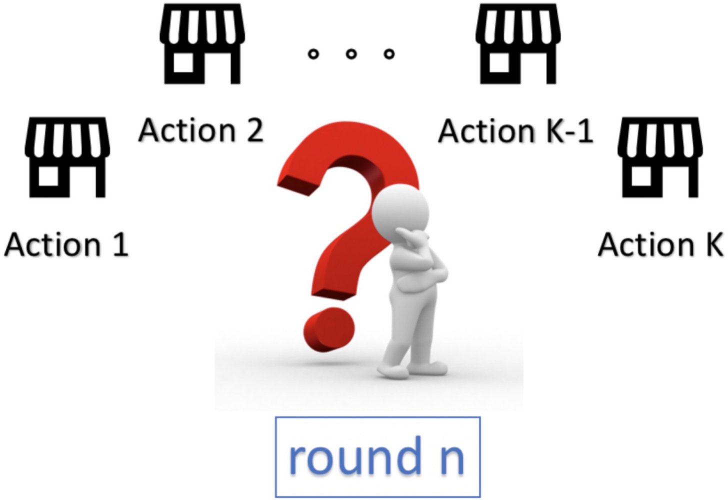
Linear Contextual Bandit



Foundational Limit: Sharp Lower Bound



Foundational Limit: Sharp Lower Bound



Theorem (informal):

$$\liminf_{n\to\infty} \frac{\text{Regret}}{\log n} \geq C$$

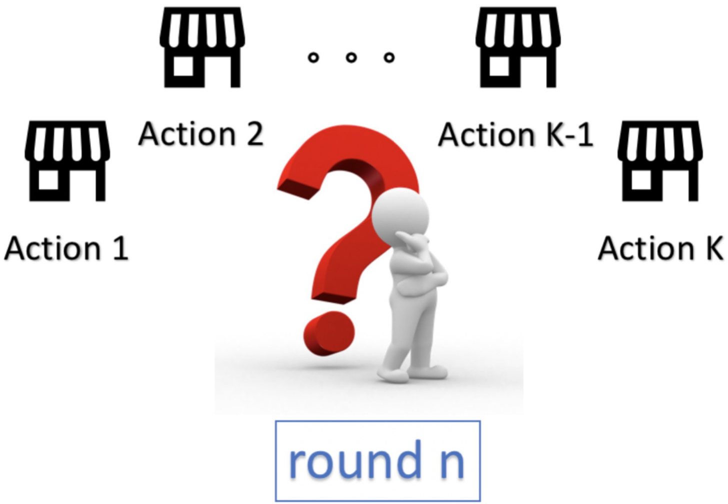
where C is optimal value of the following optimization problem,

$$\min_{\alpha} \sum \alpha_x \Delta_x$$

$$\text{subject to } \sqrt{2}\|x\|_{G_\alpha^{-1}} \leq \Delta_x$$

- Δ_x: sub-optimal gap
- $G_\alpha = \sum \alpha_x x x^\top$

Foundational Limit: Sharp Lower Bound



Theorem (informal):

$$\liminf_{n\to\infty} \frac{\text{Regret}}{\log n} \geq C$$

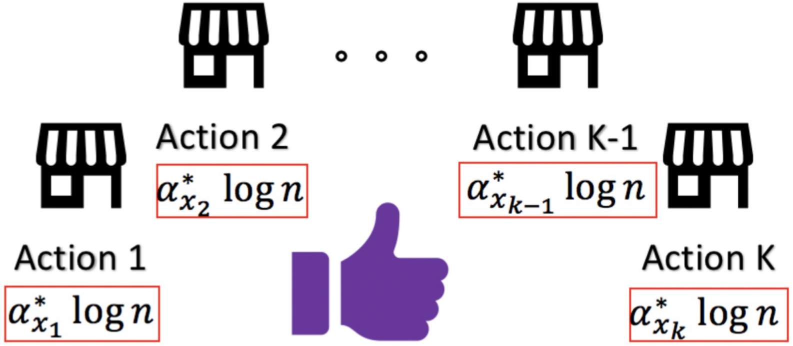
where C is optimal value of the following optimization problem,

$$\min_{\alpha} \underbrace{\boxed{\sum \alpha_x \Delta_x}}_{\text{cumulative regret}}$$

$$\text{subject to } \underbrace{\boxed{\sqrt{2}\|x\|_{G_\alpha^{-1}}}}_{\text{length of confidence interval}} \leq \Delta_x$$

- Δ_x: sub-optimal gap
- $G_\alpha = \sum \alpha_x x x^\top$

Foundational Limit: Sharp Lower Bound



Theorem (informal):

$$\liminf_{n\to\infty} \frac{\text{Regret}}{\log n} \geq C$$

where C is optimal value of the following optimization problem,

$$\min_{\alpha} \boxed{\sum \alpha_x \Delta_x}$$

cumulative regret

$$\text{subject to } \boxed{\sqrt{2}\|x\|_{G_\alpha^{-1}}} \leq \Delta_x$$

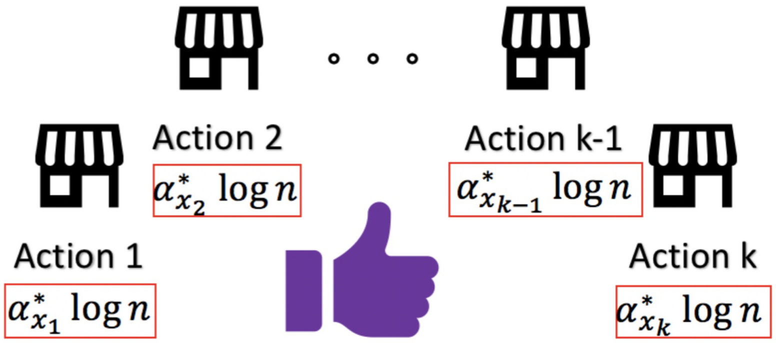
length of confidence interval

- Δ_x: sub-optimal gap
- $G_\alpha = \sum \alpha_x x x^\top$

Remark

- Asymptotical constant C is sharp.
- The allocation rule depends on the problem structure (action set/true parameter).
- When the action set enjoys some good shapes, C could be zero (sub-logarithm regret/bounded regret).
- The lower bound does not depend on the context distribution.

Foundational Limit: Sharp Lower Bound



"How to translate this resource allocation rule to a bandit algorithm? "

Theorem (informal):

$$\liminf_{n\to\infty} \frac{\text{Regret}}{\log n} \geq C$$

where C is optimal value of the following optimization problem,

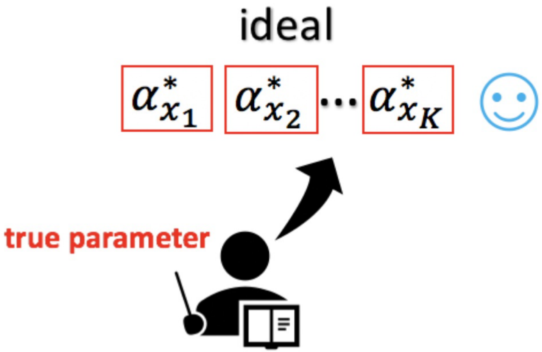
$$\min_{\alpha} \boxed{\sum \alpha_x \Delta_x}$$ cumulative regret

$$\text{subject to } \boxed{\sqrt{2}\|x\|_{G_\alpha^{-1}}} \leq \Delta_x$$

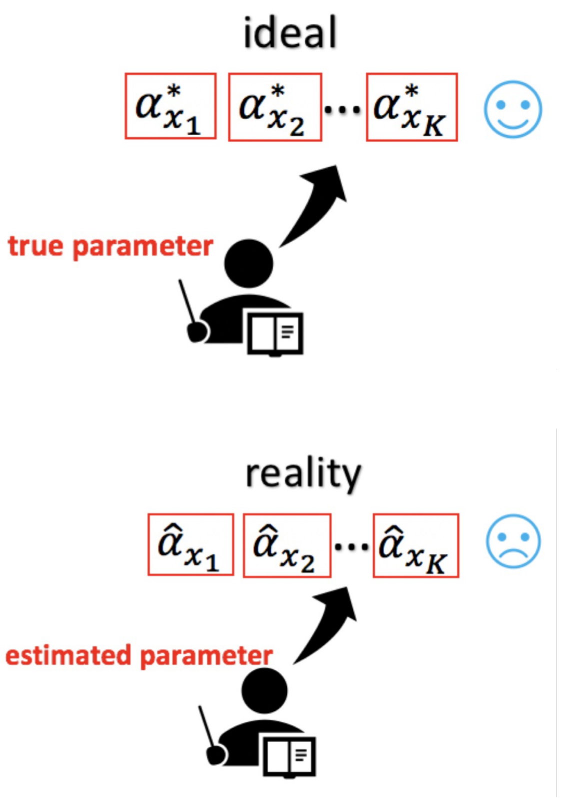
length of confidence interval

- Δ_x: sub-optimal gap
- $G_\alpha = \sum \alpha_x x x^\top$

Algorithm



Algorithm



ideal
$\alpha^*_{x_1}$ $\alpha^*_{x_2}$ ⋯ $\alpha^*_{x_K}$
true parameter
reality
$\hat{\alpha}_{x_1}$ $\hat{\alpha}_{x_2}$ ⋯ $\hat{\alpha}_{x_K}$
estimated parameter

Algorithm

<u>*Convex Optimization Problem*</u>

$$\min_{\alpha} \sum \alpha_x \Delta_x$$

$$\text{subject to } \|x\|_{G_\alpha^{-1}} \le \frac{\Delta_x}{\sqrt{2}}$$

- Δ_x: sub-optimal gap
- $G_\alpha = \sum \alpha_x x x^\top$

- Solve the optimization problem with $\widehat{\Delta}_x$, denote the solution as $\hat{\alpha}_x$
- Check if $N_x(t) \ge \hat{\alpha}_x \log t$ for all sub-optimal arms

  ($N_x(t)$:number of pulls for arm x)

- if yes, do *exploitation/greedy action*
- if not, do *exploration*

  Pull arm : $\arg\min_x \frac{N_x(t)}{\hat{\alpha}_x}$

- Update $\widehat{\Delta}_x$

Algorithm

Convex Optimization Problem

$$\min_{\alpha} \sum \alpha_x \Delta_x$$

$$\text{subject to } \|x\|_{G_\alpha^{-1}} \leq \frac{\Delta_x}{\sqrt{2}}$$

- Δ_x: sub-optimal gap
- $G_\alpha = \sum \alpha_x x x^\top$

Matching Upper Bound!

- ❖ **Solve the optimization problem with $\widehat{\Delta}_x$, denote the solution as $\hat{\alpha}_x$**
- ❖ **Check if $N_x(t) \geq \hat{\alpha}_x \log t$ for all sub-optimal arms**

($N_x(t)$:number of pulls for arm x)

- ❑ **if yes, do *exploitation/greedy action***
- ❑ **if not, do *exploration***

Pull arm : $\arg\min_x \frac{N_x(t)}{\hat{\alpha}_x}$

- ❖ **Update $\widehat{\Delta}_x$**

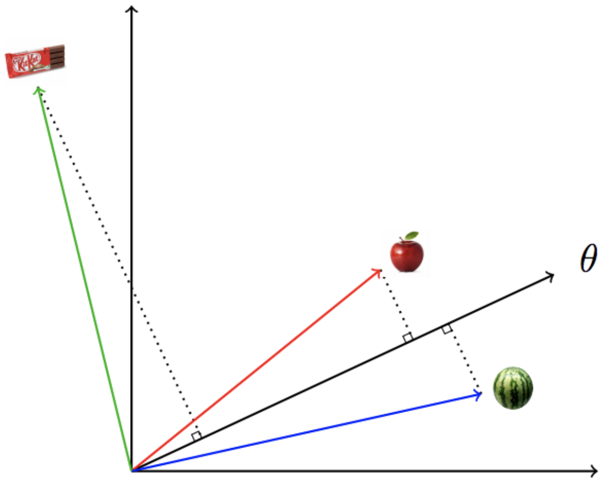
Remark

- If the distribution of contexts is well behaved, our algorithm acts mostly greedily and enjoy sub-logarithmic regret. (adaptive to the good case)

- Asymptotically, the optimal constant is independent of the context distribution. Designing algorithms that optimize for the asymptotic regret may make huge sacrifices in finite-time!

Experiments

$d = 2$ and $k = 3$ and $\mathcal{A} = \{$, , $\}$

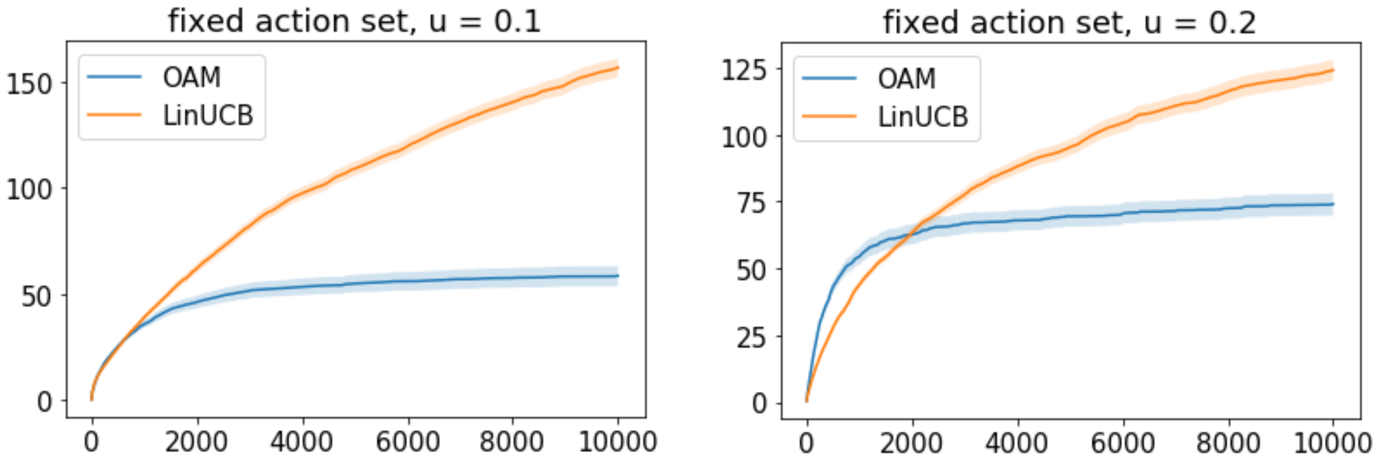


$$\theta^* = (1, 0)$$

$$x_1 = (1, 0), x_2 = (0, 1), x_3 = (1 - u, \gamma u)$$

Experiments



$$\theta^* = (1, 0)$$

$$x_1 = (1, 0), x_2 = (0, 1), x_3 = (1 - u, \gamma u)$$

Limitations and Related Work

Current limitations

- Unclear if the algorithm is minimax optimal
- Need to solve an optimization problem each round

Published Work:

- The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits (Lattimore and Szepesvari, AISTAT 2016)
- Minimal Exploration in Structured Stochastic Bandits (Combes et al., NIPS 2017)
- Exploration in Structured Reinforcement Learning (Ok et al., NIPS 2018)

thank you!