A New Framework: Short-Term and Long-Term Returns in Stochastic Multi-Armed Bandit

Abdalaziz Sawwan (Presenter) and Jie Wu

Department of Computer and Information Sciences

Temple University

Outline

- Introduction to Multi-Armed Bandit (MAB) problems
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

Outline

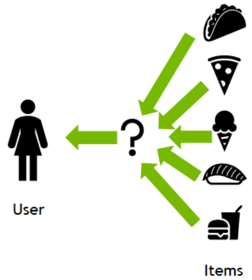
- Introduction
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

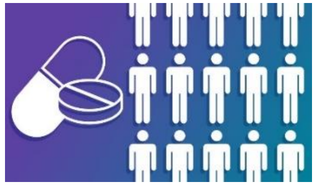
Introduction

- The **Multi-Armed Bandit (MAB) Problem** is a fundamental paradigm in sequential decision-making
- An agent must choose between multiple options (arms) to maximize the total reward
- Balancing:
  - **exploration** (trying new options)
  - **exploitation** (choosing the best-known option)

Introduction

- Attracted significant attention from researchers in various fields
- Rich literature on the theory, algorithms, and applications
- Applications:
  - Online advertising
  - Recommendation systems
  - Clinical trials and more





Outline

- Introduction
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

Challenges in the existing MAB models

- **Delayed feedback**: The true reward of an action may not be immediately observable.
- **Missing information**: Information from delayed feedback may be incomplete.
- **Exploration vs. exploitation**: Balancing the trade-off remains a challenge, especially with delayed feedback.

Outline

- Introduction
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

Previous work

- Dudik et al. [1] were the first to consider delayed feedback
  - Fixed delay
- Pike-Burke et al. [2] considered:
  - getting the sum of rewards that arrive at the same round
  - assumed that the expected delay is known
- Lancewicki et al. [3]:
  - were the first to consider unrestricted delayed feedback
  - time can be reward-dependent
  - infinite-delay is allowed
  - improved regret bounds

[1] Dudik, M., et al. "Efficient optimal learning for contextual bandits." arXiv preprint arXiv:1106.2369 (2011).
[2] Pike-Burke, C., et al. "Bandits with delayed, aggregated anonymous feedback." International Conference on Machine Learning. PMLR, 2018.
[3] Lancewicki, T., et al. "Stochastic multi-armed bandits with unrestricted delay distributions." International Conference on Machine Learning. PMLR, 2021.

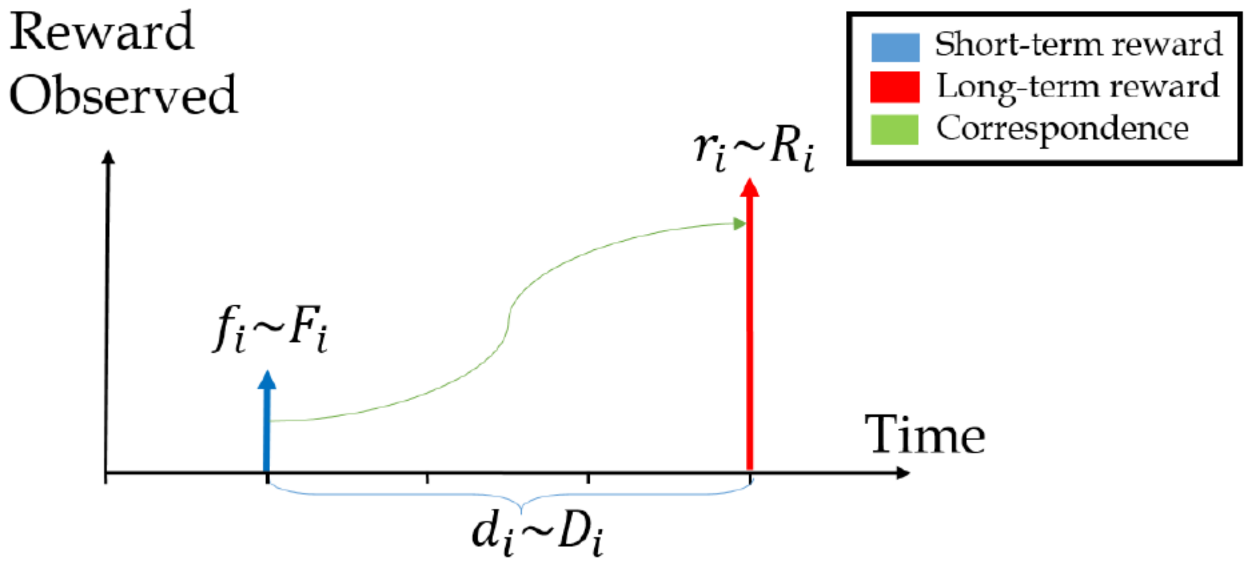
Outline

- Introduction
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

Proposed framework

- Combines **short-term** (instant) and **long-term** (delayed) rewards
- Pulling an arm i yields:
  - short-term reward drawn from distribution F_i
  - long-term reward drawn from distribution R_i
- Dominance of short-term or long-term rewards is controlled by:
  - tunable parameter κ
  - delay distribution D_i
- Known relationship between short-term and long-term reward distributions

Proposed framework



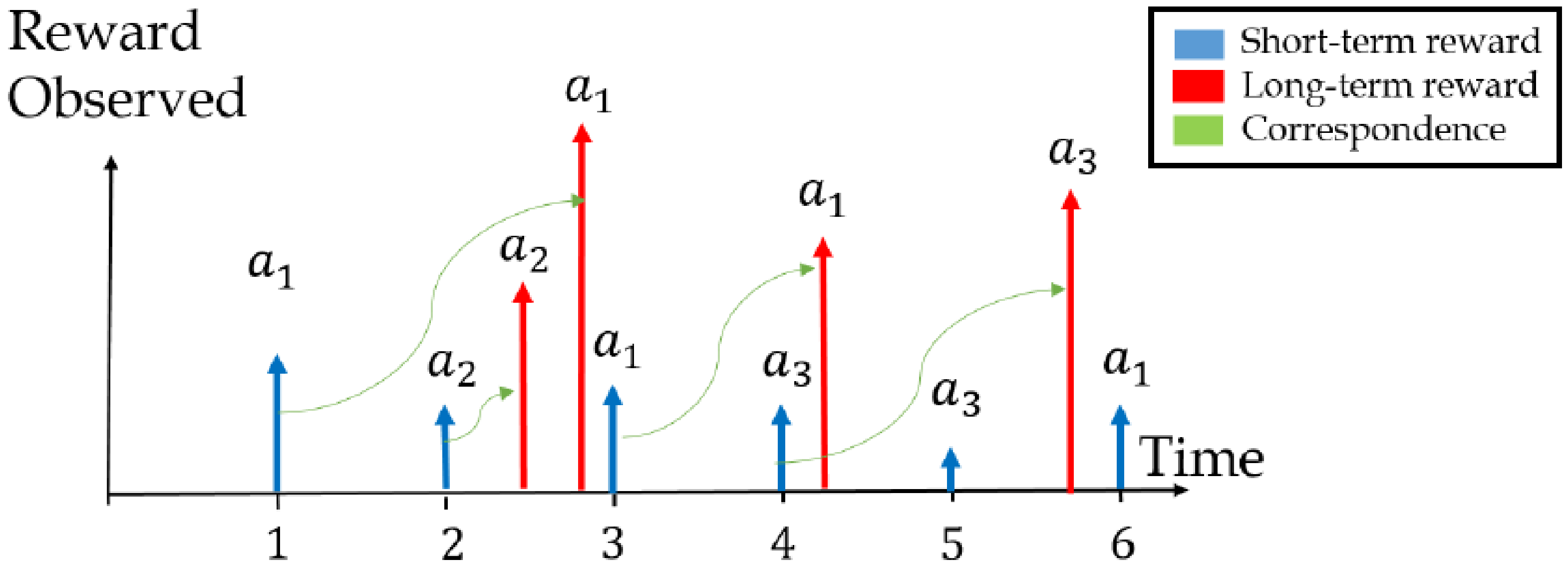
Proposed framework

- F_i and R_i are related by a **linear transformation**
- The linear transformation factor is κ
    - $\kappa \in [0, 1]$
- κ is the long-term to short-term **scaling factor**
- It makes the two rewards observed from an arm reasonably related

Proposed framework

- This makes $r_t(i) \in [0, 1]$, $f_t(i) \in [0, \kappa]$
- For the delay $d_t(i)$: its domain is $\mathbb{N} \cup \{\infty\}$
  - $d_t(i) = \infty \rightarrow r_t(i)$ **will never be observed**
- μ_i: the mean value of R_i
- $\kappa\mu_i$: the mean value of F_i

Proposed framework



Proposed framework

Relationship between Classic and New Framework:

- **Classic MAB model**: Instantaneous feedback
- **Delayed stochastic MAB model**: Rewards observed after a time delay
- **New framework:** unifies both models with tunable parameter κ

Outline

- Introduction
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

Extended UCB-based algorithms

Algorithm 1 UCB for Short-Term and Long-Term Rewards

Input: T, K. //Number of rounds and number of arms.

Output: The set of pulled arms a_t s.t. $t \in [1, T]$.

Initialization: $t \leftarrow 1$. //Start from the first round.

Pull each arm $i \in [1, K]$ one time.

Observe any incoming reward.

Let $t \leftarrow t + K$.

While $t < T$ **do**

for $i \in [1, K]$ **do**

$n_t(i) \leftarrow \Sigma_{\tau:t>\tau+d_\tau} \mathbb{I}\{a_\tau = i\}$.

$\hat{\mu}_t(i) \leftarrow \frac{1}{n_t(i)} \Sigma_{\tau:t>\tau+d_\tau} \mathbb{I}\{a_\tau = i\}(r_\tau + \frac{f_\tau}{\kappa})$.

$UCB_t(i) \leftarrow \hat{\mu}_t(i) + \sqrt{\frac{2\log(T)}{n_t(i)}}$.

Pull arm $a_t = \arg\max_i UCB_t(i)$.

Observe reward.

Let $t \leftarrow t + 1$.

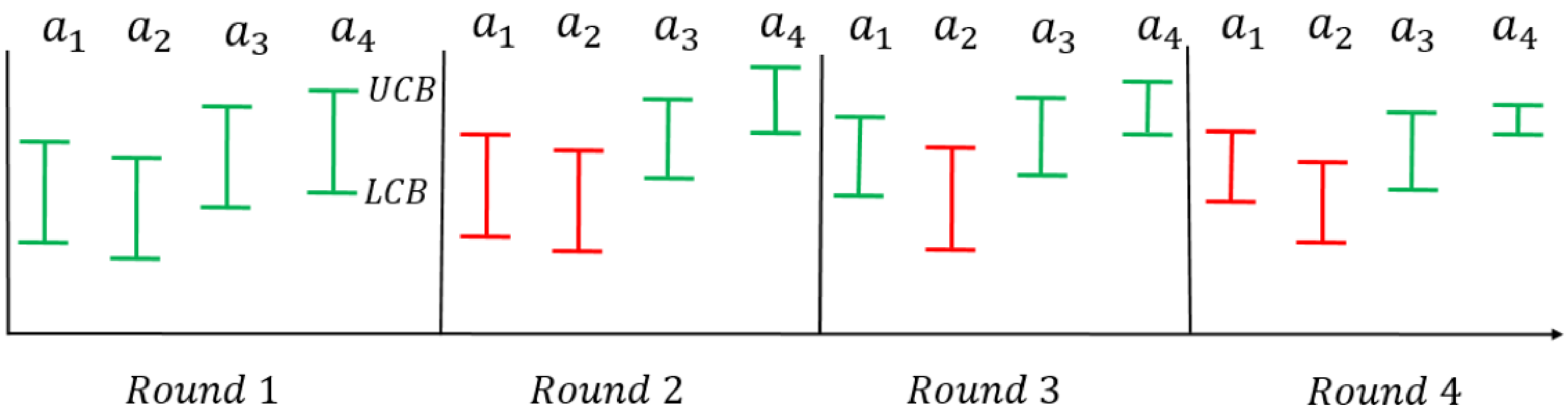
Extended UCB-based algorithms

Algorithm 2 SE for Short-Term and Long-Term Rewards

Input: T, K. //Number of rounds and number of arms.
Output: The set of pulled arms a_t s.t. $t \in [1, T]$.
Initialization: $t \leftarrow 1, S \leftarrow [1, K]$. //Start from the first round.

1: **While** $t < T$ **do**
2: Pull each arm $i \in S$.
3: Observe all incoming feedback.
4: Set $t \leftarrow t + |S|$.
5: **for** $i \in [1, K]$ **do**
6: $n_t(i) \leftarrow \Sigma_{\tau : t > \tau + d_\tau} \mathbb{I}\{a_\tau = i\}$.
7: $\hat{\mu}_t(i) \leftarrow \frac{1}{n_t(i)} \Sigma_{\tau : t > \tau + d_\tau} \mathbb{I}\{a_\tau = i\}(r_\tau + \frac{f_\tau}{\kappa})$.
8: $UCB_t(i) \leftarrow \hat{\mu}_t(i) + \sqrt{\frac{2 \log(T)}{n_t(i)}}$.
9: $ULB_t(i) \leftarrow \hat{\mu}_t(i) - \sqrt{\frac{2 \log(T)}{n_t(i)}}$.
10: Update S by including all arms except all arms i such that there exists j with $UCB_t(i) < LCB_t(j)$.

Extended UCB-based algorithms



Extended UCB-based algorithms

Algorithm 3 PSE for Short-Term and Long-Term Rewards

Input: T, K. //Number of rounds and number of arms.
Output: The set of pulled arms a_t s.t. $t \in [1, T]$.
Initialization: $t \leftarrow 1, S \leftarrow [1, K], \ell \leftarrow 0$.

While $t < T$ **do**
  Let $S_\ell \leftarrow S$, $\ell \leftarrow \ell + 1$. //Phase counting.
  While $S_\ell \neq \emptyset$ **do**
    Pull each arm $i \in S_\ell$, observe incoming feedback.
    Set $t \leftarrow t + |S_\ell|$.
    for $i \in [1, K]$ **do**
      $n_t(i) \leftarrow \Sigma_{\tau : t > \tau + d_\tau} \mathbb{I}\{a_\tau = i\}$.
      $\hat{\mu}_t(i) \leftarrow \frac{1}{n_t(i)} \Sigma_{\tau : t > \tau + d_\tau} \mathbb{I}\{a_\tau = i\}(r_\tau + \frac{f_\tau}{\kappa})$.
      $UCB_t(i) \leftarrow \hat{\mu}_t(i) + \sqrt{\frac{2\log(T)}{n_t(i)}}$.
      $ULB_t(i) \leftarrow \hat{\mu}_t(i) - \sqrt{\frac{2\log(T)}{n_t(i)}}$.
    Eliminate all arms that were observed at least $\frac{\log(T)}{2^{-2\ell-4}}$ times from S_ℓ.
  Update S by including all arms except all arms i such that there exists j with $UCB_t(i) < LCB_t(j)$.

Outline

- Introduction
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

Regret analysis

- Regret is defined as follows:

$$
\begin{aligned}
\mathcal{R}_T &= \max_i \mathbb{E}[\Sigma_{t=1}^T (r_t(i) + f_t(i))] - \mathbb{E}[\Sigma_{t=1}^T r_t(a_t) + f_t(a_t)] \\
&= (1+\kappa) \times (T\mu_{i^*} - \mathbb{E}[\Sigma_{t=1}^T \mu_{a_t}]) = (1+\kappa) \times \mathbb{E}[\Sigma_{t=1}^T \Delta_{a_t}],
\end{aligned}
$$

Regret analysis

Theorem *The regret of the strategy in Algorithm 2 is bounded under our model. The bound is given by*

$$\mathcal{R}_T \leq \min_{\vec{q}\in(0,1]^K} \Sigma_{i\neq i^*} 40(\log T/\Delta_i)(1/q_i + 1/q_{i^*})$$

$$+\log(K)\max_{i\neq i^*}\{(d_i(q_i) + d_{i^*}(q_{i^*}))\Delta_i\}\} + \kappa\sqrt{KT\log T}.$$

Furthermore, we can get another incomparable different bound for the regret, which is given by

$$\mathcal{R}_T \leq \min_{q\in(0,1]} \Sigma_{i\neq i^*} 325\frac{\log T}{q\Delta_i} + 4\max_i d_i(q) + \kappa\sqrt{KT\log T}.$$

Regret analysis

Theorem *The regret of the strategy in Algorithm 3 is bounded under our model. The bound is given by*

$$\mathcal{R}_T \leq \min_{\vec{q}\in(0,1]^K} \Sigma_{i\neq i^*} 290 \log(T)/q_i\Delta_i$$
$$+ \log(T)\log(K) \max_{i\neq i^*} d_i(q_i)\Delta_i + \kappa\sqrt{KT\log T}.$$

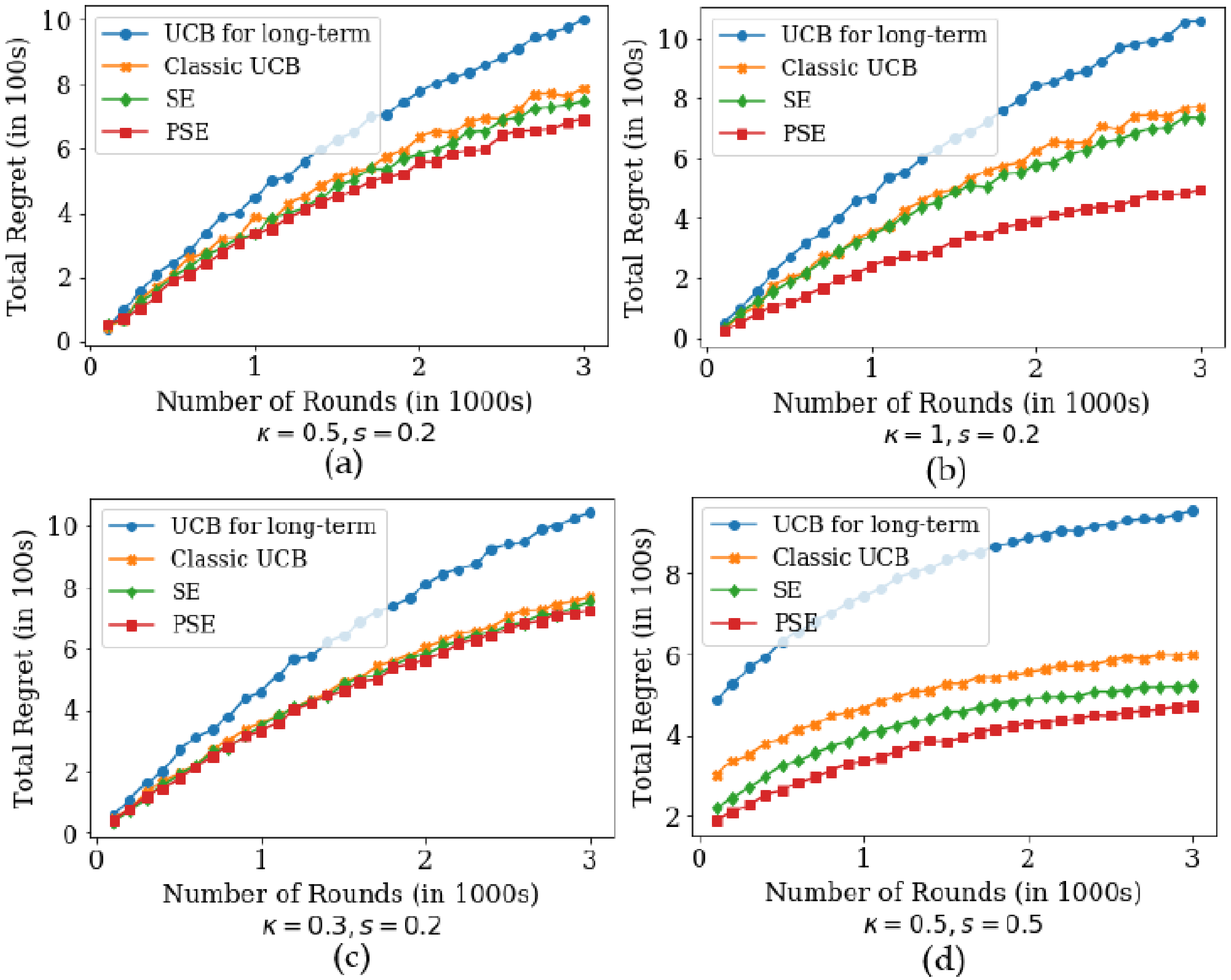
Outline

- Introduction
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

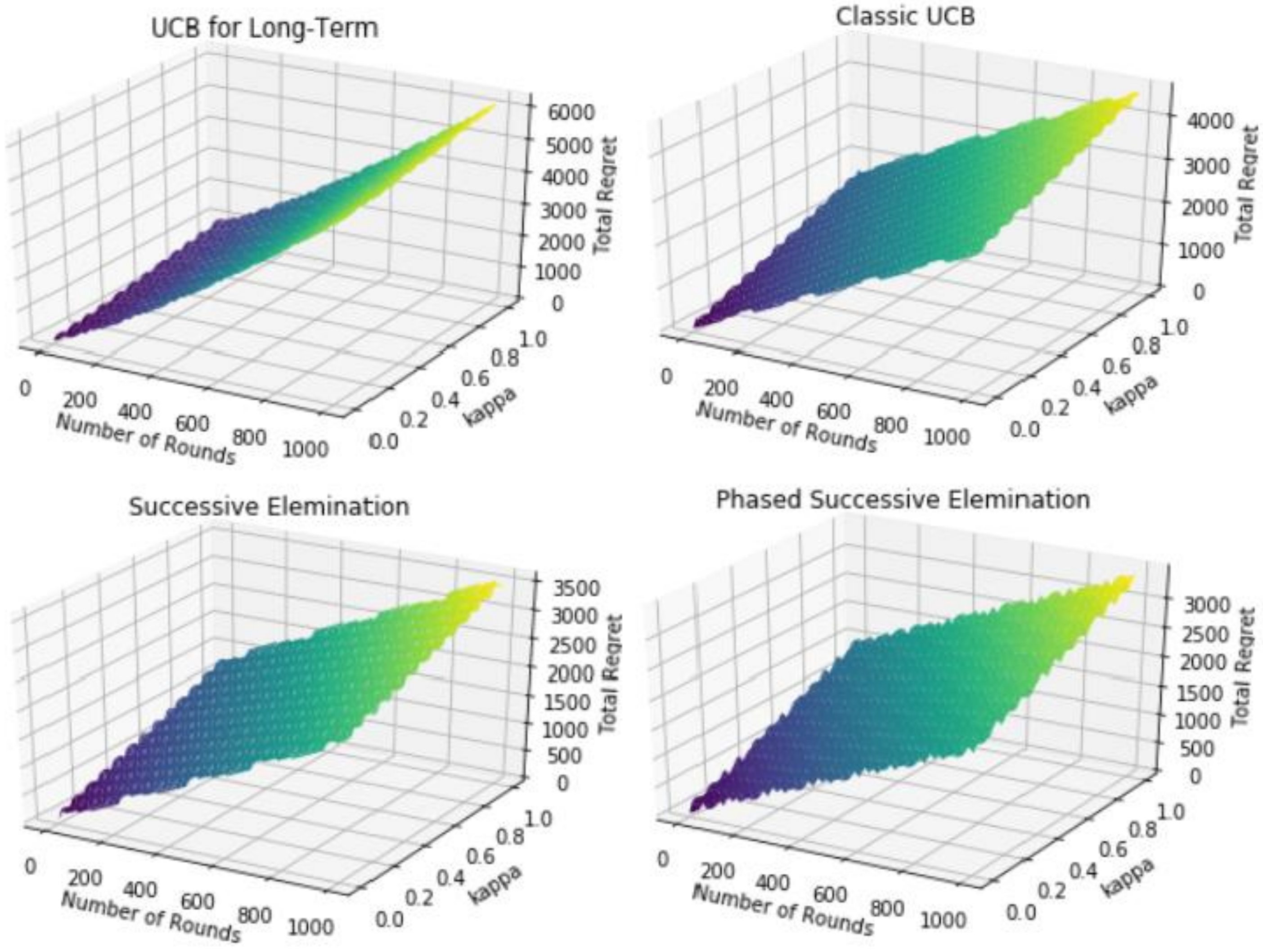
Simulations

- Synthetic data:
  - Generated to test the algorithms under controlled conditions
- Real-world data:
  - Collected from a real application to demonstrate practical performance
  - Application of sparse learning of incomplete traffic speed data
- Performance metric: **Total regret**

Simulations



Simulations



Outline

- Introduction
- Challenges in the existing MAB models
- Previous work
- Proposed framework
- Extended UCB-based algorithms
- Regret analysis
- Simulations
- Future work

Future Work

- **Explore** potential framework extensions, such as incorporating partial feedback
- **Investigate** other algorithms to be adapted to the new framework
- **Relax** the condition of having a linear transformation between the two reward distributions
- **Make** κ an unknown random variable
- **Include** multiple long-term rewards for pulling an arm
- **Apply** the framework to additional real-world problems

Conclusion

- General framework for MAB with short-term and long-term rewards
- Near-optimal Extended UCB-based algorithms
- Regret analysis of the proposed algorithms
- Evaluation on synthetic and real-world data to demonstrate the effectiveness of the proposed algorithms

Q&A

Abdalaziz Sawwan (Presenter) and Jie Wu
Department of Computer and Information Sciences
Temple University