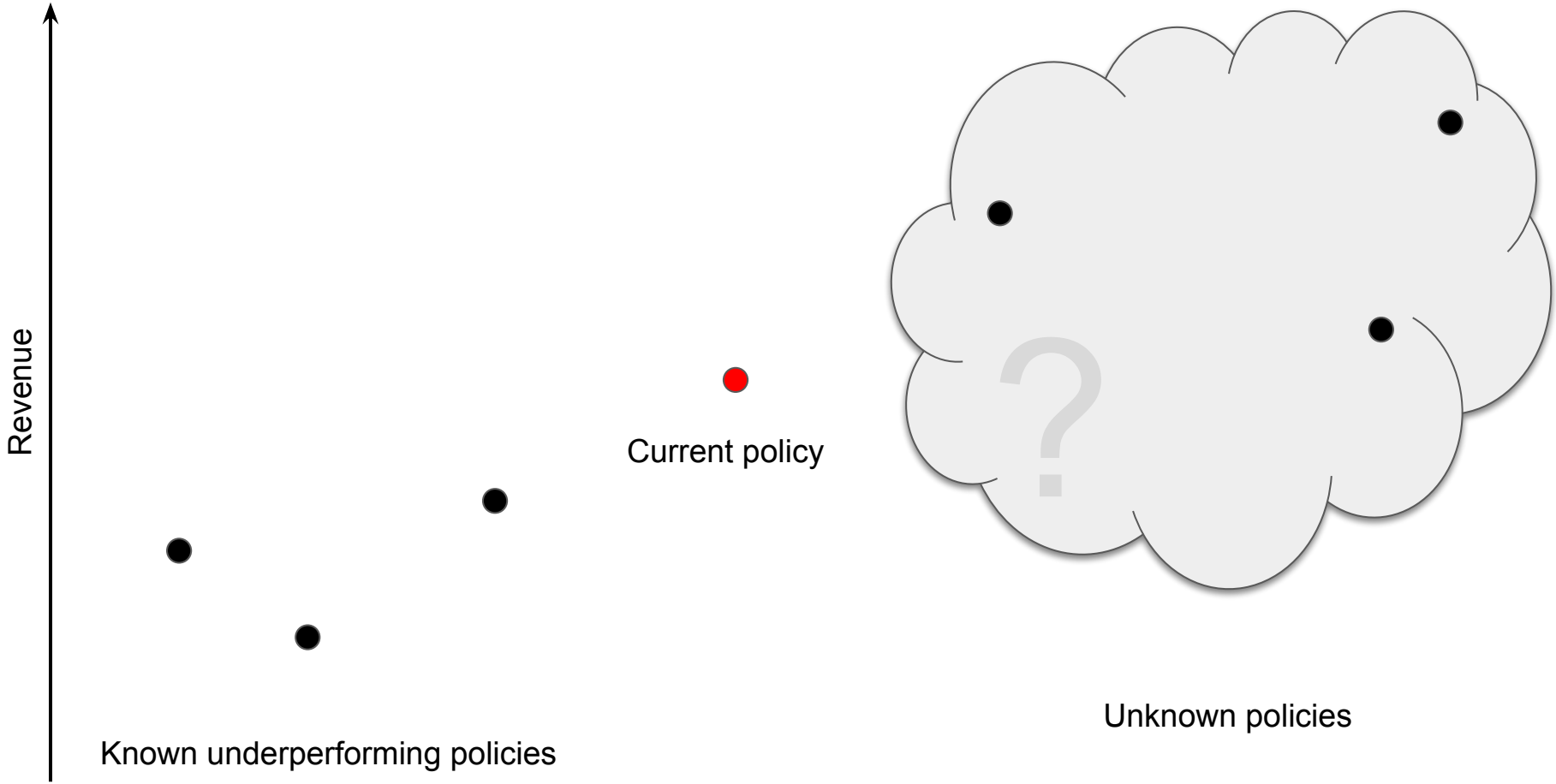
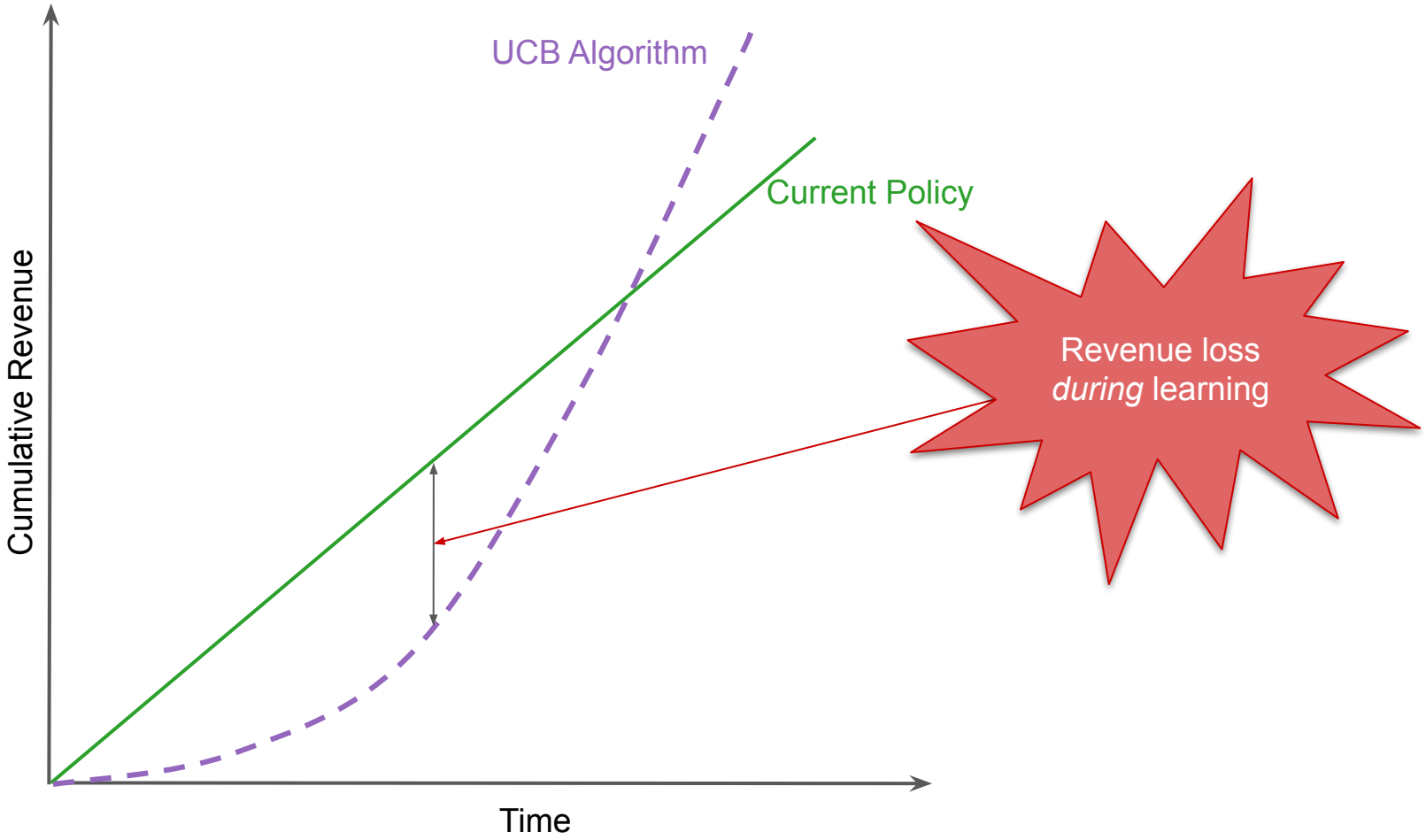
Improved Algorithms for Conservative Exploration in Bandits

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric and Matteo Pirotta

Facebook AI Research

facebook Artificial Intelligence

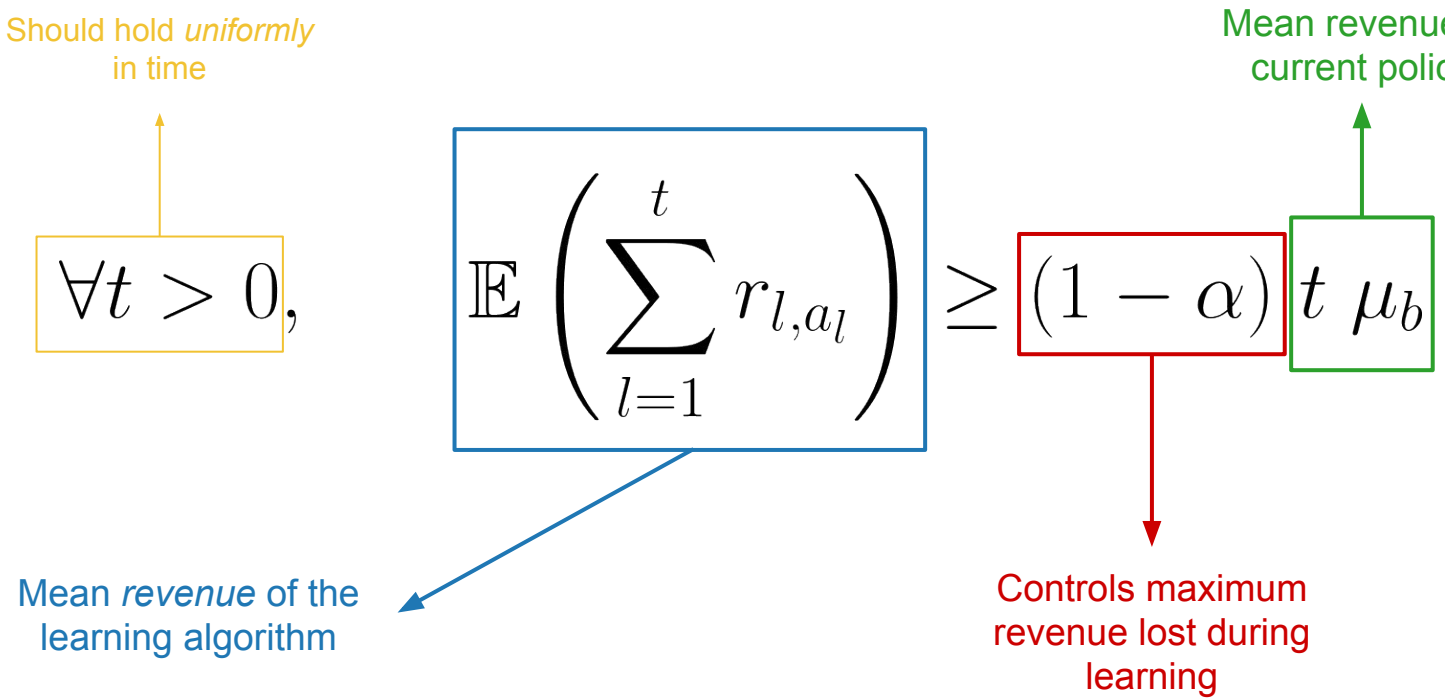

Revenue
Current policy
?
Known underperforming policies
Unknown policies


UCB Algorithm
Current Policy
Revenue loss
during learning
Cumulative Revenue
Time

Problem: How to learn an optimal policy without sacrificing much revenue?

*(aka: how to perform exploration in a **conservative** way?)*

Conservative Condition



Previous Work:

- Theoretically optimal algorithms for conservative exploration (CUCB) (Wu et al. 2016, Kazerouni et al. 2017)

Contributions:

- ➔ *Improved* empirical performance in multi-armed and linear bandit (CUCB2)
- ➔ Novel *relaxed* conservative condition

CUCB *(previous algorithm)*

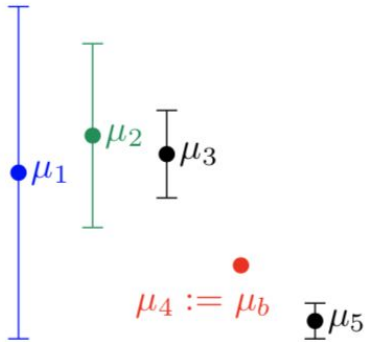
- Two phase algorithm
  a. Computes optimistic arm
  b. Checks a lower bound on the total revenue

=> impacts empirical performance!

CUCB2 *(our algorithm)*

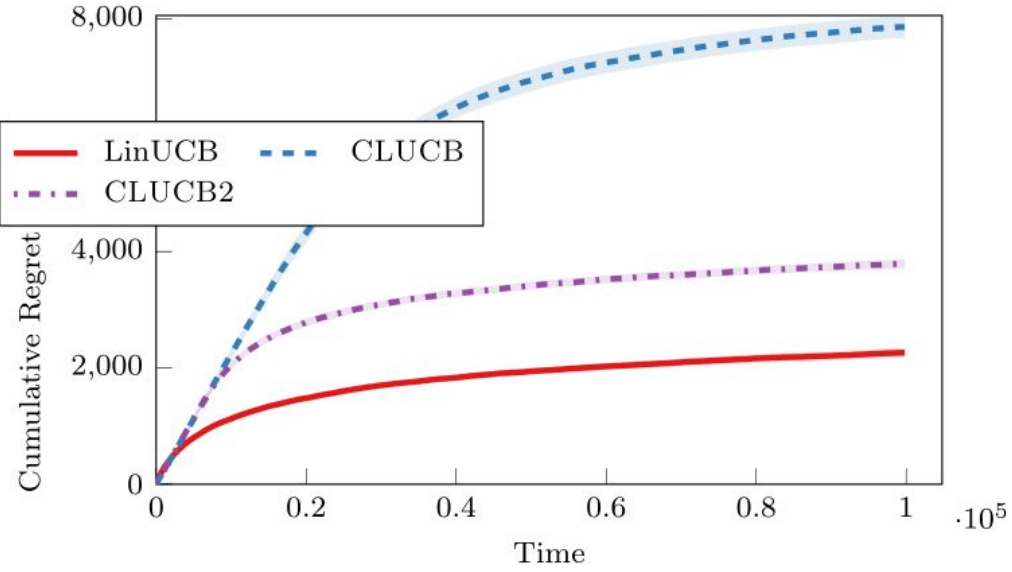
- Computes set of safe arms
- Plays the optimistic arm among safe arms

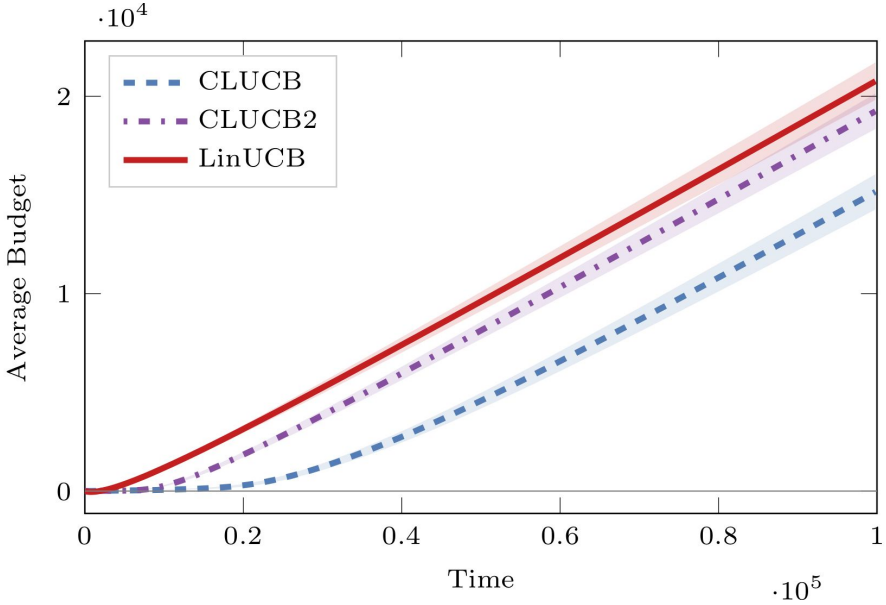
=> same regret but **better** performance!



Example: CUCB approach is suboptimal

Jester Jokes Dataset (Goldberg et al. 2001)





- Cold start problem
- Linear features