Neural Bandits With Reward-Biased Maximum Likelihood Estimation (RBMLE)

Ping-Chun Hsieh

July 13, 2020

TODOs

1. Run experiments for LinUCB, TS, LinRBMLE (bandit environment will be provided to you)

2. Understand Neural-UCB and Neural-TS and implement these two algorithms

3. Design and prototype our algorithm: Neural-RBMLE

4. Regret analysis of Neural-RBMLE

5. Empirical evaluation of our method and baselines using real datasets

Contextual Bandit Model

- Idea: Parametrize rewards with contexts

Arm 1

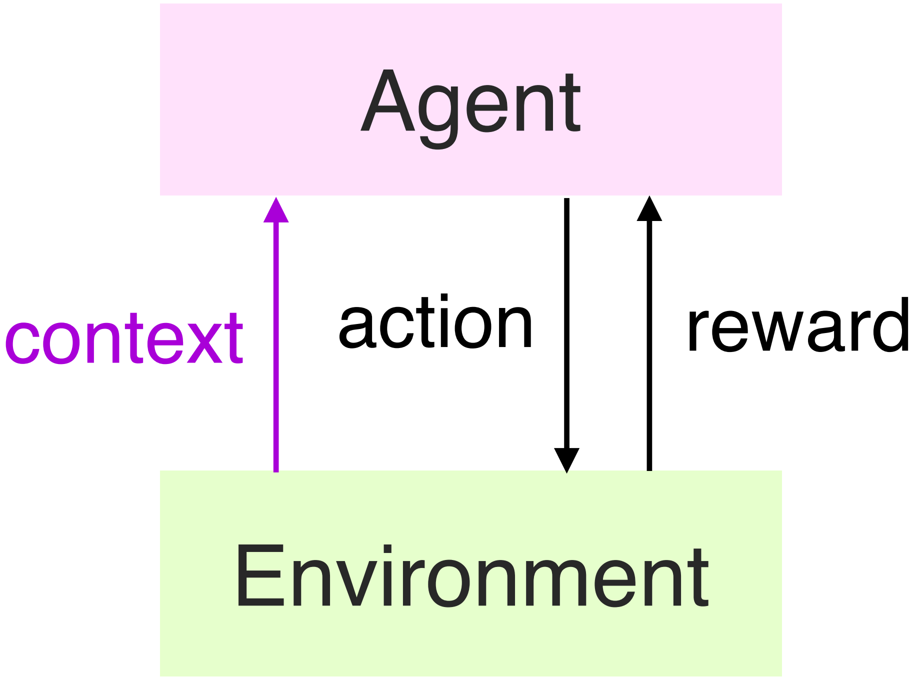
$f_1(\theta; x) + \varepsilon_1$

Arm 2

$f_2(\theta; x) + \varepsilon_2$

Arm 3

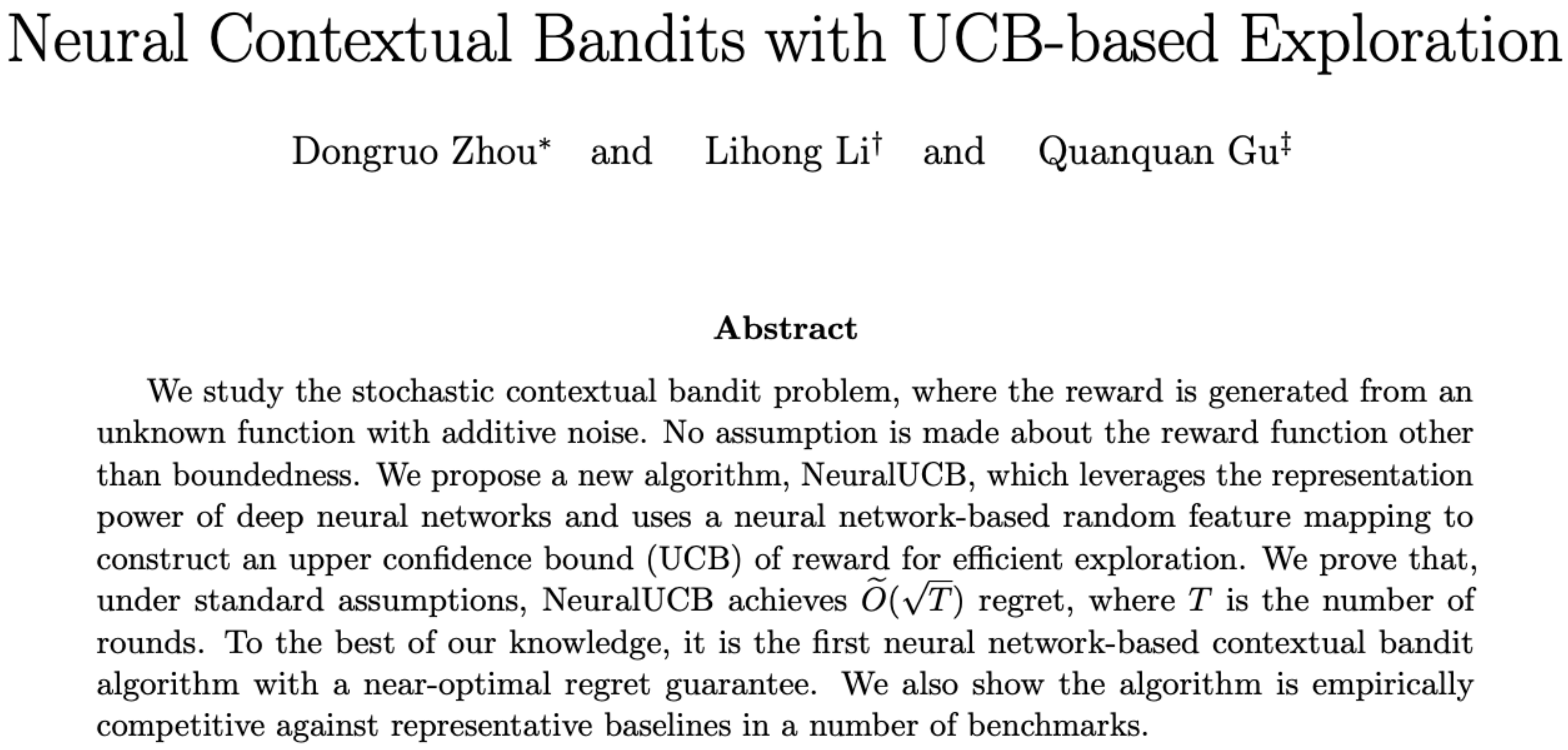
$f_3(\theta; x) + \varepsilon_3$



  - x: context (side information)
  - θ: unknown parameters (to be learned)
  - f_1, f_2, f_3 : some parametric learning model
  - Goal: Minimize cumulative regret
- Question: How to define regret for contextual bandits?

- More recently: Active on-going research on combining contextual bandits and neural networks

Zhou et al., *Neural Contextual Bandits with UCB-based Exploration, (ICML 2020)*



Neural Contextual Bandits with UCB-based Exploration

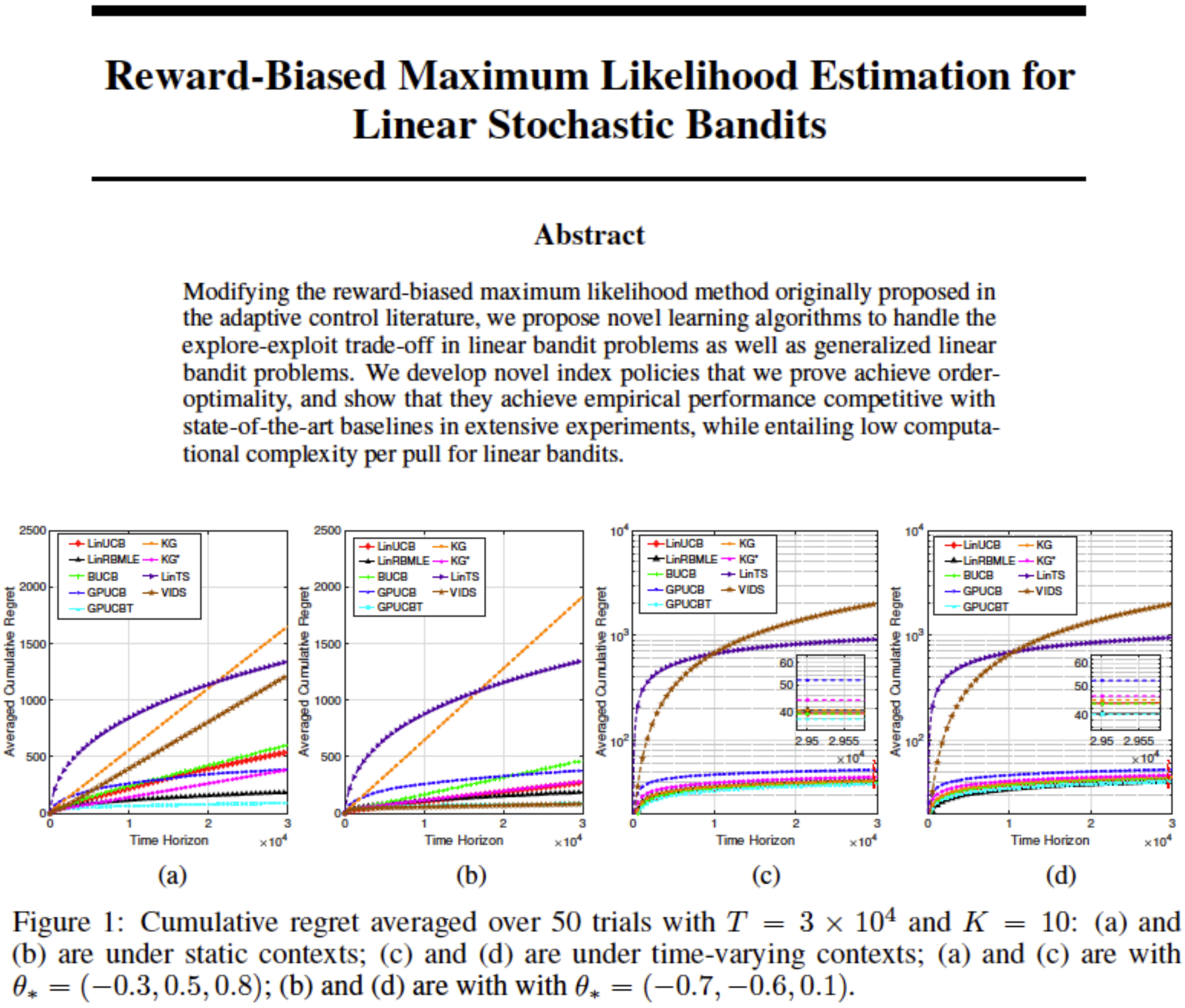
Dongruo Zhou* and Lihong Li† and Quanquan Gu‡

Abstract

We study the stochastic contextual bandit problem, where the reward is generated from an unknown function with additive noise. No assumption is made about the reward function other than boundedness. We propose a new algorithm, NeuralUCB, which leverages the representation power of deep neural networks and uses a neural network-based random feature mapping to construct an upper confidence bound (UCB) of reward for efficient exploration. We prove that, under standard assumptions, NeuralUCB achieves $\widetilde{O}(\sqrt{T})$ regret, where T is the number of rounds. To the best of our knowledge, it is the first neural network-based contextual bandit algorithm with a near-optimal regret guarantee. We also show the algorithm is empirically competitive against representative baselines in a number of benchmarks.

- Main Idea: Use a NN to approximate the reward function
- Challenge: Quantify *regret* under the existence of a NN

- Our latest work: “Reward-Biased Maximum Likelihood Estimation for Linear Stochastic Bandits”



Reward-Biased Maximum Likelihood Estimation for Linear Stochastic Bandits

Abstract

Modifying the reward-biased maximum likelihood method originally proposed in the adaptive control literature, we propose novel learning algorithms to handle the explore-exploit trade-off in linear bandit problems as well as generalized linear bandit problems. We develop novel index policies that we prove achieve order-optimality, and show that they achieve empirical performance competitive with state-of-the-art baselines in extensive experiments, while entailing low computational complexity per pull for linear bandits.

(a)

(b)

(c)

(d)

Figure 1: Cumulative regret averaged over 50 trials with $T = 3 \times 10^4$ and $K = 10$: (a) and (b) are under static contexts; (c) and (d) are under time-varying contexts; (a) and (c) are with $\theta_* = (-0.3, 0.5, 0.8)$; (b) and (d) are with with $\theta_* = (-0.7, -0.6, 0.1)$.