NEURAL INFORMATION PROCESSING SYSTEMS

TEXAS A&M UNIVERSITY

DeepTOP: Deep Threshold-Optimal Policy for MDPs and RMABs

Khaled Nakhleh, I-Hong Hou

Paper link: https://arxiv.org/abs/2209.08646

Threshold policies for MDPs

- **Problem examples**
  - Fan controller turning on when the room temperature exceeds a certain temperature.
  - Central bank raising the interest rate if inflation exceeds a particular value.

- Learning the optimal threshold function is more sample-efficient than generic reinforcement learning (RL) algorithms.

- The actions of threshold policies for MDPs are monotone. We exploit this feature to find a simple gradient expression.

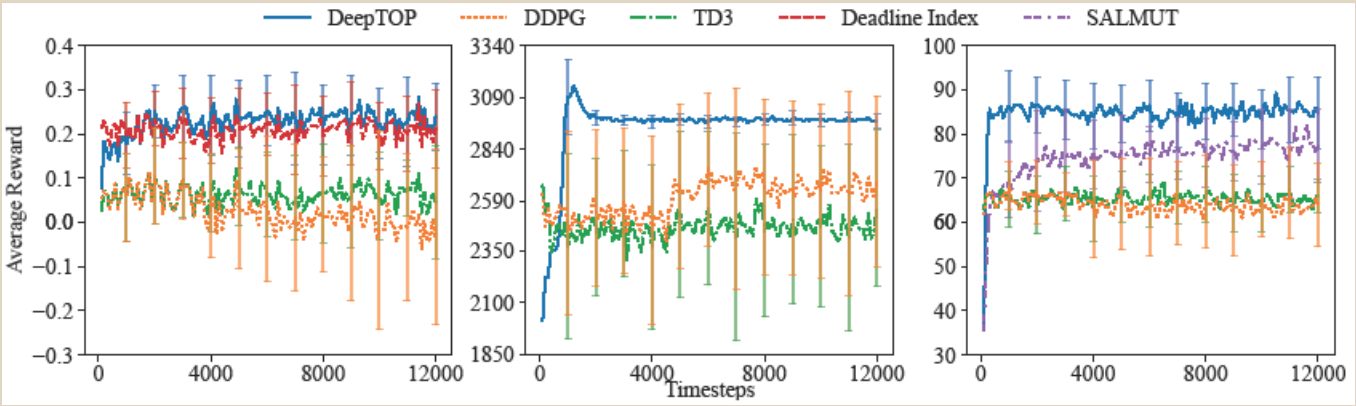
- **Goal:** design an off-policy, model-free algorithm for learning the threshold function called DeepTOP-MDP.

Restless Multi-Armed Bandits (RMABs) threshold policy

- The threshold policy theorem and DeepTOP is extended to the RMAB framework where each arm environment is an MDP.
- We formulate an alternative problem for the threshold policy and show that the objective function has a simple expression.
- We show that the Whittle index is the optimal threshold policy which maximizes the objective function of the alternative problem.
- **Goal:** obtain the optimal control policy that activates the largest V-valued arms out of N arms.

MDP problems' results

- **EV charging:** station decides whether it is optimal to charge the EV or not based on its state at time t.
- **Inventory management:** manager decides if its optimal to buy additional goods based on the season's fluctuations and in-lead times in orders.
- **Make-to-stock:** system that produces m items with W demand classes and buffer size s. system determines if it will accept class orders or not.



(a) EV charging. (b) Inventory management. (c) Make-to-stock production.

RMAB problems' description

- **One-dimensional bandits:**

Each arm has 100 states with the reward depending on the current state of arm i. If arm is activated, next state is $\min\{s_{i,t} + 1, 99\}$ with probability p_i. Otherwise, next state is $\max\{s_{i,t} - 1, 0\}$ with probability q_i.

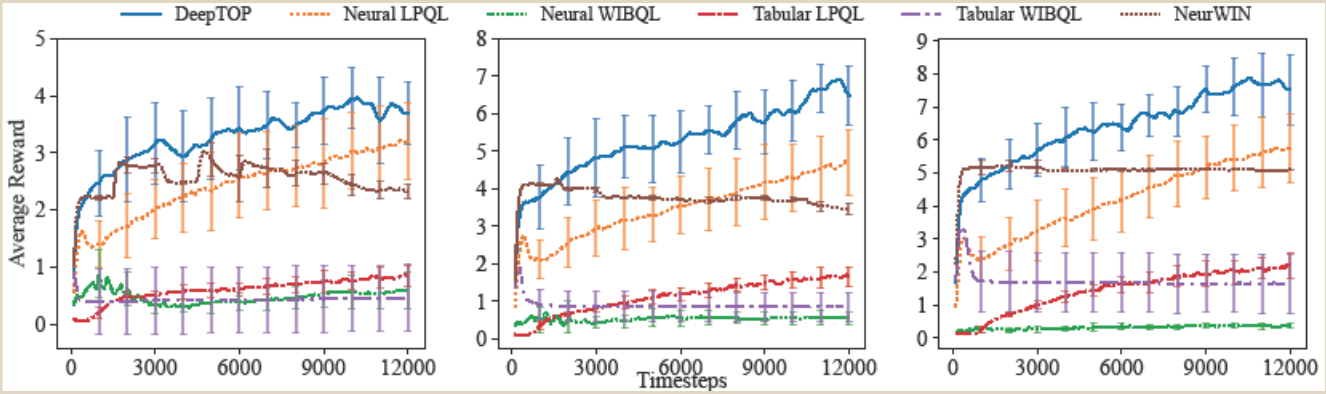
- **Recovering bandits:**

RMAB that studies the varying behavior of customers on advertisement links. If an arm is activated, the state $s_{i,t}$ is reset to state 1. Otherwise, the state increases by 1.

- DeepTOP-RMAB is tested on the two problems for arm-budget pair (N,V): (10,3), (20,5), and (30,6).

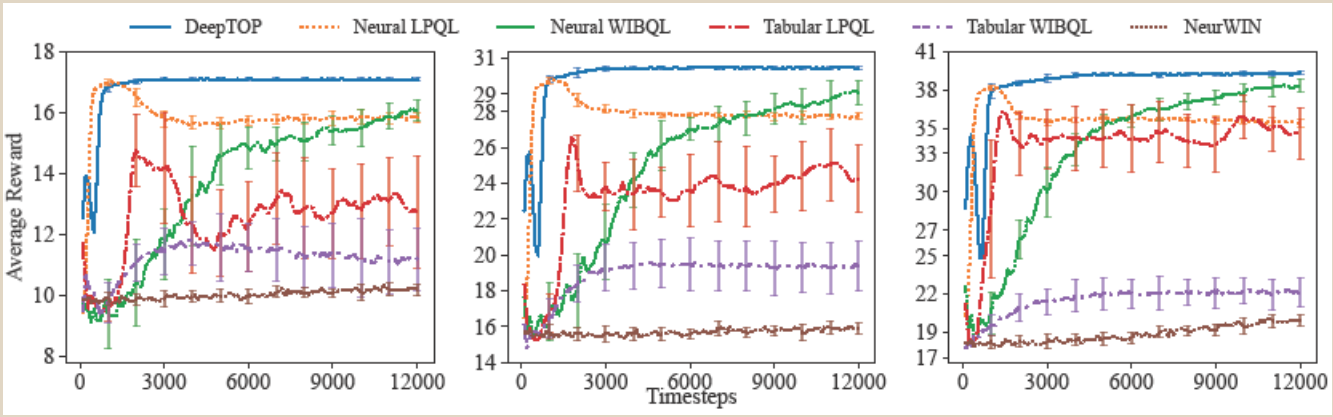
RMAB problems' results - continued

One-dimensional bandits



(a) N = 10. V = 3. (b) N = 20. V = 5. (c) N = 30. V = 6.

Recovering bandits



(a) N = 10. V = 3. (b) N = 20. V = 5. (c) N = 30. V = 6.

For more information

- **ArXiv link:**

  https://arxiv.org/abs/2209.08646

- **Source code link:**

  https://github.com/khalednakhleh/deeptop

- **Author contact info**
  - **LinkedIn:** https://www.linkedin.com/in/khalednakhleh/
  - **Email:** khaled.jamal@tamu.edu