NEURAL INFORMATION PROCESSING SYSTEMS

TEXAS A&M UNIVERSITY

1896

NeurWIN: Neural Whittle Index Network For Restless Bandits Via Deep RL

NeurIPS 2021

Paper: https://arxiv.org/abs/2110.02128

Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I-Hong Hou, Srinivas Shakkottai

Motivating Advertisement Example

Recovering bandits [Pike-Burke et al. 2019]

Classic *Restless bandits'* problem

• • •

Bandit 1 $s_1[t]$ $a_1[t] = 1$

Bandit 2 $s_2[t]$ $a_2[t] = 0$

• • •

Bandit N $s_N[t]$ $a_N[t] = 1$

States change regardless if the bandit is activated or left passive

We obtain the rewards based on active or passive actions

$R_{1,act}(s_1[t])$ $R_{2,pass}(s_2[t])$ $R_{N,act}(s_N[t])$

Setting

- Sequential decision-making problem for timesteps $t = 0, 1, \ldots, \infty$
- N choices each modelled as a restless bandit referenced by $i = 1, 2, \ldots, N$
- Control policy π can choose M out of N restless bandits in each timestep
- **Objective** is to maximize total discounted rewards

$$\mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{\boldsymbol{t=0}}^{\infty} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\beta^t r_i [t]} \right]$$

Restless Bandits' Challenges

- Restless bandits evolve for two transition kernels

$$P_{i,act}(s_i[t]) \quad \text{For} \quad a_i[t] = 1$$

$$P_{i,pass}(s_i[t]) \quad \text{For} \quad a_i[t] = 0$$

- Exponentially growing state space in N
  - K states per arm gives K^N total possible states

- Finding optimal control policy π^* for restless bandits is ***intractable***

Index Policies

- *Decompose* original N-dimensional restless bandit problem
  - Define a state index for each bandit *independently*

- Sort indices in an ascending order. Activate M highest-indexed bandits
  - Complexity becomes $N\mathrm{logN}$

- The *Whittle index* $W(s)$ for state s is a useful tool for restless bandits
  - Asymptotically optimal control performance as $N \rightarrow \infty$
  - Difficult to calculate and unknown for most problems

Whittle Index & Indexability

- System of one bandit $N = 1$

- Agent pays an *activation cost* λ when the selected action is $\mathrm{a}[\mathrm{t}] = 1$

- Activation policy goal is to maximize the *discounted net reward*

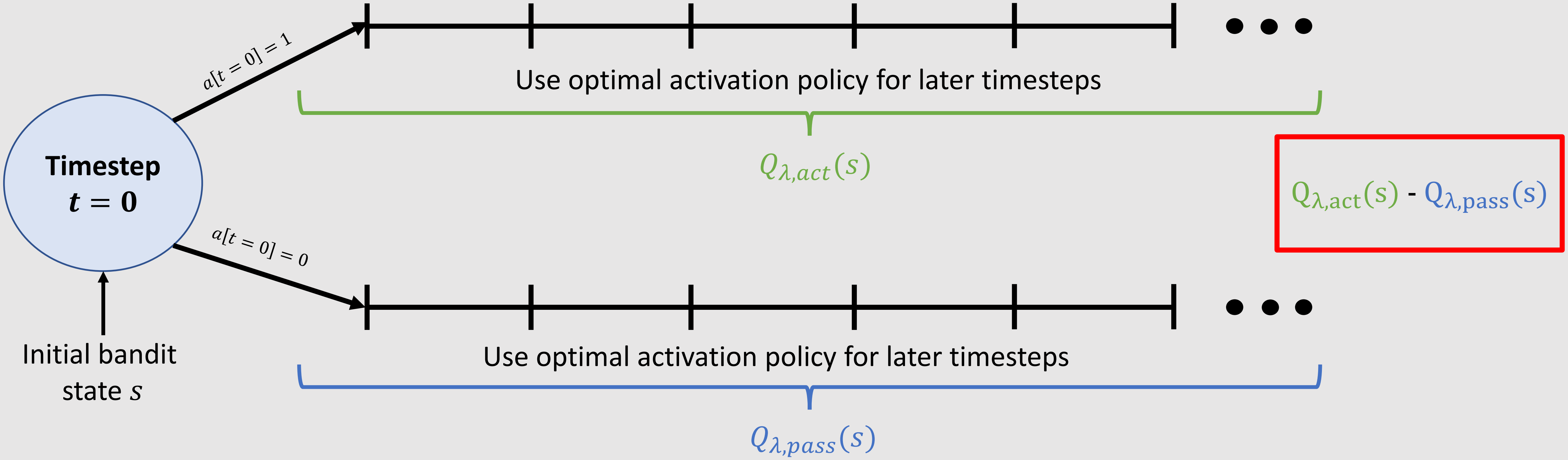
$$\mathbb{E}\,[\sum_{\boldsymbol{t=0}}^{\infty} \boldsymbol{\beta}^{\boldsymbol{t}}\,(\boldsymbol{r}[\boldsymbol{t}] - \boldsymbol{\lambda}\boldsymbol{a}[\boldsymbol{t}])]$$

- An arm becomes less likely to be activated when λ increases

$\boxed{\text{Whittle index } W(s) \text{ for state } s \text{ is the highest price the agent pays to activate the bandit}}$

Whittle Index & Indexability

Consider two activation possibilities for a single arm



For indexable arms, the difference $Q_{\lambda,act}(s)$ - $Q_{\lambda,pass}(s) \geq 0$ if $\lambda \leq W(s)$

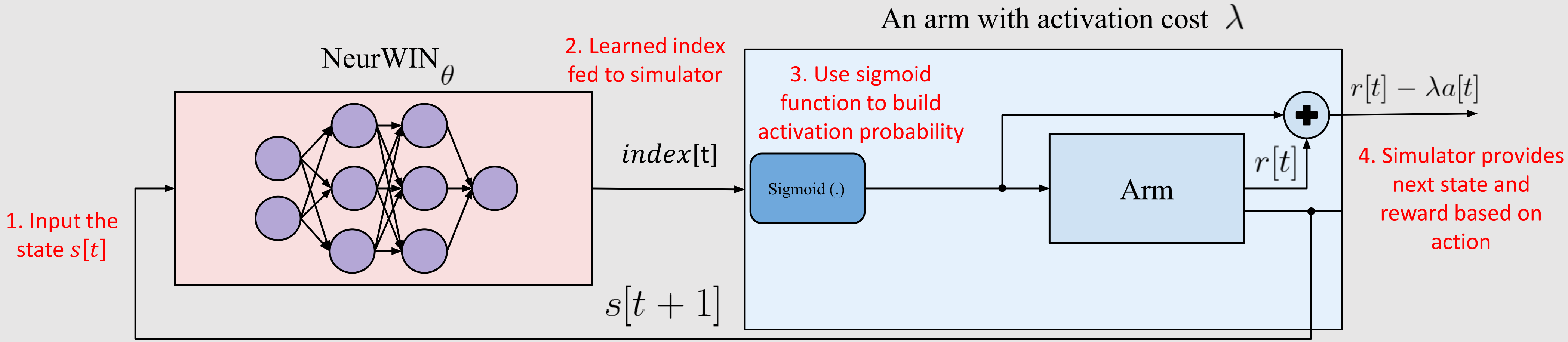
Whittle Accurate Controller

- Solving for Whittle indices yields the asymptotically optimal controller
- Use a neural network to learn Whittle index function *under all activation costs*
- Neural network that produces a *near-optimal* discounted net reward is a bandit controller

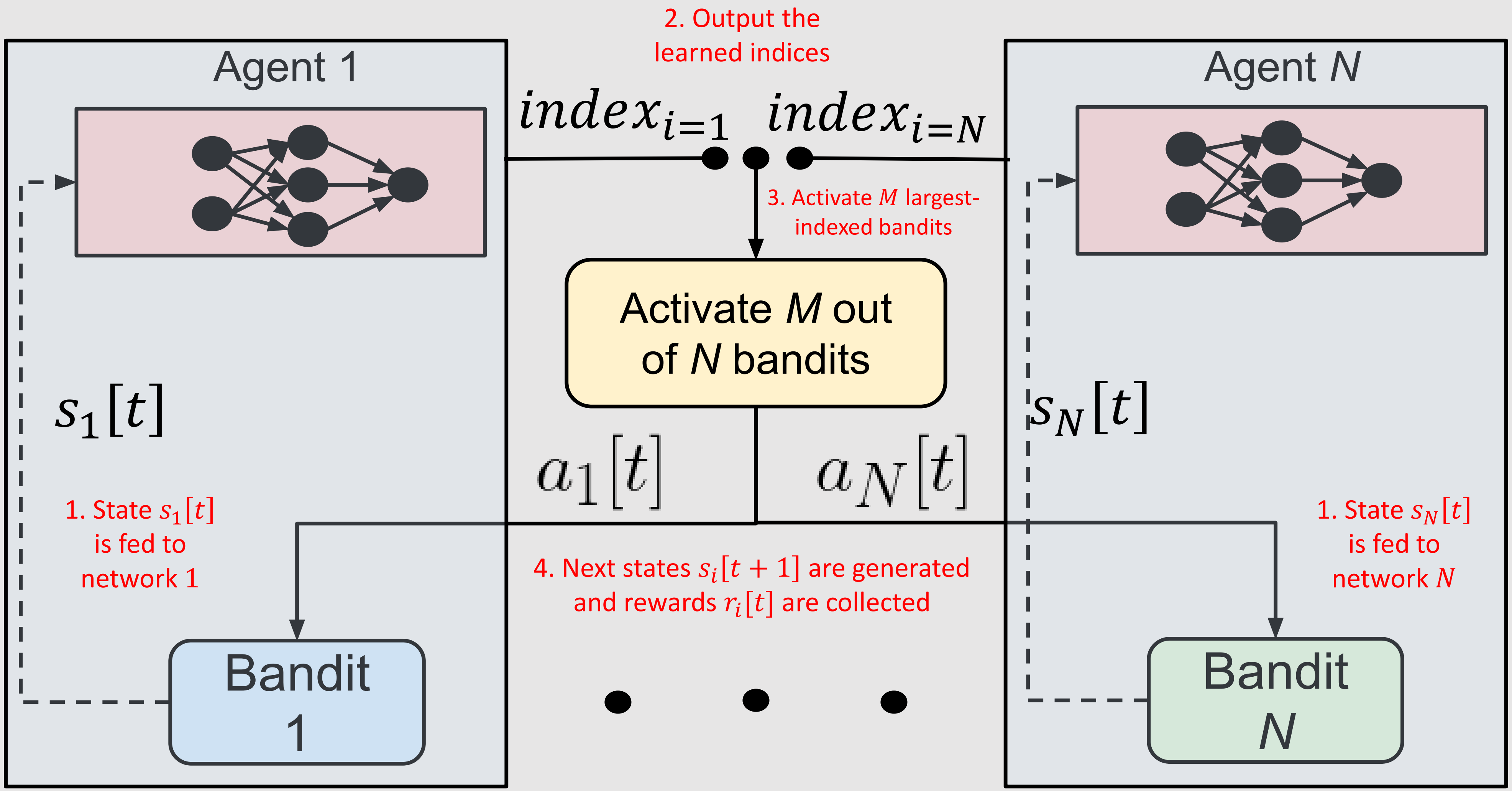
Whittle-Accurate Theorem
Near-optimal neural network for a restless bandit is also Whittle-accurate for all activation cost values and all states

NeurWIN Training

- **NeurWIN** REINFORCE-based algorithm that trains in an episodic fashion
- A minibatch of episodes have a different activation cost from previous minibatch
- Perform gradient ascent on objective function $\nabla_\theta \sum_{\lambda,s_1} \widetilde{\boldsymbol{Q}}(\boldsymbol{s_1}, \lambda)$ for all initial states s_1, λ
- Episode with time horizon T has return $\sum_{t=0}^{T-1} \beta^t \, (r[t] \, - \lambda a[t])$



NeurWIN's Control Approach



Experiment Results

- **Three recently studied restless bandits' problems**

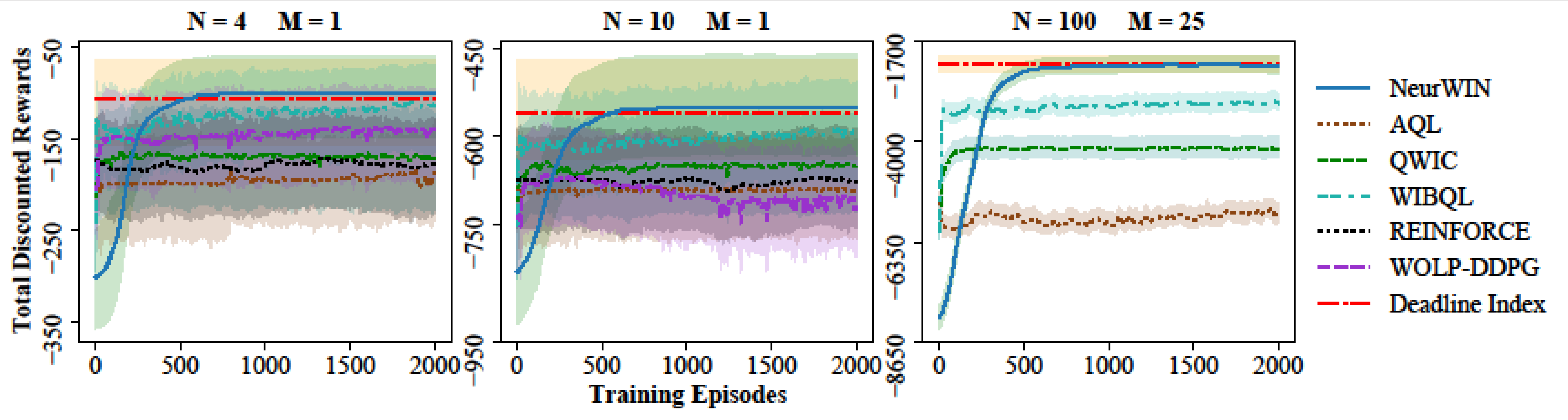
  Three sets $(N = 4, M = 1)$, $(N = 10, M = 1)$, $(N = 100, M = 25)$

- **NeurWIN training parameters**
  - Initial learning rate $L = 0.001$
  - Discount factor $\beta = 0.99$
  - Adam optimizer for gradient ascent step

- Compare with other RL algorithms and baseline from each study

Deadline Scheduling

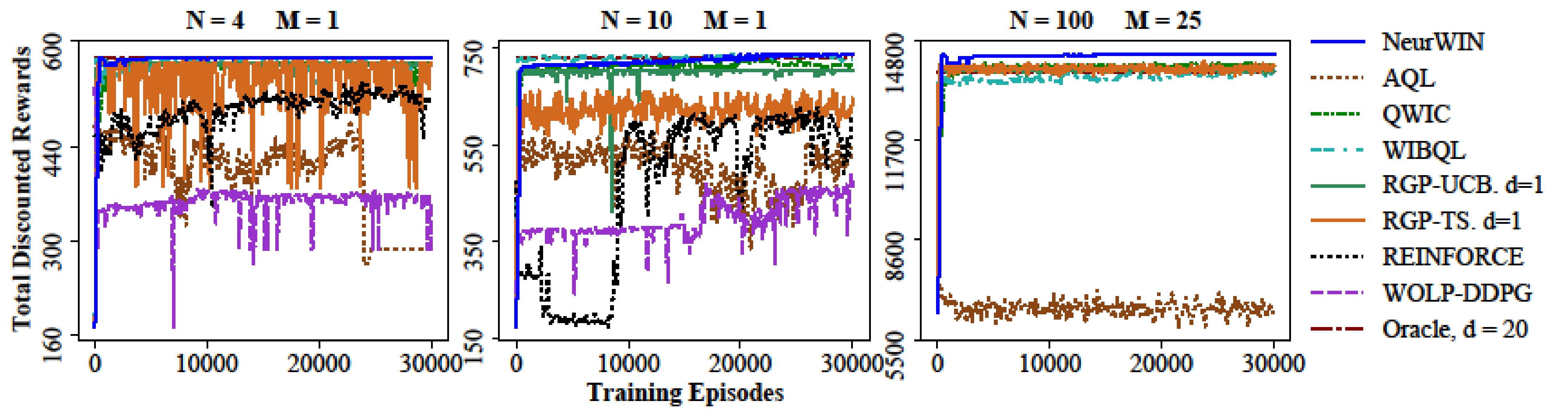
- Schedule M vehicles for N charging spots being restless bandit. Whittle index is known
- Vehicle' state is the time left until it leaves the station and electric charge needed
- Reward for charging. Penalty if car is not charged by the declared leaving time



- NeurWIN converges to Whittle index performance in approximately 600 episodes
- Other learning algorithms show no improvement

Recovering Bandits

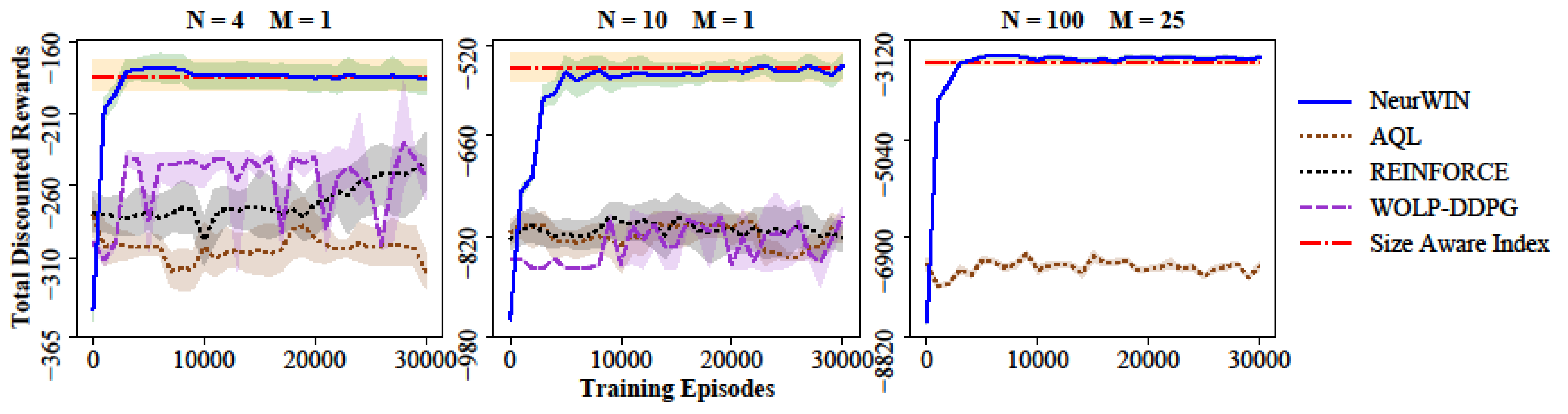
- Time-varying behavior of a customer interested in products given as N restless bandits
- Bandit state is the time since it was last activated bounded by $z_{max} = 20$ timesteps
- 20-lookahead oracle picks best leaf from a tree with 2^{20} leaves. Whittle index is unknown.



- NeurWIN outperforms all baselines in terms of total discounted rewards

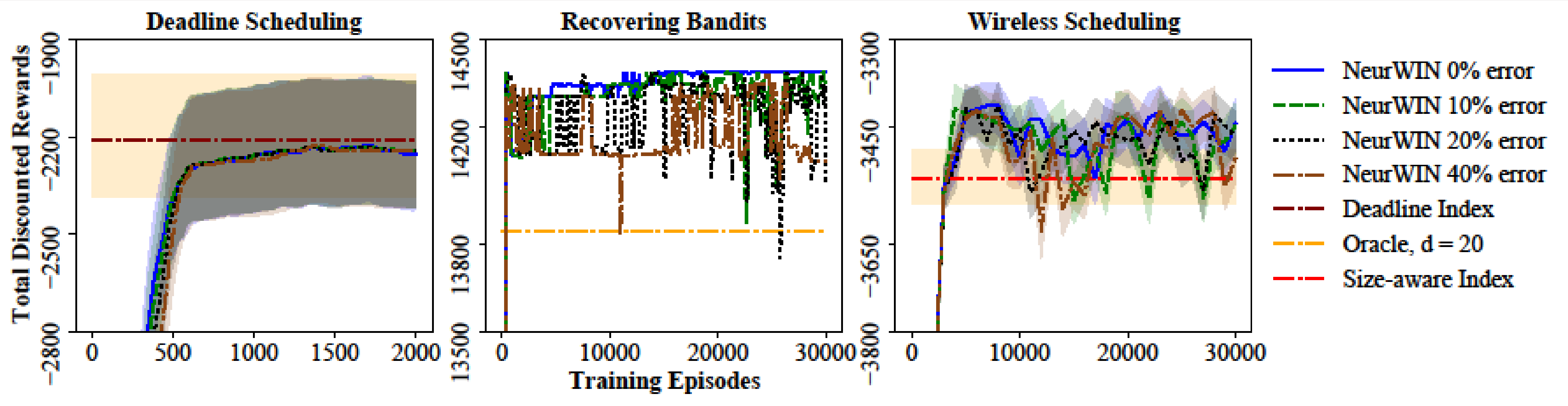
Wireless Scheduling

- Wireless scheduling over fading channels with N clients modelled as restless bandits
- Client state is the payload given in remaining bits and the current channel transmission state
- Holding cost $c = 1$ for each timestep a client's payload isn't fully transmitted.
- No known Whittle index



Results With Noisy Simulators

- Case when NeurWIN is trained on an imprecise simulator
- Added Gaussian noise of $10\%, 20\%, 40\%$
- Tested on $N = 100\ , M = 25$ setting only



- Slight performance degradation yet superior to baselines

Summary

- Asymptotically index-based optimal control policy for N restless bandits

- **Proposed NeurWIN** Deep reinforcement learning method for learning Whittle indices

- Demonstrated a superior control performance for three studies compared to baselines

Khaled Nakhleh

LinkedIn: linkedin.com/in/khalednakhleh/

Email: khaled.jamal@tamu.edu

Paper and code

ArXiv link: https://arxiv.org/abs/2110.02128

Code link: https://github.com/khalednakhleh/NeurWIN