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

Khaled Nakhleh[1], Santosh Ganji[1], Ping-Chun Hsieh[2], I-Hong Hou[1], Srinivas Shakkottai[1]

[1] Texas A&M University. [2] National Chiao Tung University

Overview

- **Setting:** N Restless bandits referenced by $i = 1,2,\dots,N$. Control policy π activates M out of N bandits in each timestep.
- **Objective:** Maximize the total discounted rewards,

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

- **Challenges:** Intractable to find the optimal control policy for restless bandits.
  - Restless bandits evolve with two kernels $P_{i,act}(s_i[t])$ and $P_{i,pass}(s_i[t])$ whether they are activated ($a[t]=1$) or left passive ($a[t]=0$).
  - Exponentially growing state space in N.
- **Approach:** Decomposition through index-based policies. Introduce NeurWIN: a deep RL algorithm that learns the *Whittle index* of a single bandit.

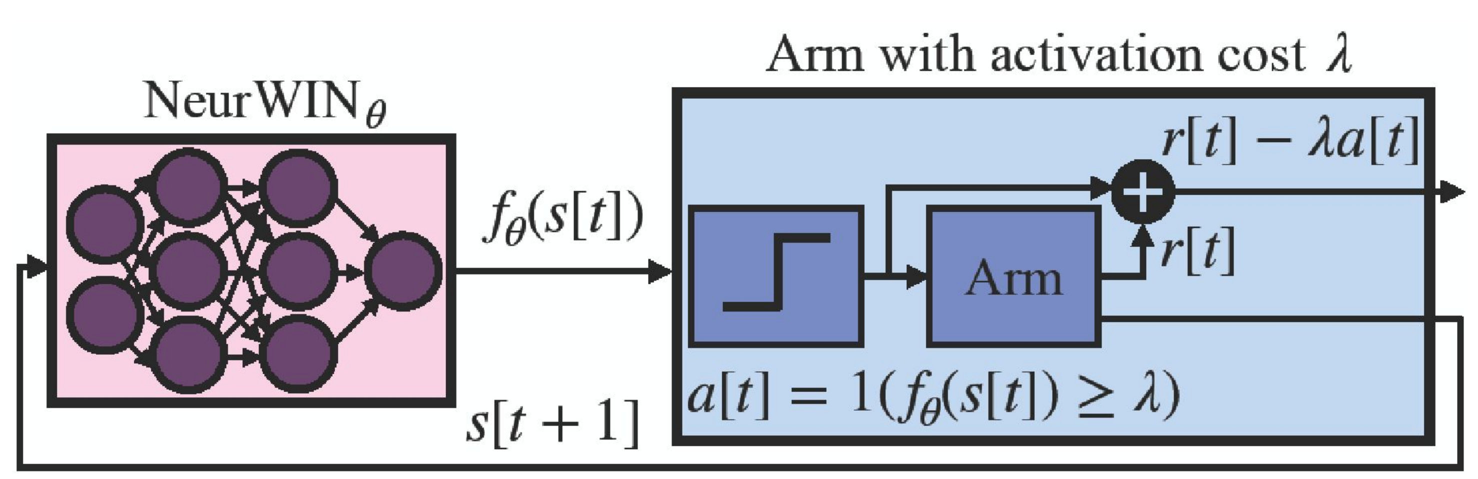
For a sequential decision-making problem, we propose an index-based control policy for *restless bandits*, which change their states at each timestep. The learning algorithm, called NeurWIN, trains a neural network on a single bandit, and assigns a *Whittle index* for each bandit's state in the state space. The control performance is asymptotically optimal in the number of bandits.

Background: Index Policies

- The index function $W_i(s_i[t])$ assigns a state index independent of other arms.
- For a single arm, an activation policy determines whether to activate the arm under a given activation cost λ.
- The activation policy objective is to maximize the total discounted net reward,

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

- Optimal activation policy activates for a states' set under a λ denoted by $S(\lambda)$.
- **Definition (Indexability):** An arm is said to be indexable if $S(\lambda)$ decreases monotonically from the set of all states to the empty set as λ increases from $-\infty$ to ∞. A restless bandit problem is said to be indexable if all arms are indexable.



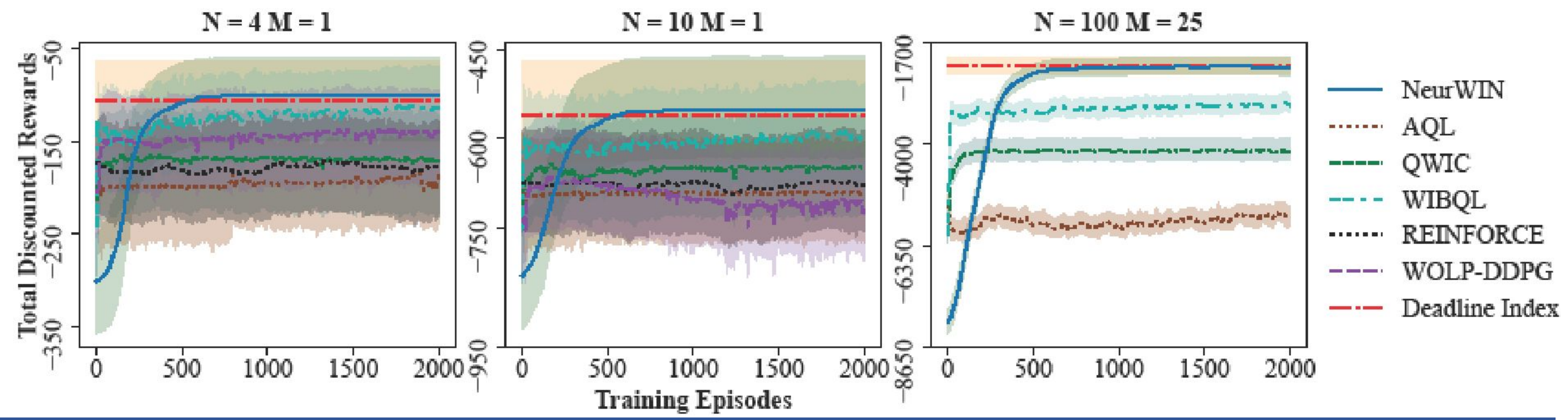
NeurWIN Training a neural network on a single simulator called $Env(\lambda)$

NeurWIN's Learned Index Performance

- Demonstrate NeurWIN's performance for three restless bandit problems.
- NeurWIN performs better than other deep RL algorithms, and respective baselines in each case

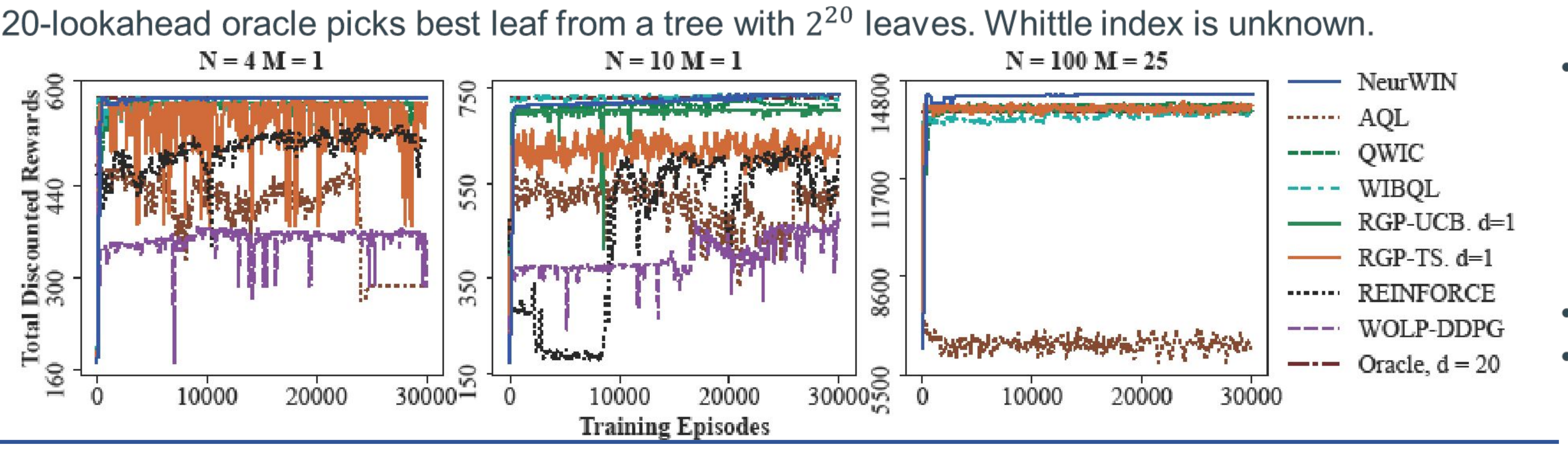
Deadline Scheduling [1]

- Vehicle charging problem, with N stations modelled as arms. M stations can be activated in a timestep.
- Problem has a closed-form Whittle index called *the deadline index*.



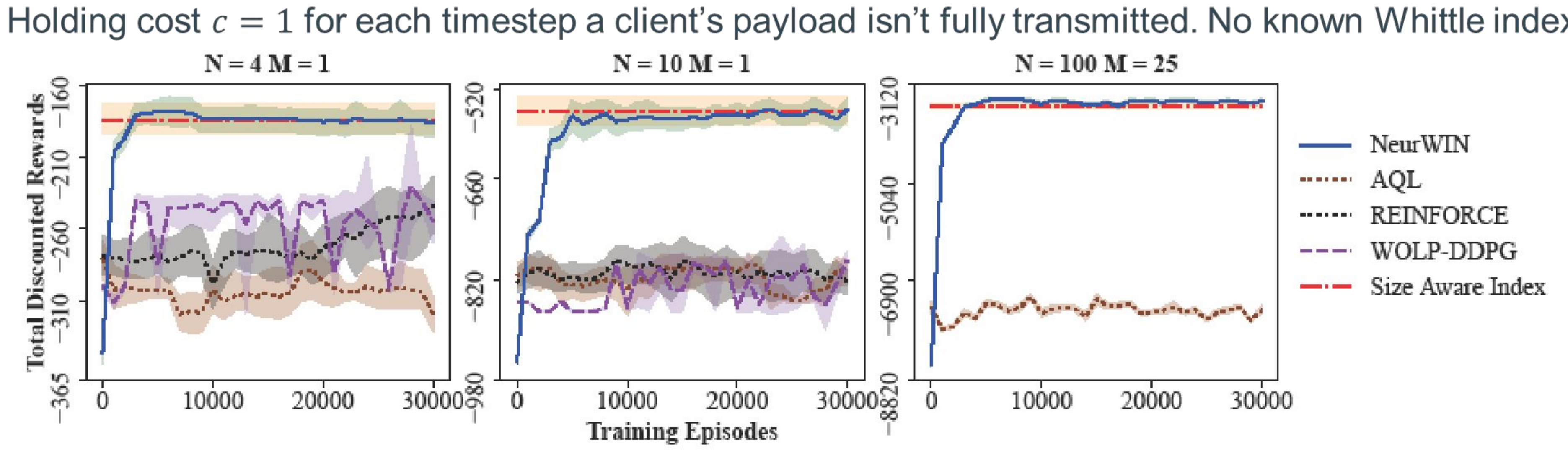
Recovering Bandits [2]

- 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.



Wireless Scheduling [3]

- 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.



γ – Whittle Accuracy

- NeurWIN trains a neural network with parameters θ to approximate the Whittle index.
- Neural network produces a real number $f_\theta(s)$ as the index while minimizing $|f_\theta(s)-W(s)|$.
- **Definition (Whittle Accurate):** A neural network with parameters θ is said to be γ-Whittle accurate if $|f_\theta(s)-W(s)|\leq\gamma$ for all s.
- Let $\tilde{Q}_\theta(s,\lambda)$ be the average reward of applying a neural network to $Env(\lambda)$ for initial state s.
- **Definition (ϵ-optimal neural network):** A neural with parameters θ is said to be ϵ-optimal if there exists a small positive number δ such that,
  $\tilde{Q}_\theta(s,\lambda)\geq\max\{Q_{\lambda,act}(s_1),Q_{\lambda,pass}(s_1)\}-\epsilon$ for all s_0, s_1, and $\lambda\in[f_\theta(s_0)-\delta, f_\theta(s_0)+\delta]$

Theorem 1. If the arm is strongly indexable, then for any $\gamma>0$, there exists a positive ϵ such that any ϵ-optimal neural network controlling $Env(\lambda)$ is also γ-Whittle-accurate.

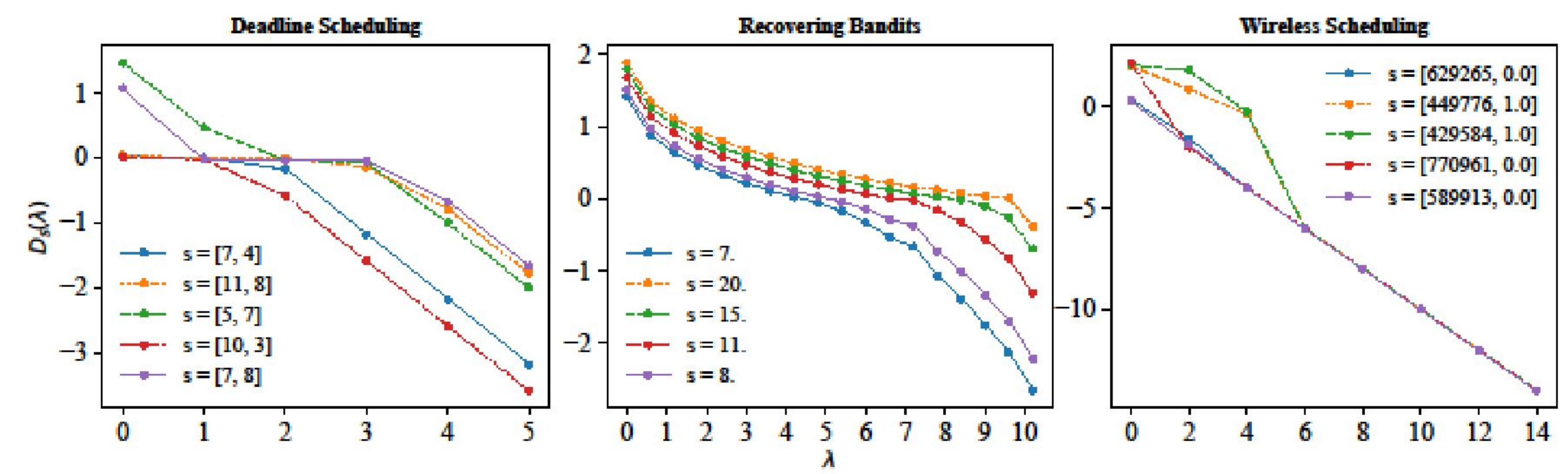


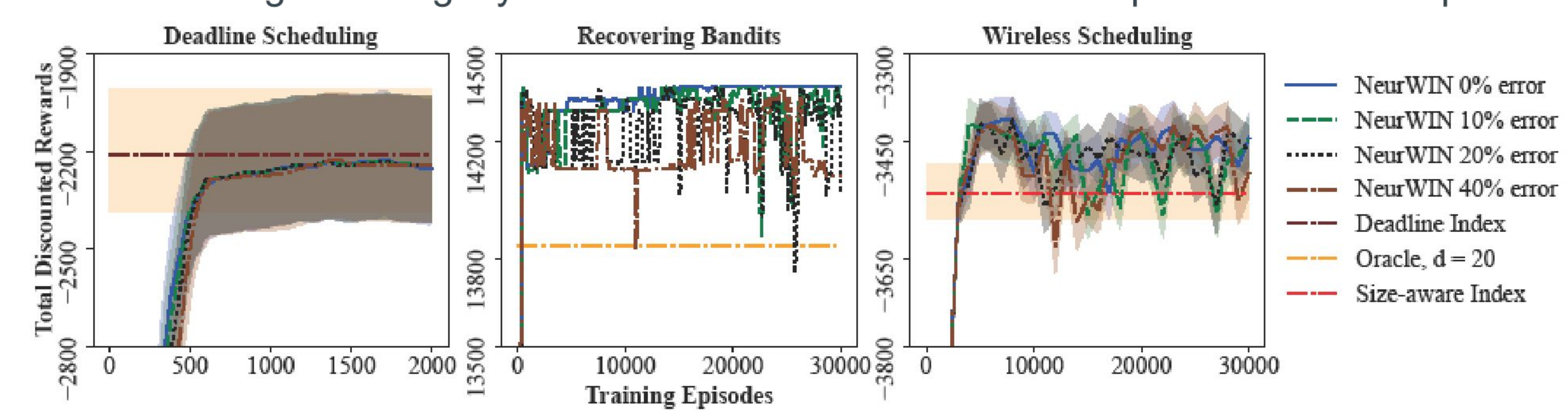
Figure showing the strong indexability condition holds for the three restless bandit problems.

NeurWIN With Noisy Simulators

- Added independent Gaussian random variables $G_{act,s}$ and $G_{pass,s}$ to rewards,

$$R'_{act}(s)=(1+G_{act,s})R_{act}(s)$$
$$R'_{pass}(s)=(1+G_{pass,s})R_{pass}(s)$$

- NeurWIN trained for different level of errors: 10%, 20%, 40%.
- Performance degrades slightly. Performance remains similar or superior to baseline policies.



Conclusion And Future Work

NeurWIN is a deep RL method for estimating the Whittle index. Performance was measured for three restless bandit problems, with it exceeding or matching state-of-the-art policies. Future work includes,

- **Offline policy extension:** construct a predictive model for each arm from offline-sampled data.
- **Non-strongly indexable cases:** adding a pre-processing step to NeurWIN to verify strong indexability. Provide performance thresholds for non-indexable arms.

References

[1] Z. Yu, Y. Xu, and L. Tong. Deadline scheduling as restless bandits. IEEE Transactions on Automatic Control, 63(8):2343–2358, 2018.

[2] Ciara Pike-Burke and Steffen Grunewalder. Recovering bandits. In Advances in Neural Information Processing Systems, pages 14122–14131, 2019.

[3] Samuli Aalto, Pasi Lassila, and Prajwal Osti. Whittle index approach to size-aware scheduling with time-varying channels. In Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pages 57–69, 2015.