Tabular and Deep Learning of Whittle Index

Francisco Robledo, Urtzi Ayesta, Konstantin Avrachenkov, Vivek S Borkar

Tabular and Deep Learning of Whittle Index

Francisco Robledo, Vivek Borkar, Urtzi Ayesta, Konstantin Avrachenkov

Introduction

- Whittle index policy is an asymptotically optimal heuristic for solving Restless Multi-Armed Bandit Problems (RMABP).
- We propose two algorithms, QWI and QWINN, for the computation of such indices.
- Both employ a two time-scale system for the computation of the indices and the Q-values of each state/action.

Motivation

Asymptotically Optimal Heuristics for Restless Multi-Armed Bandit Problems (RMABP)

- Load-balancing problems
- Machine maintenance
- Health-care systems

QWI/QWINN Algorithm

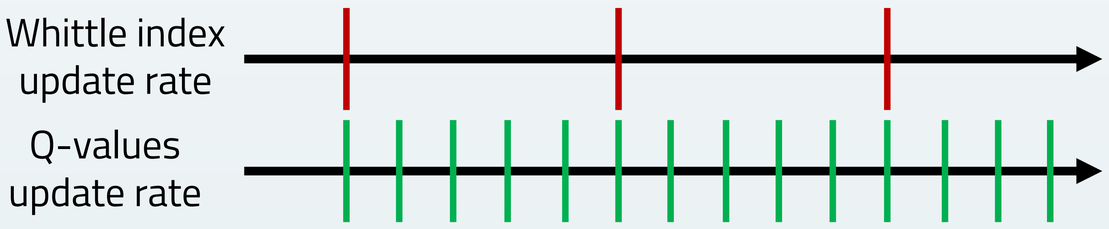
- **Algorithm**:

Evaluate the **Q-values** for a given state/action (s_n, a_n) using the index for all states $x \in \mathcal{S}$

$$Q^x_{n+1}(s_n,a_n) = \big(1-\alpha(n)\big)Q^x_{n+1}(s_n,a_n) + \alpha(n)\Big((1-a_n)\big(r_0(s_n)+\lambda_n(x)\big) + a_n r_1(s_n) + \gamma \max_{v\in\{0,1\}} Q^x_n(s_{n+1},v)\Big)$$

Update **Whittle index** for all states $x \in \mathcal{S}$

$$\lambda_{n+1}(x) = \lambda_n(x) + \beta(n) + \left(Q^x_n(x,1) - Q^x_n(x,0)\right)$$



- **Two time-scales** set through learning rate
  - Fast time-scale: Q-values

$$\alpha(n): \sum_n \alpha(n) = \infty, \qquad \sum_n \alpha(n)^2 < \infty$$

  - Slow time-scale: Whittle index

$$\beta(n): \sum_n \beta(n) = \infty, \qquad \sum_n \beta(n)^2 < \infty, \qquad \beta(n) = o\big(\alpha(n)\big)$$

Neural Network implementation (QWINN):

- Hybrid system:
  - Neural network to approximate Q-values

$$Q^x_{target}(s_n,a_n) = (1-a_n)\big(r_0(s_n)+\lambda_n(x)\big) + a_n r_1(s_n) + \gamma \max_{v\in\mathcal{A}} Q^x_{\theta'}(s_{n+1},v)$$

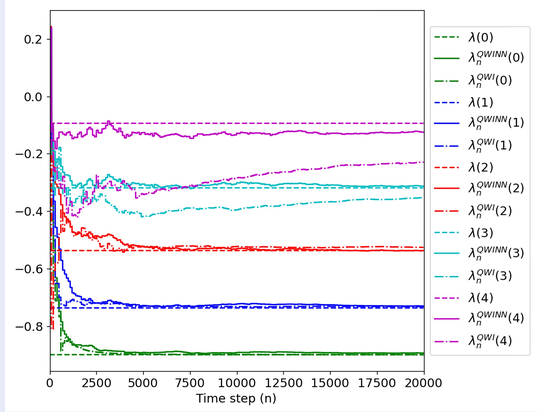
  - Tabular computation for Whittle index

$$\lambda_{\theta,n+1}(x) = \lambda_{\theta,n}(x) + \beta(n)\left(Q^x_{\theta,n}(x,1) - Q^x_{\theta,n}(x,0)\right)$$

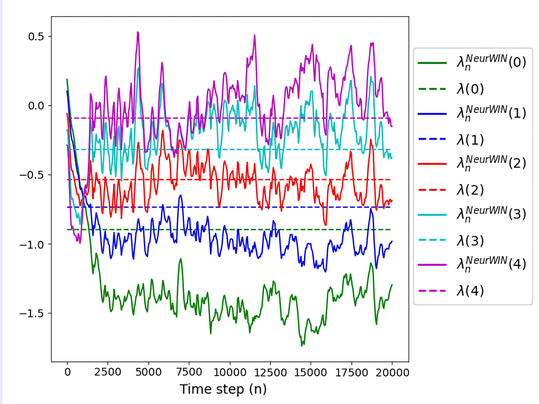
Results

Restart problem

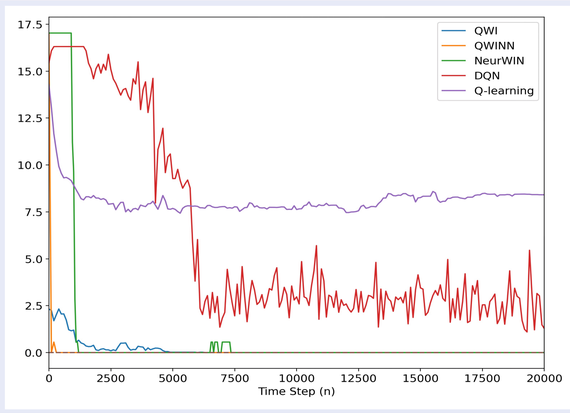
QWI/QWINN indices



NeurWIN indices [3]

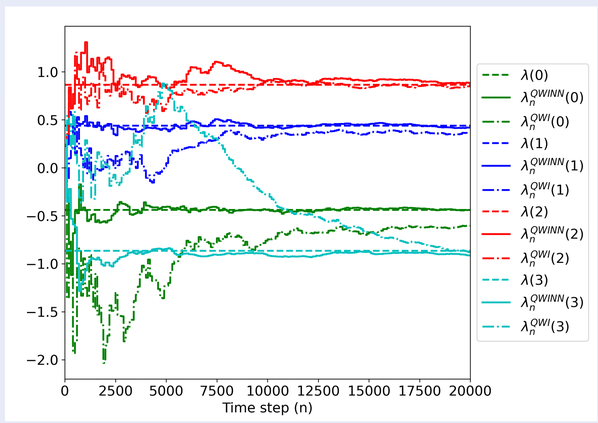


Bellman Relative Error (lower is better)

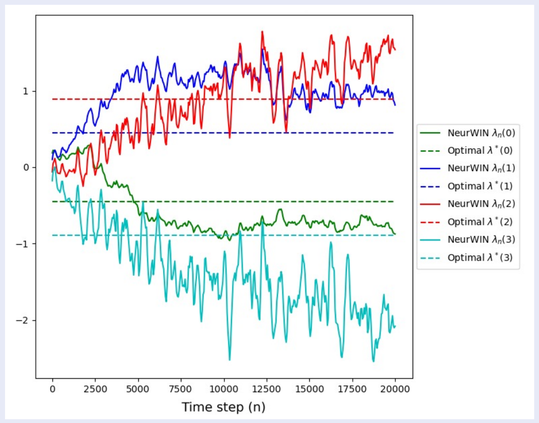


Circular problem

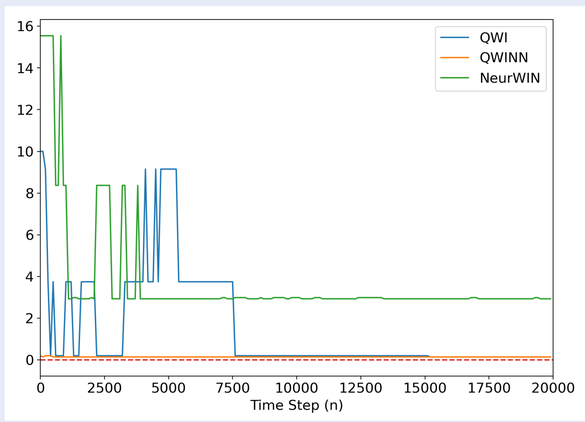
QWI/QWINN indices



NeurWIN indices [3]



Bellman Relative Error (lower is better)



Conclusion

- Two time-scale implementation to decouple Q-learning and Whittle index updates
- Fast and stable convergence to theoretical Whittle index
- Neural Network implementation improves performance in underexplored states

References

1. Abounadi, Jinane / Bertsekas, Dimitrib / Borkar, Vivek S., Learning algorithms for Markov decision processes with average cost, 2001
2. Avrachenkov, Konstantin E. / Borkar, Vivek S., Whittle index based Q-learning for restless bandits with average reward, 2022-05, Automatica , Vol. 139, p. 110186
3. Nakhleh, Khaled / Ganji, Santosh / Hsieh, Ping-Chun / Hou, I.-Hong / Shakkottai, Srinivas, NeurWIN: Neural Whittle Index Network For Restless Bandits Via Deep RL, 2021

This work was supported by the ANR LabEx CIMI (grant ANR-11-LABX-0040) within the French State Programme “Investissements d’Avenir”.