QWI: Q-learning with Whittle Index

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

ABSTRACT

The Whittle index policy is a heuristic that has shown remarkable good performance (with guaranted asymptotic optimality) when applied to the class of problems known as multi-armed restless bandits. In this paper we develop QWI, an algorithm based on Q-learning in order to learn the Whittle indices. The key feature is the deployment of two timescales, a relatively faster one to update the state-action Q-functions, and a relatively slower one to update the Whittle indices. In our main result, we show that the algorithm converges to the Whittle indices of the problem. Numerical computations show that our algorithm converges much faster than both the standard Q-learning algorithm as well as neural-network based approximate Q-learning.

1. INTRODUCTION

Markov Decision Processes (MDPs) provide a mathematical framework for sequential decision making. Formally, an MDP is a sequential stochastic control process, where a decision maker aims at maximizing its long term reward. The basic setup is as follows, at each time step, a state dependent reward is accrued, the decision maker chooses an action among the available ones, and the process randomly moves to a new state. Due to their broad applicability, MDPs are found in many areas, including artificial intelligence, economics, communications and operations research.

An MDP can be solved via dynamic programming, however, this is a computationally intractable task for realistic model sizes. As a result, classes of MDPs that are analytically tractable or possess good approximations have received a lot of attention. In this extended abstract we focus on one such class, namely, the Restless Multi-Armed Bandit Problem (RMABP), introduced in [3]. In an RMABP there are multiple concurrent projects or bandit's arms. The decision maker knows the states of all projects and the reward in every state, and aims at maximizing the long-term reward. At every decision epoch, the decision maker activates one or several projects, and the state of active and passive projects evolve stochastically.

RMABPs have become extremely popular over the years, and have been applied in many contexts, including inventory routing, machine maintenance, health-care systems, networking, etc. RMABPs cannot be solved exactly, except for some toy examples. In [3] Whittle developed a methodology to obtain heuristics by solving a relaxed version of the RMABP. The obtained heuristics, nowadays known as Whittle's index policy, rely on calculating Whittle's index for each of the projects, and activating in every decision epoch the project with highest Whittle's index. It has been reported on numerous instances that Whittle's index policy provides strikingly good performance, and it has been shown to be asymptotically optimal as the number of projects grows large [4].

Reinforcement Learning refers to the area of machine learning that aims at solving MDPs for which we do not have precise knowledge of the mathematical model. In [5], it was shown that the well known Q-learning algorithm can solve an MDP, provided that every state-action pair is visited infinitely often. In recent years there has been a technological breakthrough, and the combination of reinforcement learning with deep neural networks has allowed to accomplish large scale problems [6].

As the main contribution of this paper, we develop a reinforcement learning algorithm QWI, which is based on Q-learning and learns the Whittle Index policy for the total discounted criterion. To illustrate its performance, we consider a benchmark RMABP, restart problem, and compare numerically the performances and convergence times of Q-learning, Q-learning with neural network approximation, and QWI. The numerical results show that our algorithm converges faster than the other algorithms.

2. RESTLESS MARKOVIAN BANDITS

2.1 Problem formulation and relaxation

We consider an RMABP with N projects (or arms) and under the total discounted criterion. We denote by S_n^i the state of project i at the n-th time step, and let $r^i(S_n^i, A_n^i)$ denote the conditional expected reward obtained by the i-th project at step n. The discounting factor is given by γ.

We let A_n be the vector of actions at time-step n, with elements $A_n^i = a \in \{0,1\}$. Since we assume that we can only activate a number $M < N$ of projects at each time-step, the objective is to determine the control policy maximizing the following reward:

$$E\left[\sum_{n=1}^{\infty}\sum_{i=1}^{N}\gamma^n r^i(S_n^i, A_n^i)\right], \tag{1}$$

*Corresponding author. Authors listed in reversed alphabetical order

Copyright is held by author/owner(s).

under the constraint:

$$\sum_{i=1}^{N} A_n^i \leq M, \quad n \geq 0. \tag{2}$$

Following Whittle's development [3], we relax the constraint (2) and require that it holds only on average. The constraint can then be added to the objective function:

$$E\left[\sum_{n=1}^{\infty}\sum_{i=1}^{N}\gamma^i\left(r\left(S_n^i, A_n^i\right)+\lambda\left(1-A_n^i\right)\right)\right], \tag{3}$$

where λ is a Lagrange multiplier associated to the constrained.

The solution to (3) is obtained by combining the solution to N independent problems. In other words, for each bandit i we need to solve the associated Bellman equation given by:

$$\begin{aligned} V^i(s) = \max_{a\in\{0,1\}} \Bigg[& a\left(r^i(s,1)+\gamma\cdot\sum_j p^i(j|s,1)V^i(j)\right)+ \\ & (1-a)\left(r^i(s,0)+\lambda+\gamma\cdot\sum_j p(j,|s,0)V^i(j)\right)\Bigg] \end{aligned} \tag{4}$$

where $V^i(s)$ is the value function corresponding to the initial state s. The optimal action in (4) will be to activate the bandit if $r^i(k,1)+\gamma\cdot\sum_j p^i(j|k,1)V^i(j) > r^i(k,0)+\lambda+\gamma\cdot\sum_j p(j,|k,0)V^i(j)$, while otherwise the optimal action will be to keep that project passive. For this reason, the multiplier λ can be seen as a subsidy for passivity.

2.2 Whittle index

We can rewrite the expression in the square brackets of equation (4) as a function of the state-action pair in the following manner:

$$\begin{aligned} Q^i(s,a) =& a\left(r^i(s,a)+\gamma\cdot\sum_j p^i(j|s,a)V^i(j)\right)+ \\ & (1-a)\left(r^i(s,a)+\lambda+\gamma\cdot\sum_j p(j|s,a)V^i(j)\right), \end{aligned} \tag{5}$$

where we have omitted the dependency on project i.

Following Whittle's approach, provided that the technical condition known as *indexability* holds [3], the solution to (3) is characterized by a set of indices $\lambda^*(\cdot)$ known as Whittle indices. A good suboptimal policy will be to activate at every time-step M projects with the largest Whittle indices. Whittle then [3] introduced the index as the value of $\lambda^*(k)$ of λ such that both the passive and active actions are equally preferred for a given state k, namely:

$$Q(k,1)-Q(k,0)=0. \tag{6}$$

3. QWI ALGORITHM DESCRIPTION

We describe now our algorithm QWI that aims at learning both the state-action values in (5) and the Whittle indices of (6).

For the calculation of the Q-values we have

$$\begin{aligned} Q_{n+1}^x(S_n,A_n) \leftarrow & Q_n^x(S_n,A_n)+\alpha(n) \\ & \Big[(1-A_n)(r(S_n,0)+\lambda_n(x))+A_n r(S_n,1)+ \\ & \gamma\max_{b\in\{0,1\}} Q_n^x(S_{n+1},b)-Q_n^x(S_n,A_n)\Big], \end{aligned} \tag{7}$$

where the learning step must satisfy the usual conditions $\sum_n \alpha(n)=\infty, \sum_n \alpha^2(n)<\infty$. For the Whittle index, we have

$$\lambda_{n+1}(x)=\lambda_n(x)+\beta(n)\left(Q_n^x(x,1)-Q_n^x(x,0)\right), \tag{8}$$

where $\sum_n \beta(n)=\infty, \sum_n \beta^2(n)<\infty$.

A key requirement is that the algorithm operates in two time scales. On the fast time scale, the Q-values are updated similarly as in the standard Q-learning algorithm, whereas on the slower time scale, the Whittle index is updated. In order to achieve this we require that $\beta(n)=o(\alpha(n))$. In our numerical experiments, and based on the experience reported in [1], we will use $\alpha(n)=\frac{1}{\lceil\frac{n}{5000}\rceil}$ and $\beta(n)=\frac{1}{1+\lceil\frac{n\log n}{5000}\rceil}I\{n(\bmod 100)\equiv 0\}$. Thus, $\beta\neq 0$ only once every 100 iterations.

We make throughout the following assumption of 'sufficient exploration': $\liminf_{n\uparrow\infty}\frac{1}{n}\sum_{m=1}^{n} I\{S_m=s, A_m=a\}>0$, for all s,a. That is, each state-action pair is sampled 'comparably often'. We can now state the main theoretical result of the paper:

PROPOSITION 3.1. *Given the two time-scale iterations* (7)-(8) *with stepsizes* $\{\alpha(n)\}$ *and* $\{\beta(n)\}$ *satisfying* $\sum_n \alpha(n)=\infty$, $\sum_n \alpha^2(n)<\infty$, $\sum_n \beta(n)=\infty, \sum_n \beta^2(n)<\infty$ *and* $\beta(n)=o(\alpha(n))$, *under the assumptions of RMABP indexability and sufficient exploration,* $\lambda_n(k)\to\lambda^*(k)$, $\forall k$, *a.s. as* $n\to\infty$.

The proof of the convergence of this two-time scale system is based on the results in [2]. We also note that an algorithm converging to the Whittle indices for the average long-run reward criterion is given in [1]. The proof for the present discounted case will be presented in the full-length version this work.

4. NUMERICAL RESULTS

To assess the performance of our algorithm we consider the benchmark "restart problem" studied in [1] with a state space $S=\{0,1,2,3,4\}$. Active action ($a=1$) takes the project to the initial state 0 with probability 1. In the case of the passive action, the transition matrix is given by

$$P_0=\begin{pmatrix} 1/10 & 9/10 & 0 & 0 & 0 \\ 1/10 & 0 & 9/10 & 0 & 0 \\ 1/10 & 0 & 0 & 9/10 & 0 \\ 1/10 & 0 & 0 & 0 & 9/10 \\ 1/10 & 0 & 0 & 0 & 9/10 \end{pmatrix}.$$

The conditional expected reward is given by $r(s,a)=0.9^{s+1}$ for $a=0$, and $r(s,a)=0$ for $a=1$.

Throughout this section, we have considered the case of homogeneous projects, i.e., all projects have the same dynamics and reward functions.

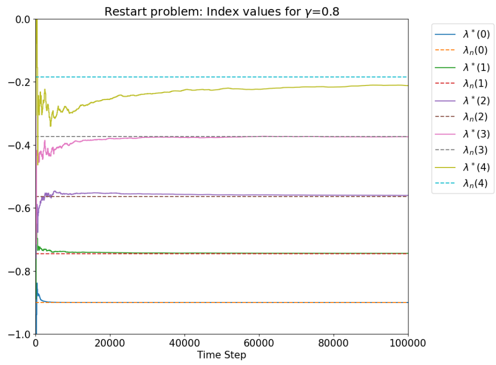

Figure 1: Convergence of the Whittle indices

We consider there are 5 projects, and we assess numerically the convergence stated in Proposition 3.1. Figure 1 depicts the evolution of the estimated Whittle indices with respect to the time step. Each of the lines represents the average value of the estimates of all the projects for a given state. We observe that the convergence for state 5 requires many more steps. However, from very early on, the estimated indices are ordered correctly.

We now compare the performance of QWI with respect to the standard Q-learning [5], and Q-learning with neural network approximation (RLNN) [6]. We use a neural network with two hidden layers network, with 16 neurons in each of them.

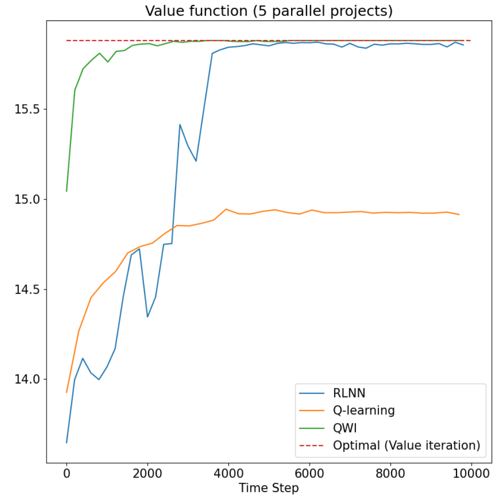

Figure 2: Bellman's Value Function (4) evolution during training for the RLNN, Q-learning and QWI algorithms using 5 projects

In Figure 2 we consider the case of 5 projects, $\gamma = 0.8$ and an exploration parameter $\epsilon = 0.3$ in QWI and RLNN algorithms and $\epsilon = 1.0$ for Q-learning. We note that Q-learning does not converge to the correct value of the value function, but with $\epsilon = 0.3$, the convergence of Q-learning was even slower than in the case of $\epsilon = 1.0$. Both RLNN and QWI algorithms succeed in converging to the optimal value function, but QWI, which takes advantage of the "restless" structure of the problem, converges faster.

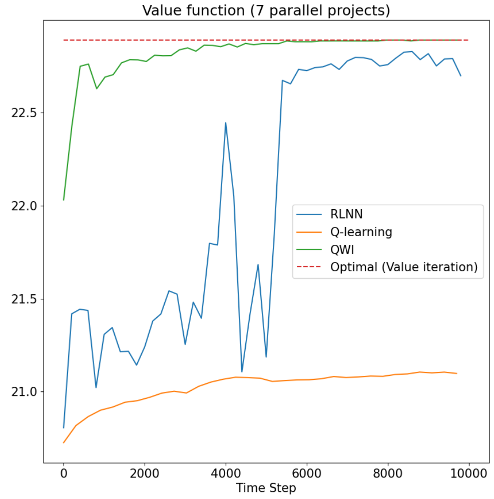

Figure 3: Bellman's Value Function (4) evolution during training for the RLNN, Q-learning and Whittle index algorithms using 7 projects

In Figure 3 we consider the case of 7 projects. We observe that the benefits of QWI with respect to the two other algorithms increases. Once again Q-learning does not converge to the optimal value function, and RLNN performs better than Q-learning, but it does not converge to the optimal values neither. On the other hand, QWI converges to the correct values.

5. CONCLUSIONS

We have presented QWI, an algorithm that converges to the Whittle indices of any indexable RMABP. An advantage of QWI with respect to standard Q-learning and neural-network based approximations, is that it uses effectively the decomposition structure of RMAB to tackle the curse of dimensionality. In our numerical section we have considered a benchmark RMABP, and showed that QWI indeed outperforms the other algorithms. In future work we plan to investigate in more depth the performance of QWI with more realistic and complex RMABPs.

6. REFERENCES

[1] K. Avrachenkov and V.S. Borkar, Whittle index based *Q*-learning for restless bandits with average reward arXiv preprint arXiv:2004.14427 (2020)

[2] C. Lakshminarayanan and S. Bhatnagar, A stability criterion for two timescale stochastic approximation schemes Automatica **79**, 108-114 (2017)

[3] P. Whittle, Restless bandits: Activity allocation in a changing world. Journal of applied probability, 287–298 (1988)

[4] R.R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27:637–648, 1990.

[5] C.J.C.H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–202, 1992.

[6] V. Mnih, et al. Human-level control through deep reinforcement learning. *Nature* 518.7540: 529-533, 2015.