

Whittle index based Q-learning for restless bandits with average reward

Konstantin Avrachenkov*, Vivek S. Borkar†

*Inria Sophia Antipolis, 2004 Route des Lucioles, Valbonne 06902, France, k.avrachenkov@inria.fr

†Indian Institute of Technology, Powai, Mumbai, 400076, India, borkar.vs@gmail.com

Abstract—A novel reinforcement learning algorithm is introduced for multiarmed restless bandits with average reward, using the paradigms of Q-learning and Whittle index. Specifically, we leverage the structure of the Whittle index policy to reduce the search space of Q-learning, resulting in major computational gains. Rigorous convergence analysis is provided, supported by numerical experiments. The numerical experiments show excellent empirical performance of the proposed scheme.

Keywords—reinforcement learning; restless bandits; Whittle index; Q-learning; average reward

I. INTRODUCTION

Restless bandits have found in recent times numerous applications for various scheduling and resource allocation problems, such as web crawling [6], [5], [30], congestion control [7], [8], [9], queueing systems [4], [16], [20], [25], cluster and cloud computing [17], [29], wireless communication [1], [26], [31], machine maintenance [19], target tracking [28] and clinical trials [36]. See [18], [22], [32] for book-length accounts of theory and applications of restless bandits. While restless bandits can be viewed as a special case of classical Markov decision processes, this suffers from curse of dimensionality because the state space grows exponentially in the number of arms. In fact it is provably hard in the sense of belonging to the complexity class PSPACE [33]. One very successful heuristic in this context has been the celebrated Whittle index policy [39], which relaxes the ‘hard’ constraint of using a certain number of arms at each time, to doing so on the average. Thereby it allows a decoupling of the problem into multiple individual controlled Markov chains via the Lagrange multiplier, using the fact that both the reward and the constrained functional are separable. This leads to a state space that grows linearly in the number of arms. These chains are coupled through the control policy based on ordinal comparison of a scalar function of their individual states, viz., the so called Whittle index. While this is known not to be optimal in general, it works very well in practice and is asymptotically optimal in a certain sense [25], [38].

The use of Whittle index policy, however, requires full knowledge of the system, both in the relatively few cases

where it is known explicitly (e.g., in [5]) or when it has to be numerically calculated. This is often not the case in practice, sometimes called the ‘curse of modelling’. The uncertainties can be either parametric, structural, or both. In either case, the classical adaptive control schemes (e.g., [21], [23]) or off-the-shelf reinforcement learning schemes (e.g., [10], [11], [35]) become computationally unmanageable if applied directly to restless bandits. These schemes typically do not exploit the special structure available in the problem, in this case the Whittle index policy. Motivated by this observation, the present authors have devised reinforcement learning schemes specifically tailored for the Whittle index policy. One of them is a scheme wherein the threshold structure implicit in arriving at the index policy in many problems is exploited by sampling candidate thresholds [15], whereas in [6], the monotone dependence between various factors are used to arrive at a parametric scheme that takes advantage of the explicit expression for the Whittle index where available. More recently, [27] proposed an ad hoc scheme motivated by Q-learning which, however, lacks convergence guarantees. Also, their numerical evidence does not suggest convergence to the correct values of the Whittle indices.

In this work, we make a departure from these works by combining the Q-learning algorithm for average reward [3] with a tuning scheme for the Whittle indices. This yields a provably convergent learning algorithm with excellent empirical performance on test cases. In case the arms are statistically identical, the algorithm is particularly economical because it learns the common Q-values and Whittle index. The algorithm has a notably simple form compared to above works, and can be executed in both on-line and off-line modes, the latter allowing for off-policy iterations. Specific comparisons with the predecessors are as follows:

- 1) In [15], the threshold is treated as an independently sampled additional state variable. This blows up the dimensionality of the problem. Further, the convergence of the parameter tuning scheme, which involves an ad hoc simplification, is not rigorously established. In contrast, the present work has a rigorous theory behind it and is economical with the state space as compared to [15].
- 2) In [6], only parametric uncertainties can be handled. That is, it is assumed that an explicit analytic expression for the Whittle index is available and only the parameters of the underlying stochastic processes are unknown.

Work is supported in part by the DST-Inria project “Machine Learning for Network Analytics” IFC/DST-Inria-2016-01/448. VB is also supported in part by a J. C. Bose Fellowship from the Government of India and KA is also supported in part by Nokia Bell Labs and ANSWER project PIA FSN2 (P15 9564-266178 \D0S0060094).

The present work requires no such assumptions. The numerical experiments therein also indicated some stability issues with the algorithm and needed additional tweaks.

- 3) The work [27] uses an ad hoc scheme based on Q-learning which lacks rigorous justification and the numerical experiments show that it has problems with convergence.

The paper is organized as follows. The next section summarizes the Whittle index formalism. Section 3 describes the algorithm in detail for statistically identical arms. Extension to non-homogeneous scenarios is straightforward. Section 4 presents numerical experiments. Section 5 provides convergence analysis, which relies upon [3], [12], [14] and [24].

II. WHITTLE INDEX FOR RESTLESS BANDITS

Consider $N > 1$ controlled Markov chains $\{X_n^i, n \geq 0\}$, $1 \leq i \leq N$, on a discrete state space $S = \{1, 2, \dots, d\}$, $1 < d < \infty$. The controlled transition kernel

$$(k, j, u) \in S^2 \times \{0, 1\} \mapsto p^i(j|k, u) \in [0, 1]$$

for the i th chain satisfies $\sum_j p^i(j|k, u) = 1 \forall i, k, u$, and has the interpretation of ‘probability of going from state k to state j under control u ’. The control variable u is binary. This has the interpretation of two modes of operation, active ($u = 1$) and passive ($u = 0$). These chains together constitute a ‘restless bandit’, deemed so because they evolve dynamically even in the passive mode unlike classical multi-armed bandits [18]. Define the increasing family of σ -fields $\mathcal{F}_n := \sigma(X_m^j, U_m^j, 1 \leq j \leq N, m \leq n)$, $n \geq 0$. The ‘controlled Markov property’ then corresponds to

$$P(X_{n+1}^i = k | \mathcal{F}_n) = p^i(k | X_n^i, U_n^i), \quad \forall n,$$

where $U_n^i, n \geq 0, 1 \leq i \leq N$, are the $\{0, 1\}$ -valued control processes, called ‘admissible controls’. A special subclass of interest to us is stationary policies wherein $U_n^i = \varphi^i(X_n^i)$ for some $\varphi^i : S \mapsto \{0, 1\}$, $n \geq 0$. The individual chains are called ‘arms’ of the restless bandit. Let $R_u^i : S \mapsto [0, \infty)$, $u = 0$, resp. 1, denote prescribed per stage reward functions for passive, resp. active, mode for the i th chain. These controlled Markov chains are assumed to satisfy:

(C0) (Unichain property) There exists a distinguished state $i_0 \in S$ that is reachable with strictly positive probability from any other state under any stationary policy.

Since $d = |S| < \infty$, this implies in particular that for $\tau := \min\{n \geq 0 : X_n^i = i_0\}$,

$$\sup_k E[\tau | X_0^i = k] < \infty. \quad (1)$$

The objective is to maximize the long run average reward

$$\liminf_{n \uparrow \infty} \frac{1}{n} E \left[\sum_{i=1}^N \sum_{m=0}^{n-1} R_{U_m^i}^i(X_m^i) \right], \quad (2)$$

subject to the constraint, for prescribed $M < N$,

$$\sum_i U_n^i = M, \quad \forall n. \quad (3)$$

That is, at each time instant, only M arms are activated.

Since the problem is provably hard, Whittle’s relaxation is to replace the ‘per time instant’ constraint (3) by a ‘time-averaged constraint’

$$\liminf_{n \uparrow \infty} \frac{1}{n} E \left[\sum_{m=0}^{n-1} U_m^i \right] = M. \quad (4)$$

This renders it a classical ‘constrained Markov decision process’ [2]. While this is a significant simplification, the problem is still unwieldy. Whittle’s ingenious observation was to use the fact that it is a problem with separable cost and constraint and invoke the Lagrangian relaxation to decouple it into individual control problems given the Lagrange multiplier λ . That is, we consider now the unconstrained control problem of maximizing

$$\liminf_{n \uparrow \infty} \frac{1}{n} E \left[\sum_{m=0}^{n-1} (R_{U_m^i}^i(X_m^i) + \lambda(1 - U_m^i)) \right] \quad (5)$$

separately for each i . (We have dropped a constant factor of $(M - 1)\lambda$ from the reward so as to match it with Whittle’s set-up.) The dynamic programming equation then is

$$\begin{aligned} V^i(k) &= \max \left(R_1^i(k) + \sum_j p^i(j|k, 1) V^i(j), \right. \\ &\quad \left. R_0^i(k) + \lambda + \sum_j p^i(j|k, 0) V^i(j) \right) - \beta^i \\ &= \max_{u \in \{0, 1\}} \left(u(R_1^i(k) + \sum_j p^i(j|k, 1) V^i(j)) + (1 - u) \times \right. \\ &\quad \left. (R_0^i(k) + \lambda + \sum_j p^i(j|k, 0) V^i(j)) \right) - \beta^i, \end{aligned} \quad (6)$$

with $(V^i(\cdot), \beta^i) \in \mathcal{R}^d \times \mathcal{R}$ the unknown variables. Under **(C0)**, β^i is unique and equals the optimal reward [34]. V is unique up to an additive constant. Since we are considering the case when all arms are statistically identical with identical reward functions, V^i, β^i, r^i , are independent of the superscript i , which will be dropped henceforth. The optimal decision $u^*(x)$ in state x then is given by the maximizer in the right hand side of (6). Define the Q-value as

$$\begin{aligned} Q(k, u) &:= u(R_1(k) + \sum_j p(j|k, 1) V(j)) + (1 - u) \times \\ &\quad (R_0(k) + \lambda + \sum_j p(j|k, 0) V(j)) - \beta. \end{aligned} \quad (8)$$

This satisfies the equation

$$\begin{aligned} Q(k, u) &= uR_1(k) + (1 - u)(\lambda + R_0(k)) - \\ &\quad \beta + \sum_j p(j|k, u) \max_v Q(j, v), \end{aligned} \quad (9)$$

for $i \in S, u \in U$. Under **(C0)**, this has a solution (Q, β) where β is uniquely specified as the optimal reward and Q

is unique up to an additive scalar, just as for (6). The set $\{j \in S : u^*(j) = 1\}$ is the set of states when the arm is active, its complement being the set of states when it is passive. Whittle's insight was to view the Lagrange multiplier as a 'subsidy' for passivity. He defined the problem to be indexable when the set of passive state increases monotonically from the empty set to all of S as the subsidy is increased from $-\infty$ to ∞ . In this case, he defines the (Whittle) index to be the value $\lambda^*(k)$ of λ for which both active and passive modes are equally preferred in the current state k . That is,

$$\begin{aligned} \lambda^*(k) &:= R_1(k) + \sum_j p(j|k, 1)V(j) \\ &\quad - R_0(k) - \sum_j p(j|k, 0)V(j) \\ &= -(Q(k, 1) - Q(k, 0)). \end{aligned} \quad (10)$$

Our algorithm does the following. It is a two time scale iteration wherein the faster timescale performs Q-learning for a 'static' subsidy λ_n , the latter in reality being only 'quasi-static', i.e., changing on a slower time scale. Thus it tracks the Q-value corresponding to the slowly changing subsidy, which in turn is updated on a slower timescale by a simple tuning scheme suggested by (10). Note that the Whittle index is a function of $k \in S$, so for large state spaces, one may compute it for a suitably chosen subset of S and extrapolate.

III. Q-LEARNING FOR WHITTLE INDEX

Q-learning is one of the oldest and most popular reinforcement learning scheme for approximate dynamic programming, due to Watkins [37]. Originally developed for infinite horizon discounted rewards, we shall be using a variant for average reward from [3]. For the controlled Markov chain $\{X_n^i\}$ above with average reward (5), the 'RVI Q-learning' algorithm of (2.7), [3], is as follows (with a key difference we highlight later). Fix a stepsize sequence $\{\alpha(n)\}$ satisfying $\sum_n \alpha(n) = \infty$ and $\sum_n \alpha(n)^2 < \infty$. For each $i \in S$, $u \in \{0, 1\}$, do:

$$\begin{aligned} Q_{n+1}(i, u) &= Q_n(i, u) + \\ &\quad \alpha(\nu(i, u, n))I\{X_n = i, Z_n = u\} \times \\ &\quad \left((1-u)(R_0(i) + \lambda) + uR_1(i) + \right. \\ &\quad \left. \max_{v \in \{0, 1\}} Q_n(X_{n+1}, v) - f(Q_n) - Q_n(i, v) \right), \end{aligned} \quad (11)$$

where¹

$$f(Q) = \frac{1}{2|S|} \sum_{i \in S} (Q(i, 0) + Q(i, 1)).$$

Here for $i \in S, u \in \{0, 1\}$,

$$\nu(i, u, n) = \sum_{m=0}^n I\{X_m = i, Z_m = u\},$$

is the 'local clock' for the pair (i, u) counting the updates of the (i, u) th component.

Our objective is to learn the Whittle index, i.e., the value $\lambda^*(x)$ of λ for which active and passive modes are equally desirable for a given $x \in S$. Hence we also have an updating scheme for λ , leading to a coupled iteration for each $x \in S$. The first component is the same as (11) except for the replacement of λ by the estimated Whittle index $\lambda_n(x)$, i.e.,

$$\begin{aligned} Q_{n+1}^x(i, u) &= Q_n^x(i, u) + \\ &\quad \alpha(\nu(i, a, n))I\{X_n = i, Z_n = u\} \times \\ &\quad \left((1-u)(R_0(i) + \lambda_n(x)) + uR_1(i) + \right. \\ &\quad \left. \max_{v \in \{0, 1\}} Q_n^x(X_{n+1}, v) - f(Q_n^x) - Q_n^x(i, u) \right) \end{aligned} \quad (12)$$

along with an update for learning the Whittle index $\lambda(x)$ for state x given by: with a prescribed stepsize sequence $\{\gamma(n)\}$ satisfying $\sum_n \gamma(n) = \infty$, $\sum_n \gamma(n)^2 < \infty$ and $\gamma(n) = o(a(n))$, do

$$\lambda_{n+1}(x) = \lambda_n(x) + \gamma(n) (Q_n^x(x, 1) - Q_n^x(x, 0)). \quad (13)$$

The control actions at time n are defined as follows: Let $0 \leq \epsilon < 1$ be prescribed. With probability $(1 - \epsilon)$, we sort arms in the decreasing order of the estimated Whittle index $\lambda_n(X_n^i)$ and render the top M arms active, the remaining arms are passive. Ties are broken according to some pre-specified convention. With probability ϵ , we render active M random arms, chosen uniformly and independently, the rest passive. We used following stepsize sequences, which gave good performance in practice:

$$\begin{aligned} \alpha(n) &= \frac{C}{\lceil \frac{n}{500} \rceil}, \\ \gamma(n) &= \frac{C'}{1 + \lceil \frac{n \log n}{500} \rceil} I\{n \pmod N \equiv 0\}. \end{aligned} \quad (14)$$

Define $F^\lambda = [F_{iu}^\lambda(\cdot)]_{i \in S, u \in \{0, 1\}} : \mathcal{R}^{2d} \mapsto \mathcal{R}^{2d}$ as follows:

$$\begin{aligned} F_{iu}^\lambda([[\Psi(j, b)]]) &:= (1-u)(R_0(i) + \lambda) + uR_1(i) \\ &\quad + \sum_j p(j|i, u) \max_{v \in \{0, 1\}} \Psi(j, v) - f(\Psi). \end{aligned}$$

Also define $\{M(n) := [M_n(i, u)]\}$ by

$$\begin{aligned} M_{n+1}(i, u) &:= (1-u)(R_0(i) + \lambda_n(x)) + uR_1(i) + \\ &\quad \max_{v \in \{0, 1\}} Q_n(X_{n+1}, v) - f(Q_n) - F_{ia}^{\lambda_n(x)}(Q_n). \end{aligned}$$

Then $\{M_n\}$ are martingale difference sequences w.r.t. $\{\mathcal{F}_n\}$, i.e., they are adapted to $\{\mathcal{F}_n\}$ and satisfy $E[M_{n+1}(i, u) | \mathcal{F}_n] = 0 \forall i, u, n$. Rewrite (12) as

$$\begin{aligned} Q_{n+1}(i, u) &= Q_n(i, u) + \alpha(\nu(i, u, n))I\{X_n = i, Z_n = u\} \\ &\quad \times (F_{iu}^{\lambda_n(x)}(Q_n) - Q_n + M_{iu}(n+1)). \end{aligned} \quad (15)$$

In view of the easily verified fact $\gamma(n) = o(\alpha(n))$, the coupled iterates (15), (13) form a two time scale stochastic approximation algorithm in the sense of [14], section 6.1, with (15) operating on the faster time scale and (13) on the slower time scale. We exploit this fact later for the convergence analysis.

¹This is not the unique choice of $f(\cdot)$, see [3].

IV. NUMERICAL EXAMPLES

Let us illustrate the proposed scheme with two examples.

A. Example with circulant dynamics

We first test our scheme on the example from [27]. The example has four states and the dynamics is circulant: when an arm is passive ($u = 0$), resp. active ($u = 1$), the state evolves according to the transition probability matrix

$$P_0 = \begin{bmatrix} 1/2 & 0 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}, \quad \text{and} \quad P_1 = P_0^T,$$

respectively. The rewards do not depend on the action and are given by $R_0(1) = R_1(1) = -1$, $R_0(2) = R_1(2) = 0$, $R_0(3) = R_1(3) = 0$, and $R_0(4) = R_1(4) = 1$. Intuitively, there is a preference to activate an arm when the arm is in state 3. Indeed, the exact values of the Whittle indices, calculated in [27], are as follows: $\lambda(1) = -1/2$, $\lambda(2) = 1/2$, $\lambda(3) = 1$, and $\lambda(4) = -1$, which give priority to state 3. Consider a scenario with $N = 100$ arms, out of which $M = 20$ are active at each time. We initialize our algorithm with $\lambda_0(x) = 0$, and $Q^x(i, u) = R_u(i)$, $\forall x \in S$.

In this example, we assumed the shared memory architecture and took full advantage of the fact that the arms are identical. This helps to collect the statistics very quickly and results in a rapid convergence of the algorithm. We first set the exploration parameter as $\epsilon = 0.1$.

In Figure 1 we present the convergence of the estimated values of the Whittle indices (see (13)) to the exact values. In Figure 2, we present the comparison of the running time averaged reward obtained by our algorithm with that of the algorithm based on the use of the exact Whittle indices from the beginning. We see that the average rewards stabilize in both approaches already after 250 iterations. The 10% loss of efficiency of our scheme with respect to the approach using the exact Whittle indices is due to the fact that we spend 10% of effort on pure exploration. This actually can be mitigated by decreasing the exploration parameter with time. We notice that as predicted by the theory and confirmed by Figure 1 the estimated Whittle indices in our algorithm converge to the true values. This is in contrast to the scheme proposed in [27] where the convergence appears to be to some random variable.

If we set the exploration parameter as $\epsilon = 0.01$, there is hardly any loss of efficiency of our scheme with respect to the scheme using the exact Whittle indices (see Figure 3). Remarkably, the convergence of the running time averaged reward does not seem to suffer. Of course, the convergence of the estimated Whittle indices to the exact values is now slower. However, since the Whittle indices form a discrete set with generous spacing, what matters is actually the ordinal ranking produced by the estimated Whittle indices, which is quite robust, and not their proximity to the exact values.

This controlled chain is not unichain, as under some stationary policies, it splits into two communicating classes. However, any state is reachable from any other under some control, as a result the optimal cost does not depend on the initial state.

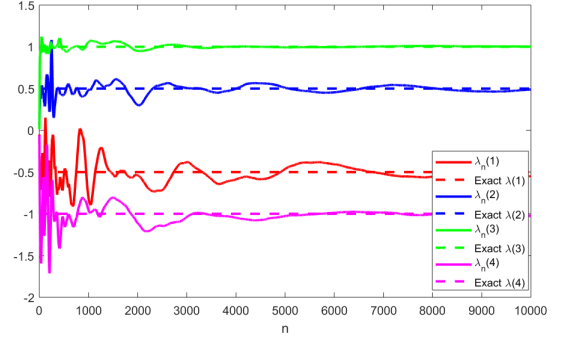


Fig. 1. Estimated (solid lines) and exact (dash lines) Whittle indices in the example with circulant dynamics.

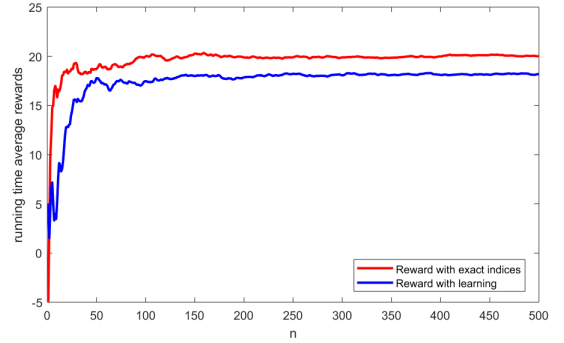


Fig. 2. Rewards comparison in the circulant dynamics ($\epsilon = 0.1$).

B. Example with restart

Now we consider an example where the active action forces an arm to restart from some state. Specifically, we consider an example with 5 states, where in the passive mode ($u = 0$) an arm has tendency to go up the state space, i.e.,

$$P_0 = \begin{bmatrix} 1/10 & 9/10 & 0 & 0 & 0 \\ 1/10 & 0 & 9/10 & 0 & 0 \\ 1/10 & 0 & 0 & 9/10 & 0 \\ 1/10 & 0 & 0 & 0 & 9/10 \\ 1/10 & 0 & 0 & 0 & 9/10 \end{bmatrix},$$

whereas in the active mode ($u = 1$) the arm restarts from state 1 with probability 1, i.e.,

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The rewards in the passive mode are given by $R_0(k) = a^k$ (in our numerical experiments, we have taken $a = 0.9$) and the rewards in the active mode are all zero.

At least three facts have motivated us to choose this example. Bandits with restarting dynamics have several applications such as congestion control [7], [8], web crawling [5], [6], [30] and machine maintenance [19]. Their Whittle indices can be

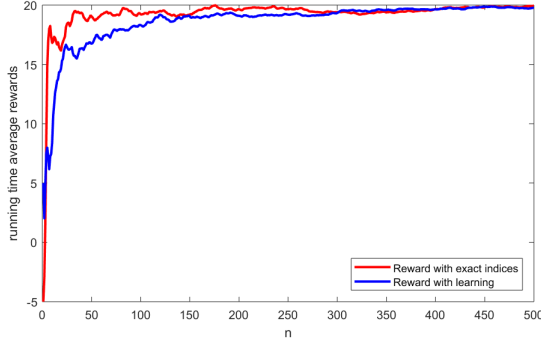


Fig. 3. Rewards comparison in the circulant dynamics ($\epsilon = 0.01$).

easily calculated, see e.g., [22], [25]. The upper states are much less visited, if at all, which poses a challenge for learning.

As in the previous example, we consider the scenario with $N = 100$ arms out of which $M = 20$ are active at each time step. The exact Whittle indices are given by: $\lambda(1) = -0.9$, $\lambda(2) = -0.73$, $\lambda(3) = -0.5$, $\lambda(4) = -0.26$, and $\lambda(5) = -0.01$. We initialize the algorithm with $\lambda_0(x) = 0$, and $Q^x(i, u) = R_u(i)$, $\forall x \in S$.

In Figure 4 we plot the evolution of the estimated Whittle indices with $\epsilon = 0.1$. As expected in this example, the non-homogeneous structure of the state space poses some problems for learning in comparison with the more symmetric example with circulant dynamics. It takes noticeably longer time to learn the Whittle indices for the upper states 4 and 5 in comparison with the lower states 1, 2 and 3.

So far, we have applied decreasing stepsizes recommended in (14). In practice one could also apply constant stepsizes. For instance, in Figure 5 we used constant stepsizes $\alpha = 0.02$, $\gamma = 0.005$. The results are fairly good for all the states except the top state 5. However, the top is visited rarely and thus the value of its Whittle index is not really relevant for good control of the system. One clear practical advantage of the constant stepsize is the possibility of using this variant for tracking purposes when the environment is changing slowly.

V. CONVERGENCE ANALYSIS

In addition to **(C0)**, we make the following assumptions:

- **(C1)** The stepsizes $\{\alpha(n)\}$ satisfy: for $x \in (0, 1)$,

$$\sup_n \frac{\alpha(\lfloor xn \rfloor)}{\alpha(n)} < \infty,$$

$$\sup_{y \in [x, 1]} \left| \frac{\sum_{m=0}^{\lfloor yn \rfloor} \alpha(m)}{\sum_{m=0}^n \alpha(m)} - 1 \right| \rightarrow 0.$$

These are satisfied, e.g., by $\alpha(n) = \frac{1}{n}$ or $\frac{1}{n \log n}$ from some n on.

- **(C2)** The problem is Whittle indexable.

We prove convergence of the above scheme to the desired limit using a combination of results from the theory of

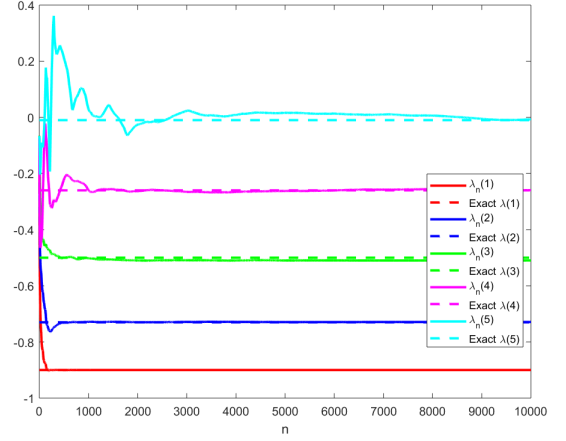


Fig. 4. Estimated (solid lines) and exact (dash lines) Whittle indices in the example with restart.

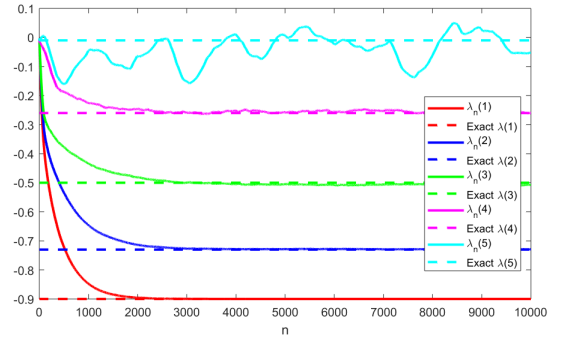


Fig. 5. Estimated (solid lines) and exact (dash lines) Whittle indices in the example with restart. Constant step sizes: $\alpha = 0.02$, $\gamma = 0.005$.

stochastic approximation, in conjunction with [3]. We sketch the key steps, omitting details which are identical to the sources being cited. We call the iteration (12) *synchronous* if all components of Q_n^x are updated at the same time, i.e., the indicator $I\{X_n = i, Z_n = u\}$ in (12) is dropped. Then $\nu(i, u, n) = n \forall i, u$. Also, for updating the i th component, X_{n+1} is replaced by $X_{n+1}^{i,u}$, a simulated random variable independent of all else with law $p(\cdot|i, u)$. The iterate becomes

$$Q_{n+1}^x(i, u) = Q_n^x(i, u) + \alpha(n) \left((1-u)(R_0(i) + \lambda_n(x)) \right. \\ \left. + uR_1(i) + \max_{v \in \{0,1\}} Q_n^x(X_{n+1}^{i,u}, v) - f(Q_n^x) - Q_n^x(i, v) \right).$$

This can be legitimate only for off-line and therefore off-policy learning. It does not cover learning based on the real or simulated run of a single controlled Markov chain, which will be performed asynchronously. We take this up later. In any case, it provides a step towards analyzing the fully *asynchronous* update (12) based on a single run $\{(X_n, Z_n)\}$, which updates only the (X_n, Z_n) th component at time n . One can consider more general forms of asynchrony where

some but not all, and not necessarily only one, components are updated at each time. The analysis will be similar.

Theorem Under the hypotheses (C0), (C1) and (C2), for either synchronous or asynchronous updates of (12)-(13),

$$\lambda_n(x) \rightarrow \lambda(x), \quad \text{a.s.}$$

Proof We split this into many steps.

- 1) *Convergence in synchronous case for a.s. bounded iterates:* Define $h(Q, \lambda) = [[h_{ia}(Q, \lambda)]]$ by:

$$h_{ia}(Q, \lambda) := F_{ia}^\lambda(Q) - Q,$$

and, for the prescribed x as above,

$$g(Q, \lambda) := Q(x, 1) - Q(x, 0).$$

Suppose that the iterates (12)-(13) remain bounded a.s. (We prove this later.) For sake of simplicity, consider the synchronous version. The two timescale argument works as follows². Rewrite (13) as

$$\lambda_{n+1}(x) = \lambda_n(x) + \alpha(n) \left(\frac{\gamma(n)}{\alpha(n)} \right) (Q_n(x, 1) - Q_n(x, 0)). \quad (16)$$

Let $\tau(n) := \sum_{m=0}^n \alpha(m)$, $m \geq 0$. Define $\bar{Q}(t), \bar{\lambda}(t)$, $t \geq 0$, by $\bar{Q}(\tau(n)) := Q_n^x$, $\bar{\lambda}(\tau(n)) = \lambda_n(x)$ with linear interpolation on each interval $[\tau(n), \tau(n+1)]$, $n \geq 0$. Then $\bar{Q}(\cdot), \bar{\lambda}(\cdot)$ track the asymptotic behavior of the coupled o.d.e.s

$$\dot{Q}(t) = h(Q(t), \lambda(t)), \quad \dot{\lambda}(t) = 0,$$

where the latter is a consequence of $\frac{\gamma(n)}{\alpha(n)} \rightarrow 0$ in (16). These o.d.e.s have Lipschitz functions on the right hand side (see, e.g., the discussion on p. 687, [3]) and therefore are well-posed. Thus $\lambda(\cdot) \equiv$ a constant (say) λ' . The first o.d.e. then reduces to $\dot{Q}(t) = h(Q(t), \lambda')$, which has a globally asymptotically stable equilibrium $Q_{\lambda'}^*$ (Theorem 3.4, p. 689, [3]) satisfying $f(Q_{\lambda'}^*) = \beta_{\lambda'}$, $\beta_{\lambda'} :=$ the optimal cost for $\lambda_n \equiv \lambda'$. What this translates into for the original iterates is that $Q_n^x - Q_{\lambda_n}^* \rightarrow 0$ a.s. That is, (12) views (13) as quasi-static and tracks $Q_{\lambda_n}^*$ as λ_n evolves on a slower time scale. This in turn can be used to argue [14] that the interpolated trajectory

$$\tilde{\lambda}(t) = \lambda_n + \left(\frac{t - \tau'(n)}{\tau'(n+1) - \tau'(n)} \right) (\lambda_{n+1} - \lambda_n)$$

for $t \in [\tau'(n), \tau'(n+1)]$, $\tau'(n) := \sum_{m=0}^n \gamma(m)$, $n \geq 0$, tracks the o.d.e.

$$\dot{\Lambda}(t) = Q_{\Lambda(t)}^*(x, 1) - Q_{\Lambda(t)}^*(x, 0). \quad (17)$$

If $\Lambda(t) > \lambda(x)$ (excess subsidy), the passive mode is preferred, i.e., $Q_{\Lambda(t)}^*(x, 0) > Q_{\Lambda(t)}^*(x, 1)$. Then the r.h.s. is < 0 and $\Lambda(t)$ decreases. Likewise,

if the opposite (strict) inequality holds, the r.h.s. is > 0 and $\Lambda(t)$ increases. Thus the trajectory $\Lambda(\cdot)$ remains bounded. Since any well-posed scalar o.d.e. with bounded trajectories must converge to an equilibrium, $\Lambda(t)$ converges to the Λ satisfying $\hat{Q}_{\Lambda}^x(x, 1) = \hat{Q}_{\Lambda}^x(x, 0)$, i.e., the Whittle index $\lambda(x)$. This is unique by hypothesis. By theory of two time scale stochastic approximation (section 6.1, [14]), we have $(Q_n^x, \lambda_n(x)) \rightarrow (Q_{\lambda(x)}^*, \lambda(x))$ a.s.

- 2) *The a.s. boundedness of iterates, general case:* Here we use the results of [24]. For this, we verify the assumptions (A1)-(A5) of [24].

- (A1) of [24] requires that h, g are Lipschitz. We have already noted that this is so.
- In the notation of [24], $M_n^{(1)}, M_n^{(2)}$ correspond to resp. M_n and the process that is identically zero. Both of these are martingale difference sequences (the latter trivially so). Furthermore, $\forall n$,

$$E [\|M_{n+1}\|^2 | \mathcal{F}_n] \leq K (1 + \|Q_n\|^2 + \|\lambda_n\|^2)$$

a.s. by the Lipschitz property of the functions involved. The zero process trivially satisfies such an inequality. The two preceding statements are precisely (A2) of [24].

- (A3) of [24] requires that $\sum_n \alpha(n) = \sum_n \gamma(n) = \infty$, $\sum_m (\alpha(n)^2 + \gamma(n)^2) < \infty$, and $\gamma(n) = o(\alpha(n))$, which hold here.
- For (A4), we first consider the ‘synchronous’ case when all components of Q_n are updated simultaneously. The limiting o.d.e. with λ_n frozen at λ is (cf. equation (3.4) in [3]) $\dot{Q}(t) = h(Q(t), \lambda)$. This has as its globally asymptotically stable equilibrium the solution $Q_\lambda^* = [[Q_\lambda^*(i, a)]]$ of (9) with $f(Q_\lambda^*) = \beta_\lambda := \beta$ with its λ -dependence made explicit (Theorem 3.4 of [3]). The limit

$$h_\infty(Q, \lambda) := \lim_{c \uparrow \infty} \frac{h(cQ, c\lambda)}{c}$$

then corresponds to the Q-learning problem for average reward control with constant running reward $\equiv \lambda$ for passive states and zero reward for active states. Recall from [3] that this Q-learning scheme converges to the unique \hat{Q}_λ^* for which $f(\hat{Q}_\lambda^*) = \hat{\beta}_\lambda := \lambda \times$ the stationary probability of the set of passive states under the optimal policy. Thus, for $\lambda = 0$, both the active and passive running rewards and hence $\hat{\beta}_\lambda$ are zero. So the unique solution to (9) with $f(\hat{Q}_\lambda) = \hat{\beta}_\lambda$ is the zero vector. This establishes (A4) of [24].

- Consider the limit

$$g_\infty(\lambda) = \lim_{c \uparrow \infty} \frac{g(\hat{Q}_{c\lambda}^*, c\lambda)}{c}.$$

Letting

$$r_c(i, u) := \left(uR_1(i) + (1-u)(c\lambda + R_0(i)) \right) / c$$

²We only sketch the main steps, see [14], Chapter 6, for details.

denote the scaled running reward and $\beta^c := \beta/c$ the scaled optimal reward, both are seen to be uniformly bounded for $c \geq 1$. Divide both sides of equation (9) by c and let $c \uparrow \infty$. For each $c \geq 1$, it becomes the counterpart of (9) for running reward r_c that remains uniformly bounded over $c \in [1, \infty)$. Let SP be the set of stationary policies $v : S \mapsto \{0, 1\}$, V_c be the value function for the average reward problem with running reward r_c , and τ be the first hitting time of a fixed state $i_0 \in S$ accessible from every other state as per (C0). For now, we write Q -values as $Q^c(\cdot, \cdot)$ to show the c -dependence explicitly. Using a standard representation for the value function ([13], p. 79),

$$\begin{aligned} & Q^c(i, u)/c \\ &= r_c(i, u) - \beta^c + \sum_j p(j|i, u) V_c(j) \\ &= r_c(i, u) - \beta^c + \sum_j p(j|i, u) \times \\ & \quad \max_{v \in SP} E \left[\sum_{m=0}^{\tau} (r_c(X_m, v(X_m)) \right. \\ & \quad \left. - \beta^c) | X_0 = j \right] \\ &\leq C \left(1 + \max_{v \in SP, j \in S} E[\tau | X_0 = j] \right) < \infty \end{aligned}$$

by (1), for a suitable constant C . Thus $Q^c(\cdot, \cdot)/c, \beta^c$ remain bounded as $c \uparrow \infty$. Any limit point $(Q_\lambda^\infty(\cdot, \cdot), \beta_\lambda^\infty)$ thereof (with the λ -dependence rendered explicit again) satisfies

$$Q_\lambda^\infty(i, 0) = \lambda - \beta_\lambda^\infty + \sum_j p(j|i, 0) \times \max_a Q_\lambda^\infty(j, a), \quad (18)$$

$$Q_\lambda^\infty(i, 1) = -\beta_\lambda^\infty + \sum_j p(j|i, 1) \times \max_a Q_\lambda^\infty(j, a). \quad (19)$$

For $\lambda > 0$ as $c \uparrow \infty$, eventually $u = 0$ is optimal for all states. Then $\beta_\lambda^\infty = \lambda$ and (18) leads to

$$Q_\lambda^\infty(i, 0) = \sum_j p(j|i, 0) Q_\lambda^\infty(j, 0) \quad \forall i,$$

implying $Q_\lambda^\infty(i, 0) \equiv \text{a constant}$. Then from (19), we have $Q_\lambda^\infty(i, 1) - Q_\lambda^\infty(i, 0) = -\beta_\lambda^\infty = -\lambda$. For $\lambda < 0$ as $c \uparrow \infty$, eventually $u = 1$ is optimal for all states. Equation (19) then implies that $\beta_\lambda^\infty = 0$, otherwise the equation does not have a solution: Iterating (19) leads to $Q_\lambda^\infty(i, 1) = -n\beta_\lambda^\infty + \text{a bounded quantity}$. This becomes unbounded unless $\beta_\lambda^\infty = 0$. In turn, (19) with $\beta^\infty = 0$ leads to $Q^\infty(i, 1) \equiv \text{a constant independent of } i$. From

(18), we then have

$$Q_\lambda^\infty(i, 1) - Q_\lambda^\infty(i, 0) = -\lambda + \beta_\lambda^\infty = -\lambda.$$

For $\lambda = 0$, $\beta_\lambda^\infty = 0$ and the unique up to additive scalar solution to (18), (19) is again the zero vector. This leads to

$$Q_\lambda^\infty(i, 1) - Q_\lambda^\infty(i, 0) = 0 = -\lambda.$$

We have proved that $g_\infty(\lambda) = -\lambda \quad \forall \lambda$. The limiting o.d.e. $\dot{\lambda}(t) = g_\infty(\lambda(t)) = -\lambda(t)$ has zero as its unique globally asymptotically stable equilibrium. This verifies (A5) of [24].

The results of [24] then imply a.s. boundedness of the iterates, i.e.,

$$\sup_n |\lambda_n(x)| < \infty, \quad \sup_n |Q_n(i, u)| < \infty \quad \forall i, u, \quad \text{a.s.}$$

3) Asynchronous case:

For the asynchronous case, further tweaks are needed for the above analysis and the analysis of [24] to become applicable, because one of the iterations, viz., (15), is asynchronous. The a.s. stability of iterates can be established as in [12], combining it with [24] to include the two timescale effect. Note that in the present case, only the faster iteration (15) is asynchronous, so only the analysis of section 5.1 of [24], which deals with the fast timescale, that has to be replaced by the analysis for the asynchronous case in [12]. Thereafter asynchrony only changes the intermediate analysis of the o.d.e. without changing the conclusion as shown in [12], as long as

$$\liminf_{n \uparrow \infty} \frac{\nu(k, u, n)}{n} > 0 \quad \forall k, u, \quad \text{a.s.}$$

But this is ensured by our choice of $\{Z_n\}$ which picks every possible action with probability at least $\epsilon > 0$. We omit the details. It is worth highlighting that assumption (C1) plays a crucial role here, in particular it implies that $\forall i, j, u, v$,

$$\lim_{n \uparrow \infty} \frac{\sum_{m=0}^n \alpha(\nu(i, u, n)) I\{X_m = i, Z_m = u\}}{\sum_{m=0}^n \alpha(\nu(j, v, n)) I\{X_m = j, Z_m = v\}} = 1.$$

Intuitively, this means that the algorithmic time scales across different components are ‘balanced’, which ensures that the asynchronous updates asymptotically track a time-scaled version of the same limiting o.d.e. as the synchronous updates ([14], Chapter 7). Since time scaling does not alter the trajectories and affects only the speed with which they are traversed, the asymptotic behavior is identical for both. The results of section 3 of [3] then imply that for $\lambda_n \equiv \text{a constant } \lambda$, the iterates (15) converge to Q_λ^* a.s. in both synchronous and asynchronous cases. The rest of the argument is identical to that for the synchronous case. \square

VI. CONCLUSIONS

We have presented a novel Q-learning algorithm for Whittle indexable restless bandits and justified it analytically and through numerical experiments. The general philosophy extends easily to related problems such as discounted rewards and related algorithms such as SARSA. An interesting future direction is to explore function approximation for Q-values in order to handle large state spaces.

REFERENCES

- [1] Aalto, S., Lassila, P., and Taboada, I. "Whittle index approach to opportunistic scheduling with partial channel information," *Performance Evaluation*, 136, p. 102052, 2019.
- [2] Altman, E. *Constrained Markov Decision Processes*, CRC Press, Boca Raton, FL, 1999.
- [3] Abounadi, J., Bertsekas, D. P. and Borkar, V. S. "Learning algorithms for Markov decision processes with average cost", *SIAM Journal on Control and Optimization*, 40(3), pp. 681-698, 2001.
- [4] Archibald, T. W.; Black, D. P. and Glazebrook, K. D. "Indexability and index heuristics for a simple class of inventory routing problems," *Operations Research*, 57.2, pp. 314-326, 2009.
- [5] Avrachenkov, K. and Borkar, V. S., "Whittle index policy for crawling ephemeral content," *IEEE Trans. on Control of Network Systems*, 5.1, 2018, pp. 446-455.
- [6] Avrachenkov, K. and Borkar, V. S. "A learning algorithm for the Whittle index policy for scheduling web crawlers", *Proc. 57th Allerton Conference on Communication, Control, and Computing*, Monticello, IL, 2019.
- [7] Avrachenkov, K., Ayesta, U., Doncel, J., and Jacko, P. "Congestion control of TCP flows in Internet routers by means of index policy," *Computer Networks*, 57(17), pp. 3463-3478, 2013.
- [8] Avrachenkov, K., Borkar, V. S. and Pattathil, S. "Controlling G-AIMD using index policy," *56th IEEE Conference on Decision and Control*, Melbourne, Dec., 2017.
- [9] Avrachenkov, K., Piunovskiy, A., and Zhang, Y. "Impulsive Control for G-AIMD Dynamics with Relaxed and Hard Constraints". *57th IEEE Conference on Decision and Control*, pp.880-887, 2018.
- [10] Bertsekas, D. P. *Reinforcement Learning and Optimal Control*, Athena Scientific, Belmont, Mass., 2019.
- [11] Bertsekas, D. P. and Tsitsiklis, J. N. *Neurodynamic Programming*, Athena Scientific, Belmont, Mass., 1996.
- [12] Bhatnagar, S. "The Borkar-Meyn theorem for asynchronous stochastic approximations", *Systems and Control Letters*, 60(7), pp. 472-478, 2011.
- [13] Borkar, V. S. *Topics in Controlled Markov Chains*, Longman Scientific and Technical, Harlow, UK, 1991.
- [14] Borkar, V. S. *Stochastic Approximation: A Dynamical Systems Viewpoint*, Hindustan Publishing Agency, New Delhi, and Cambridge University Press, Cambridge, UK, 2008.
- [15] Borkar, V. S. and Chadha, K. "A reinforcement learning algorithm for restless bandits", in *Proc. 5th Indian Control Conference (ICC)*, IIT Kanpur, pp. 89-94, 2018.
- [16] Borkar, V. S., and Pattathil, S. "Whittle indexability in egalitarian processor sharing systems," *Annals of Operations Research* (2017) (available online).
- [17] Borkar, V. S. Ravikumar, K. and Saboo, K. "An index policy for dynamic pricing in cloud computing under price commitments," *Appl. Math. Comput.*, 44.2, pp. 215-245, 2017.
- [18] Gittins, J.; Glazebrook, K. and Weber, R. *Multi-armed bandit allocation indices*, John Wiley & Sons, 2011.
- [19] Glazebrook, K. D., Mitchell, H. M. and Ansell, P. S. "Index policies for the maintenance of a collection of machines by a set of repairmen," *Europ. J. Oper. Res.*, 165, pp. 267-284, 2005.
- [20] Glazebrook, K. D., Kirkbride, C. and Ouenniche, J. "Index policies for the admission control and routing of impatient customers to homogeneous service stations", *Operations Research*, 57.4, pp. 975-989, 2009.
- [21] Hernández-Lerma, O. *Adaptive Markov Control Processes*, Springer Verlag, New York, 1989.
- [22] Jacko, P. *Dynamic Priority Allocation in Restless Bandit Models*, Lambert Academic Publishing, 2010.
- [23] Kumar, P. R. and Varaiya, P. *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice-Hall, 1986.
- [24] Lakshminarayanan, C. and Bhatnagar, S. "A stability criterion for two timescale stochastic approximation schemes", *Automatica*, 79, 2017, pp. 108-114.
- [25] Larrañaga, M., Ayesta, U., and Verloop, I. M. "Asymptotically optimal index policies for an abandonment queue with convex holding cost," *Queueing Systems*, 81(2-3), pp. 99-169, 2015.
- [26] Liu, K. and Zhao, Q. "Indexability of restless bandit problems and optimality of Whittle Index for dynamic multichannel access", *IEEE Transactions on Information Theory*, 56.11, pp. 5547-5567, 2010.
- [27] Fu, J.; Nazarathy, Y.; Moka, S. and Taylor, P. G. "Towards Q-learning the Whittle Index for Restless Bandits", *preprint available at https://people.smp.uq.edu.au/YoniNazarathy/pub/20190502.pdf*
- [28] Niño-Mora, J. and Villar S. S. "Sensor scheduling for hunting elusive hiding targets via Whittle's restless bandit index policy", *5th International Conference on Network Games, Control and Optimization (NetGCoP)*, 2011.
- [29] Niño-Mora, J. , "Admission and routing of soft real-time jobs to multi-clusters: Design and comparison of index policies." *Computers & Operations Research*, 39.12, pp. 3431-3444, 2012.
- [30] Niño-Mora, J. , "A dynamic page-refresh index policy for web crawlers." In *Proceedings of Analytical and Stochastic Modeling Techniques and Applications*, Springer, pp. 4660, 2014.
- [31] Raghunathan, V.; Borkar V. S.; Cao, M. and Kumar, P. R. "Index policies for real-time multicast scheduling for wireless broadcast systems", *Proceedings of INFOCOM 2008, 27th IEEE Conference on Computer Communications*, 2008.
- [32] Ruiz-Hernandez, D. *Indexable restless bandits*. VDM Verlag, 2008.
- [33] Papadimitriou, C. H. and Tsitsiklis, J. N. "The complexity of optimal queueing network control", *Mathematics of Operations Research* 24.2, pp. 293-305, 1999.
- [34] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley, New York, 1994.
- [35] Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction* (2nd ed.), MIT Press, Cambridge, Mass., 2018.
- [36] Villar, S. S., Bowden, J., and Wason, J. "Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges", *Statistical Science*, 30(2), pp. 199-215, 2015.
- [37] Watkins, C. I. C. J. *Learning from delayed rewards*, Ph.D. Thesis, Cambridge University, Cambridge, UK, 1988.
- [38] Weber, R. R., and Weiss, G. "On an index policy for restless bandits", *Journal of Applied Probability*, 27.3, pp. 637-648, 1990.
- [39] Whittle, P. "Restless bandits: activity allocation in a changing world", *Journal of Applied Probability*, 25, pp. 287-298, 1988.