# Scheduling users in drive-thru Internet: a multi-armed bandit approach

Thi Thuy Nga NGUYEN[1,3], Urtzi AYESTA[2], and Balakrishna PRABHU[3]

[1]Continental Digital Service in France, Toulouse, France
[2]CNRS-IRIT, Univ. Toulouse, Toulouse, France
[3]LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France

***Abstract*—We consider the problem of allocating a wireless channel to mobile users moving on a straight road. The objective is to maximize a given function of the total data transmitted. We develop a model within the multi-armed bandit framework and formulate an optimization problem under the constraint that at most one user can be served at a time. We solve the relaxed optimization problem, in which one user is served on the average, and show it to be indexable. A simple and easy-to-compute expression is given for the Whittle index. We then propose a heuristic policy for the original optimization problem using Whittle's index policy. The proposed heuristic is shown to perform well compared to some other heuristics in various settings including dynamic scenarios with arrivals of new users and the presence of heterogeneous users.**

***Index Terms*—Markov Decision Process, restless multi-armed bandit problem, Whittle's index, scheduling, drive-thru internet**

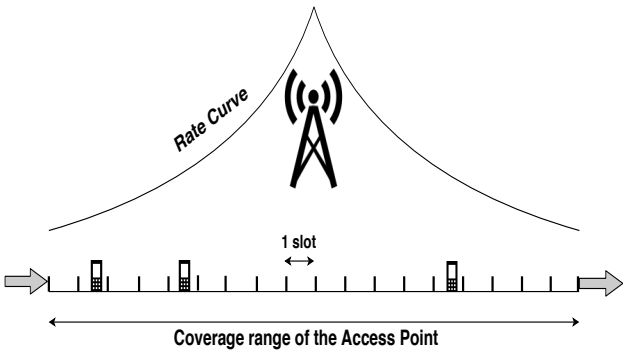


Fig. 1: A drive-thru internet network.

## I. Introduction

Drive-thru internet has seen a recent resurgence due to an increase in demand for high-speed internet access from mobile users [1], [2]. A typical scenario for drive-thru internet is a WiFi hotspot or access point (AP) that servers users moving along a straight line as shown in Figure 1. For example, these users can be cars or pedestrians moving on (or along) a long avenue. Various question related to link-layer scheduling and resource allocation [3], [4], MAC layer retransmissions [5], message scheduling using network coding [6] have recently been investigated by taking into account the specific mobility pattern of the drive-thru internet systems.

In this paper, we revisit the multi-class scheduling problem for Markovian queues [7], [8], [9] in the context of a drive-thru internet. Consider users of different classes (i.e, different mean service requirements) moving along a straight line in the coverage area of an AP (Fig. 1). Users enter the coverage range from the left and leave from the right. In each time-slot, the AP has to determine which user to serve in order to maximize a given long-term objective. The AP can serve at most one user in each time-slot. Users receive a rate depending upon their distance from the AP: users who are closer have a higher rate (as shown in Fig. 1). The trade-off is between serving users with a higher rate and users who leave first.

### A. Contributions

We model the problem as a Markov Decision Process whose solution can be computed numerically for small number of users but becomes computationally intractable for large instances. We shall rely on the multi-armed bandit approach of Whittle [10], [11] to obtain a heuristic based on the Whittle index. In order to compute the index, one needs the technical condition called *indexability* which cannot always be proven. Even when one can prove indexability, the indices are not easy to compute ([12], [13] and references therein). Using the special mobility pattern (users move one spatial-slot to the right in each time-slot), we show the indexability of a simplified model with no arrivals and give a simple computational procedure for these indices. The indices are obtained as a function of the position and the class of the user. The heuristic then serves the user with the highest index. It will be shown that this index is not the same as the greedy algorithm (or the $c\mu$-rule [14]) that assigns the channel to the user with the largest product of channel rate and mean service requirement. Several numerical experiments will be presented to show the performance improvements of the proposed heuristic with respect to the greedy policy. In particular, the improvements are seen to be more pronounced when there are more classes of users.

### B. Related work

In [6] the authors develop an information-theoretic formula for the total amount of information that a vehicle can receive (for only one user) when it passes the broadcast zone of a BS. Vehicles moving in a road may have blind zones in which they might not receive signal from the base station. To circumvent this problem, they propose and analyse the benefits of cooperative scheme for joint V2V and V2I communications in order to improve the system capacity.

In [15] the authors investigate a utility maximization problem for video transmissions for multiple vehicles and one AP. The setting is similar to ours. They formulate the problem as integer programming problem when the future arrivals are known in order to obtain a benchmark. Since the integer programming problem is intractable for large instances, they propose a heuristic based on the utility potential. For each user this potential is

computed as the maximum amount of utility that a user can get if it were served continuously for a certain number of slots. The user with this highest utility potential is served. This algorithm is similar to the Gittins' index policy which does not take into account the change in state of users.

Index policies have long been known in scheduling theory. In general, the solution to a scheduling problem will be a complex function of all the input parameters and the number of competing jobs. In practice such problems can be solved only for very specific instances. Remarkably, in some cases, a so-called *index policy* is optimal. For example, the optimality of $c\mu$ and SRPT [16] can be cast in the framework of Multi-Armed Bandit Problems (MABP), a broad class of resource allocation problems for which index policies are known to be optimal. A MABP is a particular case of a Markov Decision Process: at every decision epoch the scheduler needs to select one *bandit*, and an associated reward is accrued. The state of this selected bandit evolves stochastically, while the state of all other bandits remains *frozen* or *rested*. The scheduler knows the state of all bandits and aims at maximizing the total average reward. In a ground-breaking result Gittins showed that the optimal policy that solves a MABP is an index rule, nowadays commonly referred to as Gittins' index policy [11]. Thus, for each bandit, one calculates Gittins' index, which depends only on its own current state and stochastic evolution. The optimal policy activates in each decision epoch the bandit with highest current index.

Despite its generality, in multiple cases of practical interest the problem cannot be cast as a MABP. For example, mobility of users directly invalidates the requirement that non-selected bandits remain *frozen*. In a seminal work [10], Whittle introduced the so-called Restless Bandit Problem (RBP), a generalization of the standard MABP in which all bandits might evolve over time according to a stochastic kernel that depends on whether the bandit is made active. RBP provides a powerful modeling framework, but its solution has in general a complex structure that might depend on the entire state-space description. In fact, it is known that RBP are PSPACE-hard even in its deterministic variant [17], and typically suffer from the curse of dimensionality.

Whittle considered a relaxed version of the problem (where the restriction on the number of *active* bandits needs to be respected on average only, and not in every decision epoch), and showed that the solution to the relaxed problem is of index type, referred to as *Whittle's index*. Whittle then defined a heuristic for the original problem, referred to as Whittle's index policy, where in every decision epoch the bandit with highest Whittle index is selected. It has been shown that the Whittle index policy performs strikingly well, see [18] for a discussion, and is asymptotically optimal under certain conditions, see [7], [19].

As mentioned earlier, Gittins' index is optimal for the 'rested' bandit problem, i.e, in which the bandits do not change state if they are not served. Moving users, on the other hand, are 'restless', since users change state (or position) even when they are not served. Thus, a better approach for drive-thru internet is the Whittle's relaxation based method for restless bandits which we shall follow in this paper.

### C. Organization

The rest of the paper is organized as follows. Section II formally describes the general setting and casts the problem as an MDP. It also proposes the simpler model of no arrivals. Section III states the main result on the indexability of the model of no arrivals and gives an easy to compute formula for the indices. The heuristic Whittle-index policy based upon the main result is presented in Section IV. This section also contains numerical comparisons of the proposed policy with other policies. The conclusions and further research directions appear in Section V. Some of the proofs have been moved to the appendix for improved readability, and some other have been omitted due to lack of space.

## II. Problem formulation

Consider an AP with a coverage range of length $L$ (see Fig. 1). The users enter the coverage range from left, move at a constant velocity, and leave from the right. Every $\Delta$ time units the AP has to decide which user to serve. Let $v$ be the velocity of the users. Then, the coverage range can be divided into spatial-slots on length $v\Delta = \sigma$. Let $\mathcal{S} = \{0, 1, 2, \ldots, N\}$, where $N = L/\sigma - 1$, denote the set of spatial-slots with the convention that slot 0 is the leftmost slot. The length of the time-slot is assumed to be much smaller than the coverage range of the AP (in the order of hundreds of meters). This is a reasonable assumption since scheduling decisions are made every 10-20 ms during which a car inside a city would move a distance of less than a meter.

The data rate received by a user in spatial-slot $s$ depends on the distance between the AP and $s$. Users that are closer to the AP will get a higher rate than the users that are closer to the end points. We shall assume that the Signal-to-Noise Ratio (SNR) has a polynomial decay: $SNR(s) = C_1 d(s)^{-\gamma}$ and that the data rate in slot $s$, $C(s)$ can be obtained using Shannon's law:

$$C(s) = C_2 \log(1 + SNR(s)). \tag{1}$$

For more information on these formulae, we refer to [6]. The amount of data that is transmitted in a time-slot, $r(s)$, to a user served at rate $C(s)$ will thus be $C(s)\Delta$.

**Assumption 1.** *The function $r(s)$ is unimodal with maximum at $s = N/2$ (assuming $N$ is even). It is non-decreasing on the left and non-increasing on the right.*

The assumption is quite natural and is satisfied by the rate function derived from (1).

The total volume of data requested by user $i$ is assumed to be an independent and exponentially distributed random variable with rate $\eta_{i,b}$. Here $b$ is the class of user $i$ and $b \in \mathcal{B} := \{1, 2, ...B\}$, where $B$ is the number of classes. Thus, the probability that a user of class-$b$ who is served in slot $s$ finishes its data transfer in that slot is $1 - \exp(-\eta_b r(s))$. The assumption of exponential data volumes ensures that this probability is independent of past allocations.

In each time-slot, users arrive in spatial-slot 0 according to a categorical distribution on $\mathcal{B} \cup \{0\}$. The outcome 0 corresponds to no arrival in that time-slot. The probability that a user of class-$b$ arrives in time-slot will be denoted $p_b$ for $b \in \mathcal{B}$. If $\sum_b p_b < 1$, then there is a non-zero probability of there being no new arrival in a time-slot.

### A. Objective

In each time-slot, the AP can choose *at most* one user (or a spatial-slot) to serve, that is, its set of actions is $\mathcal{A} = \{e_i\}_{i\in\mathcal{S}}$ with $e_i$ being the unit vector for the $i$th coordinate.

For a given policy $\pi$ of the AP, let $S^\pi(t) \in (\{0,1\} \times B)^{\mathcal{S}}$ be the stochastic process that indicates whether a spatial-slot is occupied by a user or not and tells the class of the user if it is occupied. The objective of the AP is:

$$\max_{\pi} \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} R(S^\pi(t), a^\pi(S^\pi(t), t)), \qquad \text{(OBJEC)}$$

where $a^\pi(s,t) \in \mathcal{A}$ is the action prescribed by policy $\pi$ in the state $s$ in time-slot $t$. And $R^\pi(s,a)$ is reward one-step after choosing action $a$ under $\pi$ for state $s$ which is described in this below assumption.

**Assumption 2.** *The reward in a time-slot is sum of the rewards of each user, where the reward of a user is a strictly positive and increasing function of the rate if it is served and* 0 *if it is not served.*

In our problem, reward one-step of a user is its departure probability in that slot, which is a strictly positive and increasing function of that user's rate.

From the assumptions on the data volumes and the arrival process, it can be seen that this problem is a classical average-cost MDP [20]. There exists an optimal stationary (time-independent) policy that can be computed numerically. The drawback of this formulation, however, is that the number of states in any practical scenario is too large to allow numerical computation. As mentioned in the Introduction, for time-slots of 10–20 ms and a coverage length of 100–200 m, the number of spatial-slots, $N$, is of the order of a thousand. The state space of $S(t)$ will have $\approx B2^{|N|} = B2^{1000}$ elements making the problem intractable. Even for 20–30 spatial-slots, the problem is not computationally tractable in reasonable time.

Instead of treating the problem in its full generality, as a first step, we shall focus on a simplified instance of the problem in which there are no arrivals, that is $p_b = 0$, $\forall b \in \mathcal{B}$. This will allow us to obtain certain heuristics that will then be used for the general problem.

### B. Finite horizon MDP for problem with no-arrivals

Let there be $K$ users at time 0, and let $X_k(t) \in \mathcal{N} := \mathcal{S} \cup \{N+1\}$ be the position of user-$k$ in time-slot $t$. The special state $N+1$ indicates that the user has departed the system either because it has moved out of the coverage range or because its demand has been satisfied. We shall assume that the parameter of the exponential distribution for user-$k$ is $\eta_k$. That is, each user could potentially be of a different class.

Since there are no arrivals, the process $S(t)$ can be replaced by the process $\mathbf{X}(t) := (X_1(t), \ldots X_K(t))$. Let $a_k(t) \in \{0,1\}$ denote whether user-$k$ was served in slot $t$ or not.

With these definitions and Assumption 2, it can be seen that the problem (OBJEC) is equivalent to the following problem finite-horizon MDP when there are no arrivals:

$$\begin{cases} \max_{\pi} & \frac{1}{N+1} \sum_{t=0}^{N} \sum_{k=1}^{K} \mathbb{E}^{\pi}_{\mathbf{x}}(R_k(X_k(t), a_k(t)) \\ \text{subject to} & \\ & \sum_{k=1}^{K} a_k(t) \le 1,\ t = 0,1,2,...,N, \\ & a_k(t) \in \{0,1\},\ \forall k, t \end{cases} \qquad \text{(NOARR)}$$

Here $\mathbf{X}(0) = \mathbf{x}$ is the initial position of the users and $R_k(x,a)$ is the reward obtained (i.e., data transferred) by the user-$k$ when action $a$ is taken in state $x$. The constraints on the actions indicate that at most one user can be served in a time-slot. Further, the horizon of the problem can be constrained to $N$ since all users would have left the coverage range by that time.

In general, optimal policies for finite-horizon problem need not be stationary. However, (NOARR) is a particular type of finite-horizon problem known as the stochastic shortest path problem ([21], e.g.) for which, under certain assumptions, there exists a stationary optimal policy that is the solution of Bellman's equation.

**Lemma 1.** *Problem* (NOARR) *admits a stationary optimal policy that satisfies Bellman's equation.*

The proof is based on showing that (NOARR) satisfies the sufficient conditions (e.g., Assumptions 1 and 2 in [21]) for a stochastic shortest path problem to have a stationary optimal policy satisfying Bellman's equation.

This result will be important later on when we shall derive a heuristic based on Whittle's index.

With some abuse of notation, let $r_x = r(x)$. For a user-$k$, given $a_k(t) = a$, $X_k(t)$ has the transition probabilities:

$$\mathbf{P}_k(y|x,a) = \begin{cases} ae^{-r_x\eta_k} + (1-a), & y = x+1, x \neq N+1; \\ a(1-e^{-r_x\eta_k}), & y = N+1, x \neq N+1; \\ 1, & y = N+1, x = N+1; \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

It now follows that *(i)* the dynamics of each user is Markovian and is independent of that of the other users conditioned on the action. Further, each user can change state whether it is served or not. *(ii)* the reward function is decomposable into sum of rewards of the individual users.

Problem (NOARR) is thus an instance of the RBP framework considered by Whittle [10], [11]. The bandits in that framework correspond to users in our problem[1]. We note that (NOARR) is a finite-horizon problem whereas most of the work in the literature on RBPs is on the infinite-horizon setting. The fact that (NOARR) is also a classical stochastic shortest path problem allows us to use results of the infinite-horizon setting for the present problem as well.

## III. Whittle's relaxation and indexability

One of the difficulties in solving (NOARR) comes from the constraints that need to be satisfied in each time-slot. To overcome this, Whittle proposed to relax the constraint that at most one user is served (or active) per time and to replace it by the constraint that at most one user is active in average over time. He then considered the Lagrange relaxation of the problem with relaxed constraints and arrived at $K$ sub-problems—one for each of the $K$ users—thus reducing the dimension of the problem considerably (from $N^K$ to $N$).

Following Whittle's approach, we obtain the following $K$ sub-problems:

$$\begin{array}{ll} \max_{\pi_k} \sum_{t=0}^{N} & \mathbb{E}^{\pi_k}_{x_k}(R_k(X_k(t), a_k(t)) - \nu \sum_{t=0}^{N} \mathbb{E}^{\pi_k}_{x_k}(a_k(t)) \\ \text{subject to} & a_k(t) \in \{0,1\}, \forall t \end{array} \qquad \text{(SUBP-}k\text{)}$$

[1]From now on, we shall use bandits and users interchangeably to mean the same thing. Similarly activating a bandit will mean serving a user.

where $\nu$ is the Lagrange multiplier of the relaxation of the constraint of one active user per slot on an average. Sub-problem-$k$, (SUBP-$k$), can be seen as the problem solved by the AP when user-$k$ is alone in the system and there is a penalty $\nu$ on the actions. Note that each (SUBP-$k$) is again a stochastic shortest path problem which can be solved independently of the other users, and to which Lemma 1 can be applied to argue the existence of a stationary optimal policy.

Intuitively, $\nu$ can be seen as the penalty for being active (or being served) because it reduces the reward for taking $a_k > 0$. If $\nu = -\infty$, then it is optimal to activate all the bandits while if $\nu = +\infty$, then the optimal policy is to inactivate the bandits.

From now on, we concentrate on (SUBP-$k$), and omit index $k$ in the variables to simplify the notation. For a given $\nu$, any stationary policy, $\pi$, can be characterized by its active set $\Omega^\pi(\nu) = \{x : a(x) = 1\}$ which is the set of states in which the bandit is active. Let $\Omega^*(\nu) \subset \mathcal{N}$ to be the active set for the optimal policy of (SUBP-$k$). It can be seen that $\Omega^*(0) = \mathcal{N}$ is the set of all states. This is because there is no penalty for taking $a = 1$ and in each state this action gives at least as much immediate reward as $a = 0$. Similarly, $\Omega^*(\infty) = \emptyset$ since $a = 1$ has too high a penalty.

**Definition 1.** *For $\nu \in [0, \infty)$, a bandit is said to be indexable if $\Omega^*(\nu)$ is monotonically decreasing in $\nu$, that is $\nu_1 \leq \nu_2 \Leftrightarrow \Omega^*(\nu_1) \supseteq \Omega^*(\nu_2)$.*

If a bandit is indexable, we can define the Whittle index of a state (for more details see [10], [11]).

**Definition 2.** *Given an indexable bandit, the Whittle index $\nu_x$ of a state $x$, is the largest value of $\nu$ such that it is optimal to activate the bandit in that state. That is, $v_x = \sup\{\nu | x \in \Omega^*(\nu)\}$.*

From the above definition, we have

$$\nu_x \geq \nu \Leftrightarrow x \in \Omega^*(\nu). \tag{3}$$

The Fig. 2 illustrates the indexablity of an instance of (SUBP-$k$) with $N = 200$ and $\eta = 1/3$. Once a state is in the passive zone it never comes back into the active zone when the multiplier $\nu$ is increased. Later, we shall prove formally that (SUBP-$k$) is indeed indexable. The index $\nu_x$ gives us an indication of how profitable it is to activate the bandit (or serve the user) in state $x$. If $\nu_x > \nu_y$, it means that a higher penalty is required not to serve in state $x$ compared to state $y$. That is, it is more profitable to serve in a state with a higher Whittle index.

This motivates the following *heuristic policy*: given the state, the data rate, and the class of each user in the coverage, the AP serves the user with the highest current Whittle index. For this heuristic to be work, the bandits need to be indexable. In the next section, we show that this is true for the bandits defined by (SUBP-$k$) and give a relatively cheap method for the computation of the Whittle indices.

### A. Indexability

In the rest of the paper, we shall make the following assumptions which will simplify the presentation. The results carry over under the more general conditions mentioned in Assumptions 1 and 2.

**Assumption 3.** *1) The probability of leaving due to service in a time-slot in state $x$ is approximated by $\eta_b r_x$.*

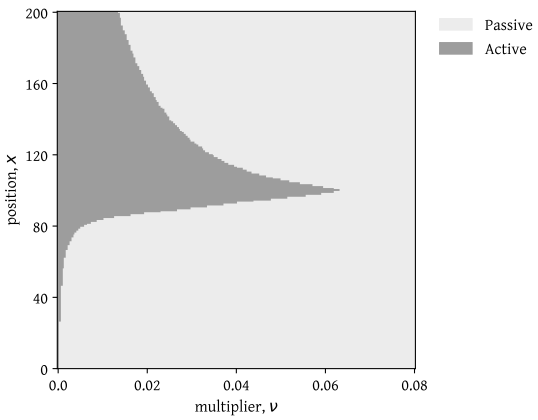


Fig. 2: Illustration of indexability.The darker (resp. lighter) region indicates the bandit is active (resp. passive). $\eta = 1/3$ and $N = 200$.

*2) The reward function is $R_k(x, a) = r_x \cdot \eta_k \cdot a$. Here, the factor $\eta_k$ can be seen as the weight of user-k.*

The approximation of probability is justified when the duration of time-slot is small compared to the average time required for service completion. Hence, the probability of departure due to service completion, $1 - e^{-\eta_b r_x}$ can be approximated by $\eta_b r_x$.

For indexability, we shall restrict the domain of $\nu$ to $[0, +\infty)$. Denote by $V(x, \nu)$ the value function, i.e.

$$V(x, \nu) = \max_\pi \sum_{t=0}^{N} \mathbb{E}_x^\pi (R(X(t), a(t)) - \nu \sum_{t=0}^{N} \mathbb{E}_x^\pi (a(t)).$$

From Lemma 1, we know that even though (SUBP-$k$) is a finite-horizon problem, its value function satisfies Bellman's equation. That is, $V$ is the solution of

$$V(x, \nu) = \max_{a \in \{0,1\}} \left( (R(x, a) - \nu)a + \sum_{y \in \mathcal{N}} \mathbf{P}(y|x, a) V(y, \nu) \right).$$

Replacing $R(x, a)$ and $\mathbf{P}(y|x, a)$ with values from Assumption 3, the above equation simplifies to:

$$V(x, \nu) = \max_{a \in \{0,1\}} \{(r_x \eta - \nu)a + (1 - r_x \eta a) V(x + 1, \nu)\}. \tag{4}$$

Recall that state $N+1$ is the terminal state in which the user has left, so $V(N + 1, \nu) = 0$ for any $\nu$. For $x = 0, 1, ...N$, define:

$$\begin{aligned} V^0(x, \nu) &= V(x + 1, \nu), \\ V^1(x, \nu) &= (r_x \eta - \nu) + (1 - r_x \eta) V(x + 1, \nu). \end{aligned}$$

Eqn. (4) is then $V(x, \nu) = \max_{\{0,1\}} \{V^1(x, \nu), V^0(x, \nu)\}$.

**Remark 1.** *$V^1(x, \nu), V^0(x, \nu), V(x, \nu)$ are continuous and non-increasing in $\nu$ since each is the maximum of finite number of continuous and non-increasing functions of $\nu$.*

From the definition of indexability (see Definition 2), Whittle's index of state $x$, is the value of $\nu$ such that

$$V^1(x, \nu_x) = V^0(x, \nu_x) \text{ with } \nu \in [0, +\infty). \tag{5}$$

We shall prove that for any $x$, (5) has *exactly one* solution, called $\nu_x$, and thus it is the Whittle index of the state $x$. The existence

and uniqueness of the solution implies indexability. Indeed, if (5) has a unique solution, then due to continuity of $V^1(x,\nu)$ and $V^0(x,\nu)$ in $\nu$ it implies that the sign of $V^1(x,\nu) - V^0(x,\nu)$ changes only once in $[0,\infty)$ and this change happens at $\nu_x$. Since $V^1(x,\infty) < V^0(x,\infty)$, we have $V^1(x,\nu) < V^0(x,\nu)$ for $\nu \in [\nu_x, \infty)$ and $V^1(x,\nu) \geq V^0(x,\nu)$ otherwise. This argument will be made formal in Theorem 1 below.

Assume $N$ is even (the arguments of the proof also work when $N$ is odd). It will be convenient to divide the state-space, $\mathcal{N}$, into two subsets: one on the left of the AP, $\mathcal{N}^- = \{0,1,2,...,N/2-1\}$ and one on the right of the AP (including in front of the AP), $\mathcal{N}^+ = \{N/2, N/2+1, ..., N+1\}$. For convenience, define $f(x,\Delta)$ for $x \in \mathcal{N}^-, \Delta \in \mathcal{N}^+$ as follows:

$$f(x,\Delta) := \frac{r_x\eta\big(1 - \sum_{i=x+1}^{\Delta} r_i\eta \prod_{j=x+1}^{i-1}(1-r_j\eta)\big)}{1 - r_x\eta(\sum_{i=x+1}^{\Delta}\prod_{j=x+1}^{i-1}(1-r_j\eta))}. \quad (6)$$

The following theorem shows the indexability and gives the formula for the unique solution and characterizes the behavior of the indices.

**Theorem 1** (Indexability)**.** *For each state $x$, the equation (5) has a unique solution denoted by $\nu_x$.* (SUBP-$k$) *is thus indexable. Further, the index is given by:*

*1.1 For $x \in \mathcal{N}^+$, that is, on the right, $\nu_x = r_x\eta$, and $\nu_{N/2} > \nu_{N/2+1} > \cdots > \nu_N$.*

*1.2 For $x \in \mathcal{N}^-$, that is, on the left, $\nu_x = f(x,\Delta(x))$ where $\Delta(x) \in \mathcal{N}^+$ such that $f(x,\Delta(x)) \in [r_{\Delta(x)+1}\eta, r_{\Delta(x)}\eta)$, and $\nu_0 < \nu_1 < \cdots < \nu_{N/2-1} < \nu_{N/2}$.*

The index of the states on the right-hand side ($x \in \mathcal{N}^+$) is straightforward, and for $x \in \mathcal{N}^-$ a simple linear search yields $\Delta(x)$ which can be plugged into (6) to obtain the index.

Consider two symmetric states $x$ and $N-x$, with one on the left and the other on the right of the AP. The following proposition shows that Whittle's index policy always gives more priority to the state on the right hand side.

**Proposition 1.** *(Right priority) Suppose $r_x$ is symmetric about $x = N/2$, that is $r_{N/2-x} = r_{N/2+x}$. If $x$ and $y$ are symmetric ($x + y = N$), with $x$ on the left ($x < N/2$) and $y$ on the right ($y \geq N/2$), then $\nu_x < \nu_y$.*

For symmetric states, the Whittle index gives priority to the users on the right-hand side because they leave the system earlier than users on the left who will pass through much more favorable channel conditions later on. Thus, one can wait to serve them later and hope to get a better reward.

## IV. Whittle-index based policy

We now come back to the original optimization problem (OBJEC) for which we propose the following heuristic based on Theorem 1.

**Whittle-Index Policy (WIP)** (See Algorithm WIP): In each time-slot, the AP takes as input the current position, the data rate and the mean service requirement of each user in its coverage range. Using Theorem 1, it computes the Whittle index for each user. The channel is allocated to the user that has the highest current Whittle index. If there are two or more users with the same index, one is chosen arbitrarily.

The WIP shall be compared with the following policies.

- *Optimal*: obtained by solving (OBJEC) (or (SUBP-$k$) depending upon the scenario). The optimal policy can only be computed for small number of time-slots so will not be shown when this number is large.
- *Greedy*: chooses the user with the best one-step reward, that is, the user with the largest $r_x\eta_k$.
- *Gittins*: serves the user with the best Gittins' index, which is defined as follows. Let $\tau_s = \min(s, \hat{\tau})$ with $\hat{\tau} = \inf\{t \geq 0 : X(t) = N+1\}$. Then, for state $x$,

$$Gi(x) = \sup_{s \geq 1} \frac{\sum_{t=0}^{\tau_s - 1} \mathbb{E}_x(R(X(t), a(t) = 1)}{\tau_s}. \quad (7)$$

- *RMS*: gives priority to the right-most user.
- *LMS*: gives priority to the left-most user.

**Algorithm WIP:** Heuristic based on Whittle indices

```
for every time step t do
    Input  : Vectors X(t), r_{X(t)}, and η
    Output: a*
    a* ← 0
    i = arg max_{k∈K(t)} ν_{k,X_k(t)}  /* choose one
        arbitrarily, if more than one          */
    a*_i ← 1
end
```

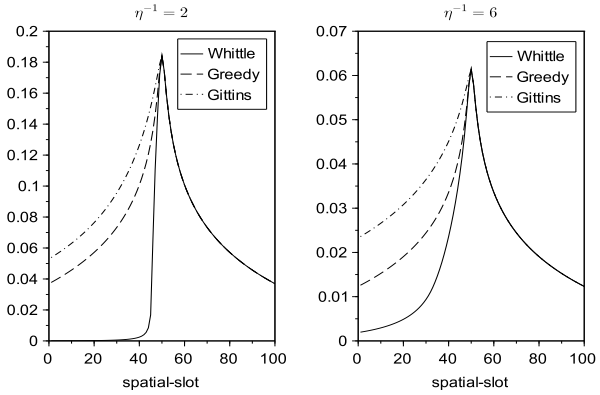


Fig. 3: Comparison of the one-step reward curve (Greedy), Gittins' index and the Whittle's index for two different values of $\eta$.

Fig. 3 illustrates the Whittle index, and compares it with the one-step reward (which is also the departure probability), and Gittins' index. The one-step reward can be seen as the index of the greedy algorithm which chooses the user with the highest one-step reward. To the right of the AP, all the three indices coincide. On the left, however, Whittle index is below the index for greedy (Prop. 1) and Gittins' index is above greedy.

In the first setting, there are no arrivals and all users belong to the same class. The number of time-slots is $N = 100$, the mean service requirement is $\eta^{-1} = 1$, and the rate curve, $r_x$, is given in Fig. 3. There are $K$ cars at time 0 and their initial position is chosen randomly. Several runs were performed, each with a different initial condition.

For various values of $K$, Table I gives the average total reward obtained after averaging over 1000 experiments for different policies. The optimal policy is not evaluated because of the large size of the state-space. The last column of the table gives the percentage improvement of WIP with respect to the greedy policy. WIP outperforms all the other policies, except for RMS

TABLE I: Comparison of policies in case of no arrivals

| # users $K$ | **Whittle** (W) | **Greedy** (G) | **Gittins** | **RMS** | **LMS** | **% gain** W vs G |
|---|---|---|---|---|---|---|
| 2 | 0.0195 | 0.0188 | 0.0187 | 0.0197 | 0.0186 | 3.7 |
| 5 | 0.043 | 0.039 | 0.039 | 0.045 | 0.037 | 10.3 |
| 10 | 0.082 | 0.070 | 0.069 | 0.077 | 0.061 | 17.1 |
| 20 | 0.129 | 0.114 | 0.113 | 0.080 | 0.086 | 13.2 |
| 40 | 0.195 | 0.190 | 0.189 | 0.080 | 0.109 | 2.6 |
| 60 | 0.226 | 0.224 | 0.224 | 0.080 | 0.122 | 0.9 |

for low values of $K$. At the two extreme values of $K$, both WIP and Greedy have the same performance but for moderate number of users one can gain up to $17\%$ with WIP.

**Numerical results when there are new arrivals:** We now move to setting with new arrivals to the systems. Recall that in each time-slot a new user of class-$b$ arrives with a probability $p_b$ in the left-most spatial-slot. We first compare the policies for a small number of spatial-slots, $N = 11$, and one class of users. This allows us to compute the average reward of the optimal policy. In Fig. 4, the average total reward is plotted for the policies as a function of the probability of new arrival. We observe that Whittle policy almost overlaps with the optimal policy, and outperforms all the others. The closeness to optimality is a coincidence and need not always happen.

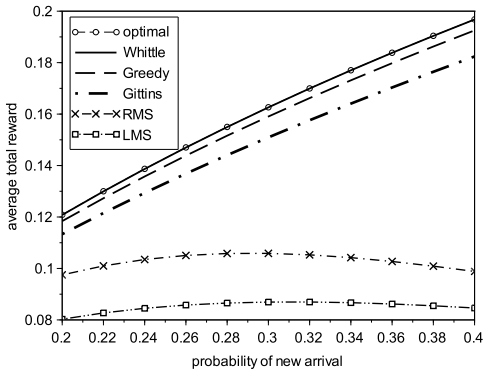


Fig. 4: Comparison of policies when there are new arrivals, number of spatial-slots is small ($N = 11$) and one class of users.

As a final comparison, for $N = 100$ and three classes of users, we show in Fig. 5 the average total reward of different policies as a function of the probability of new arrival denoted by $p$, after averaging over $1000$ experiments for different policies. Here $p_b = 1/3p$, $\forall b$, that is, a new arrival belongs to one of the three classes with equal probability. The mean service requirements of the three classes are: $\eta_1^{-1} = 0.8$, $\eta_2^{-1} = 1.4$, and $\eta_3^{-1} = 4.2$. The optimal policy cannot be computed for this $N$. This time the optimal policy is not shown because the state-space is too big to allow for its computation.

It is observed that the Whittle policy performs much better than the greedy policy when there are more number of classes. The improvement is visible for a larger range of the probability of new arrivals compared to when there is only one class of users.

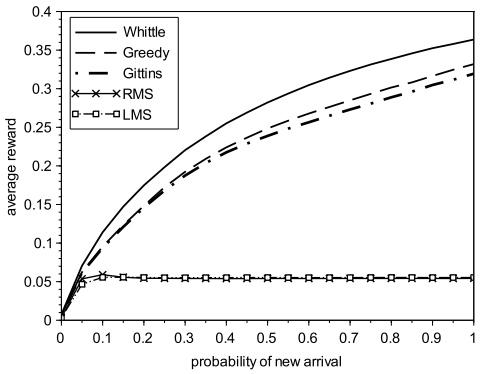


Fig. 5: Policy comparison when there are new arrivals, number of spatial-slots is large ($N = 100$) and three user classes.

## V. Conclusions and future work

Whittle's restless multi-armed bandit provides an elegant framework for computing scheduling decisions in a drive-thru internet scenario. The mobility model of this scenario makes the problem indexable and allows for an easy computation of the Whittle index. This index has the property that, between two users who have the same rate but who are on the opposite sides on the access point, it gives priority to the user on the right because the user on the left can be served later on. It was seen from numerical experiments that the heuristic policy based on Whittle index (WIP) outperforms the greedy policy in various settings including dynamic arrivals and heterogeneous users.

This framework opens several interesting questions related to the optimality-gap of the proposed heuristic as well as generalizations of indexability to models with users moving on larger networks and with varying speeds. A formal proof showing that WIP is better than greedy is also open.

## References


[1] Nan Cheng, Ning Lu, Ning Zhang, Xuemin (Sherman) Shen, and Jon W. Mark. Vehicular WiFi offloading. *Veh. Commun.*, 1(1):13–21, January 2014.

[2] D. Jia, K. Lu, J. Wang, X. Zhang, and X. Shen. A survey on platoon-based vehicular cyber-physical systems. *IEEE Communications Surveys Tutorials*, 18(1):263–284, Firstquarter 2016.

[3] J. J. Alcaraz, J. Vales-Alonso, and J. Garcia-Haro. Link-layer scheduling in vehicle to infrastructure networks: An optimal control approach. *IEEE Journal on Selected Areas in Communications*, 29(1):103–112, January 2011.

[4] Qiang Zheng, Kan Zheng, Periklis Chatzimisios, and Fei Liu. Joint optimization of link scheduling and resource allocation in cooperative vehicular networks. *EURASIP Journal on Wireless Communications and Networking*, 2015(1):170, Jun 2015.

[5] D. Jia, R. Zhang, K. Lu, J. Wang, Z. Bi, and J. Lei. Improving the uplink performance of drive-thru internet via platoon-based cooperative retransmission. *IEEE Transactions on Vehicular Technology*, 63(9):4536–4545, Nov 2014.

[6] Q. Wang, P. Fan, and K. B. Letaief. On the joint V2I and V2V scheduling for cooperative vanets with network coding. *IEEE Transactions on Vehicular Technology*, 61(1):62–73, Jan 2012.

[7] Richard R. Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

[8] 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*, SIGMETRICS '15, pages 57–69, New York, NY, USA, 2015. ACM.

[9] M. Larrnaaga, U. Ayesta, and I. M. Verloop. Dynamic control of birth-and-death restless bandits: Application to resource-allocation problems. *IEEE/ACM Transactions on Networking*, 24(6):3812–3825, December 2017.

[10] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.
[11] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, 2011.
[12] Vivek S. Borkar and Sarath Pattathil. Whittle indexability in egalitarian processor sharing systems. *Annals of Operations Research*, 2017.
[13] Arjun Anand and Gustavo de Veciana. A Whittle's index based approach for qoe optimization in wireless networks. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '18, pages 39–39, New York, NY, USA, 2018. ACM.
[14] C. Buyukkoc, P. Varaya, and J. Walrand. The $c\mu$ rule revisited. *Adv. Appl. Prob.*, 17:237–238, 1985.
[15] M. Xing, J. He, and L. Cai. Maximum-utility scheduling for multimedia transmission in drive-thru internet. *IEEE Transactions on Vehicular Technology*, 65(4):2649–2658, April 2016.
[16] L.E. Schrage and L.W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14:670–684, 1966.
[17] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of optimal queueing network. *Mathematics of Operations Research*, 24(2):293–305, 1999.
[18] J. Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15(2):161–198, 2007.
[19] I.M. Verloop. Asymptotically optimal priority policies for indexable and non-indexable restless bandits. *Annals of Applied Probability*, 2016.
[20] Dimitri P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1987.
[21] Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

## APPENDIX

We divide the proof of Theorem 1 into three steps:

- Proof of the existence of the solution of (5), which is given in Proposition 2.
- Proof of Theorem 1.1 for the states on the right. The statement is given in Proposition 3.
- Proof of Theorem 1.2 for the states on the left. This is presented in Proposition 4.

**Proposition 2.** *For each $x$, the* (5) *has at least one solution* $\nu \in (0, +\infty)$.

For the proof of Prop. 2, we need the following two lemmas.

**Lemma 2.** *If $\nu = 0$, $V^1(x,0) > V^0(x,0)$, for $x = 0, 1, \dots N$.*

The proof of this lemma has been omitted due to space restrictions.

**Lemma 3.** *For any $\nu > \max_{x=0,1,\dots N}\{r_x\eta\}$, $V^1(x,\nu) < V^0(x,\nu)$, $x = 0, 1, \dots N$.*

*Proof:* We prove by induction in reverse direction from state $N$ to state 0. For $x = N$,

$$V^1(N,0) = r_N\eta - \nu < 0 = V^0(N,0),$$

since $r_x\eta < \nu$. So, $V(N,0) = V^1(N,0)$.

Suppose the claim is true until state $x$, i.e, $V^1(y,0) < V^0(y,0)$ so that, for any $y \geq x$, $V(y,0) = V^0(y,0)$. Then, by value interaction we have:

$$V(y,0) = 0$$

for any $y \geq x$. Thus, for state $x-1$, we have

$$V^1(x-1,0) = r_{x-1}\eta - \nu < 0 = V^0(x-1,0).$$

■

*Proof of Prop. 2:* Using these two above lemmas and the continuity $V^1(x,\cdot)$ and $V^0(x,\cdot)$ in the second variable, we conclude that (5) each has at least one solution. ■

For the other two parts of the proof of Theorem 1, we need the following lemmas. The proofs have been omitted due to lack of space.

**Lemma 4.** *If $A_1, B_1, A_2, B_2, ..., A_k, B_k > 0, k \geq 2$ and $\frac{A_1}{B_1} < \frac{A_2}{B_2} < \dots < \frac{A_k}{B_k}$ then*

$$\frac{A_1}{B_1} < \frac{A_1 + A_2}{B_1 + B_2} < \dots < \frac{A_1 + A_2 + \dots + A_k}{B_1 + B_2 + \dots + B_k}.$$

**Lemma 5.** *If $A_1 > A_2 + A_3 + \dots + A_k > 0, B_1 > B_2 + B_3 + \dots + B_k > 0, k \geq 2$ and $\frac{A_1}{B_1} > \frac{A_2}{B_2} > \dots > \frac{A_k}{B_k}$ then*

$$\frac{A_1}{B_1} < \frac{A_1 - A_2}{B_1 - B_2} < \dots < \frac{A_1 - A_2 - \dots - A_k}{B_1 - B_2 - \dots - B_k}.$$

For every $\Delta > x$, define:

$$a(x,\Delta) := 1 - \sum_{i=x+1}^{\Delta} r_i\eta \prod_{j=x+1}^{i-1}(1 - r_j\eta),$$

$$b(x,\Delta) := \sum_{i=x+1}^{\Delta} \prod_{j=x+1}^{i-1}(1 - r_j\eta).$$

Then we can rewrite $f$ as:

$$f(x,\Delta) = \frac{r_x\eta a(x,\Delta)}{1 - r_x\eta b(x,\Delta)}.$$

We now state the result on the uniqueness of the solution of (5) on the right hand side, and give its properties. We also characterize the behaviour of the value function in this region which will be used later for the proof of Theorem 1.

**Proposition 3.** *(For Theorem 1.1) For any $x \geq N/2$, Eqn. (5) has a unique solution $\nu_x$. Moreover, we have:*

1) *$\nu_x = r_x\eta$. Thus, $\nu_{N/2} > \nu_{N/2+1} > \dots > \nu_N$.*
2) *The value function takes the following form:*
   - *If $\nu \geq r_x\eta$, then $V(x,\nu) = 0$,*
   - *If $0 \leq \nu < r_x\eta$, then*

$$V(x,\nu) = \sum_{i=x}^{y} \left( \prod_{j=x}^{i-1}(1 - r_j\eta) \right) (r_i\eta - \nu), \qquad (8)$$

*where $y \in \{x, x+1, ..., N\}$ is such that $r_{y+1}\eta \leq \nu < r_y\eta$, with the convention $r_{N+1} = 0$.*

The proof of this proposition has been omitted due to space restrictions. Next, we move to the proof of Theorem 1.2.

**Proposition 4.** *(For Theorem 1.2) For every state on LHS $x \leq N/2 - 1$, Eqn. (5) has a unique solution $\nu_x$. Moreover, we have:*

1) $$\nu_x = f(x, \Delta^*(x)),$$

*where $\Delta^*(x) \geq N/2$ is chosen such that $f(x,\Delta^*(x)) \in [r_{\Delta^*(x)+1}\eta, r_{\Delta^*(x)}\eta)$.*

2) *$\nu_x$ increases in x on LHS, i.e,*

$$\nu_0 < \nu_1 < \nu_2 < \dots < \nu_{N/2-1} < r_{N/2}\eta = \nu_{N/2}.$$

3) *The value function has the following form:*
   - *If $\nu \geq \nu_x$, then $V(x,\nu) = V(x+1,\nu)$.*
   - *If $0 \leq \nu < \nu_x$, then*

$$V(x,\nu) = (r_x\eta - \nu) + (1 - r_x\eta)V(x+1,\nu).$$

Before proving Prop. 4, we need to characterize properties of function $f(x,\Delta)$ which are described in the following lemma.

**Lemma 6.** *For a fixed $x \in \mathcal{N}^-$, define $D_x = \{\Delta | \Delta \geq N/2, f(x,\Delta) \geq 0\}$. Recall from (6) that $f$ is defined only on integers. Let $\Delta^*(x) \in D_x$ be the smallest value for which $r_{\Delta(x+1)}\eta \leq f(x,\Delta^*(x)) < r_{\Delta^*(x)}\eta$. Then,*

$$\Delta^*(x) = \arg\min_{\Delta \in D_x} f(x,\Delta).$$

*Moreover, for fixed $x$ and considering $f(x,\Delta)$ as a function of $\Delta$ in $D_x$, then $f$ decreases from $N/2$ to $\Delta^*(x)$ and increases from $\Delta^*(x)$ to b. Outside of $D_x$, $f(x,\Delta)$ is either negative or infinity.*

The existence of $\Delta^*(x)$ will be proved in Lemma 7 for $x = N/2-1$ while, for other values of $x$, existence will be shown in the proof of Prop. 3.

**Lemma 7.** *$\Delta^*(N/2-1) \geq N/2$ exists. Further, the equation $V^1(N/2-1,\nu) = V^0(N/2-1,\nu)$ has a unique solution which is: $\nu_{N/2-1} = f(N/2-1,\Delta^*(N/2-1))$.*

*Moreover,*
*1. If $\nu < \nu_{N/2-1}$, then $V(N/2-1,\nu) = (r_{N/2-1}\eta - \nu) + (1 - r_{N/2-1}\eta)V(N/2,\nu)$,*
*2, If $\nu \geq \nu_{N/2-1}$, then $V(N/2-1,\nu) = V(N/2,\nu)$,*
*with $V(N/2,\nu)$ given in Prop. 3.*

We will prove Prop. 4 by induction on states $x \leq N/2-1$. That is, we will show that (5) has a unique solution, $\nu_x$, and that, on the left-hand side, $\nu_x$ increases in $x$, i.e if $x < y \leq N/2-1$ then $\nu_x < \nu_y$.

*Proof of Prop. 4:* We shall prove the claim by induction in the reverse direction. For state $N/2-1$, the claim follows from Lemma 7. Suppose the claim is true until state $x \leq N/2-1$. We now prove for state $x-1$. Consider the equation:

$$r_{x-1}\eta - \nu = r_{x-1}\eta V(x,\nu).$$

By Lemma 2, we know that it has at least one solution. Suppose $\nu$ is a solution of this equation.

- If $\nu \geq r_{N/2}\eta$, then

  $$r_{x-1}\eta - \nu = r_{x-1}\eta V(x,\nu) = \cdots = r_{x-1}\eta V(N+1) = 0,$$

  which follows by the induction hypothesis for all states in $\{x, x+1, ..., N/2-1\}$ and by Prop. 3, for all states in $\{N/2, N/2+1, ..., N\}$. This implies that $\nu = r_{x-1}\eta < r_{N/2}\eta$, which leads to a contradiction with $\nu \geq r_{N/2}\eta$.
- If $\nu_x \leq \nu < \nu_{N/2}$ then there exist $y_1, y_2$ such that:

  $$\begin{cases} x \leq y_1 \leq N/2-1, N/2 \leq y_2 \leq \Delta^*(y_1) \\ \nu \in [\nu_{y_1}, \nu_{y_1+1}) \cap [r_{y_2+1}\eta, r_{y_2}\eta). \end{cases} \quad (9)$$

  So, by induction hypothesis and Prop. 3, we can develop $V(x,\nu)$ to get:

  $$\begin{aligned} V(x,\nu) &= V(x+1,\nu) = \cdots = V(y_1+1,\nu) \\ &= (r_{y_1+1}\eta - \nu) + (1 - r_{y_1+1})V(y_1+2,\nu) \\ &= \cdots \\ &= \sum_{i=y_1+1}^{y_2} \left( \prod_{j=y_1+1}^{i-1} (1 - r_j\eta) \right) (r_i\eta - \nu). \end{aligned}$$

  Now, the equation $r_{x-1}\eta - \nu = r_{x-1}\eta V(x,\nu)$ becomes linear in $\nu$, which can be solved to get:

  $$\nu = \frac{r_{x-1}\eta a(y_1,y_2)}{1 - r_{x-1}\eta b(y_1,y_2)}. \quad (10)$$

  We have:

  $$\nu = \frac{r_{x-1}\eta a(y_1,y_2)}{1 - r_{x-1}\eta b(y_1,y_2)} < \frac{r_{y_1}\eta a(y_1,y_2)}{1 - r_{y_1}\eta b(y_1,y_2)} = f(y_1,y_2), \quad (11)$$

  and we remark that $y_2 \leq \Delta^*(y_1)$. Now, there are two sub-cases:
  - If $y_2 = \Delta^*(y_1)$ then $\nu_{y_1} = f(y_1,\Delta^*(y_1)) = f(y_1,y_2) > \nu$, where the last inequality is due to (11). This contradicts $\nu \geq \nu_{y_1}$ in (9).
  - Suppose $y_2 < \Delta^*(y_1)$. From (9), we have $\nu \in [r_{y_2+1}\eta, r_{y_2}\eta)$, and by (11) we have $\nu < f(y_1,y_2)$. Therefore, $f(y_1,y_2) > r_{y_2+1}\eta$. Note that the statement $f(y_1,y_2) < r_{y_2}\eta$ cannot be true because this will imply that $\Delta^*(y_1) = y_2$. On the other hand, by induction hypothesis, $\Delta^*(y_1)$ is the unique state that satisfies $f(y_1,\Delta^*(y_1)) \in [r_{\Delta^*(y_1)+1}\eta, r_{\Delta^*(y_1)}\eta)$ and we have assumed that $y_2 < \Delta^*(y_1)$. Thus,

    $$f(y_1,y_2) \geq r_{y_2}\eta. \quad (12)$$

    Define $A_1 = r_{y_1}\eta \cdot a(y_1,y_2), B_1 = 1 - r_{y_1}\eta \cdot b(y_1,y_2)$, and for $k = 2, \ldots, \Delta^*(y_1) - y_2 + 1$,

    $$B_k = r_{y_1}\eta \prod_{j=y_1+1}^{y_2+k-2} (1 - r_j\eta), A_k = r_{y_2+k-1} \cdot B_k.$$

    From the above definitions and (12),

    $$\begin{aligned} \frac{A_1}{B_1} &= f(y_1,y_2) \geq r_{y_2}\eta > \frac{A_2}{B_2} = r_{y_2+1}\eta > \cdots \\ &> \frac{A_{\Delta^*(y_1)-y_2+1}}{B_{\Delta^*(y_1)-y_2+1}} = r_{\Delta^*(y_1)}\eta. \end{aligned}$$

    Applying Lemma 5 on $A_k$ and $B_k$, we get:

    $$\begin{aligned} r_{y_2}\eta &\leq f(y_1,y_2) = \frac{A_1}{B_1} \\ &< \frac{A_1 - A_2 - \cdots A_{\Delta^*(y_1)-y_2+1}}{B_1 - B_2 - \cdots - B_{\Delta^*(y_1)-y_2+1}} \\ &= f(y_1,\Delta^*(y_1)) < r_{\Delta^*(y_1)}\eta, \end{aligned}$$

    which is in contradiction to $N/2 \leq y_2 < \Delta^*(y_1)$.
- Finally, suppose $0 < \nu < \nu_x$. Then, following similar arguments as in the proof of Lemma 7 and by existence of the solution of (5), there exists a unique $\Delta^*(x-1) \geq \Delta^*(x)$ such that

  $$\nu = f(x-1,\Delta^*(x-1)) \in [r_{\Delta^*(x-1)+1}\eta, r_{\Delta^*(x)}\eta).$$

We have proved the first two claims of Proposition 4.

Now, we prove the third claim. From Lemma 2, we have $V^1(x,0) > V^0(x,0)$ and from Lemma 3, we know that $V^1(x,\infty) < V^0(x,\infty)$. Since $\nu_x$ is the unique solution of $V^1(x,0) = V^0(x,0)$ and $V^1(x,\nu)$ and $V^0(x,\nu)$ are continuous in $\nu$, we can infer that $V^1(x,\nu) \geq V^0(x,\nu)$ in $[0,\nu_x]$ and $V^1(x,\nu) \leq V^0(x,\nu)$ in $[\nu_x,\infty)$. This implies the claimed form of the value function for state $x$. ■