On the computation of Whittle's Index for Markovian restless bandits

Urtzi Ayesta, Manu K. Gupta, Ina Maria Maaike Verloop

On the computation of Whittle's Index for Markovian restless bandits

Urtzi Ayesta[1], Manu K. Gupta[2]*, and Ina Maria Verloop[1]
[1]IRIT, 2 rue C. Camichel, Toulouse, France
[2]Indian Institute of Technology Roorkee, Uttarakhand-247667, India

October 21, 2020

Abstract

The multi-armed restless bandit framework allows to model a wide variety of decision-making problems in areas as diverse as industrial engineering, computer communication, operations research, financial engineering, communication networks etc. In a seminal work, Whittle developed a methodology to derive well-performing (Whittle's) index policies that are obtained by solving a relaxed version of the original problem. However, the computation of Whittle's index itself is a difficult problem and hence researchers focused on calculating Whittle's index numerically or with a problem dependent approach.

In our main contribution we derive an analytical expression for Whittle's index for any Markovian bandit with both finite and infinite transition rates. We derive sufficient conditions for the optimal solution of the relaxed problem to be of threshold type, and obtain conditions for the bandit to be indexable, a property assuring the existence of Whittle's index. Our solution approach provides a unifying expression for Whittle's index, which we highlight by retrieving known indices from literature as particular cases. The applicability of finite rates is illustrated with the machine repairmen problem, and that of infinite rates by an example of communication networks where transmission rates react instantaneously to packet losses.

keywords: Restless bandits, Whittle index.

1 Introduction

Markov Decision Processes (MDPs) provide a mathematical framework for sequential decision making where outcomes are random. Formally, an MDP is a sequential stochastic control process, where a decision maker aims at minimizing its long term cost. The basic setup is as follows, at each time step, a state dependent cost 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, 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 have received a lot of attention. One such class is the Multi-Armed Bandit Problem (MABP) framework. In an MABP, there are multiple concurrent projects or bandits. The decision maker knows the states of all bandits and the cost in every state, and aims at minimizing the average cost. At every decision epoch, the decision maker needs to select one bandit, the state of this selected bandit evolves stochastically, while the states of all other bandits remain *frozen*. In a ground-breaking result, Gittins showed that the optimal policy that solves an MABP is an index rule, nowadays referred to as Gittins' index policy [Gittins et al., 2011]. Thus, for each bandit, one calculates Gittins' index, which depends only on its own current state and stochastic evolution. The optimal policy activates the bandit with highest current index in each decision epoch.

The Restless Multi-Armed Bandit Problem (RMABP), introduced in Whittle [1988], is a more general class of problems, in which the states of non-selected bandits also evolve randomly. That is, in contrast with an MABP, here the non-selected bandits do not remain *frozen*. RMABPs have become extremely popular over the years, and have been applied in many contexts, including inventory routing, machine maintenance, health-care systems, outsourcing warranty repair, etc. RMABPs can not be solved analytically, except for some toy examples. 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 bandits, and activating in every decision epoch the bandit 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 bandits grows large.

*Corresponding author: Manu K. Gupta (manu.gupta@ms.iitr.ac.in)

Fundamental questions regarding Whittle's index policy concern their existence and their complexity in computation. To prove existence, one needs to establish a technical property known as *indexability*. Computing Whittle's index might be involved, and in practice the indices are computed on a problem-to-problem basis (see more in Section 1.1), either numerically or analytically.

An important feature of the present paper is that – in addition to actions that yield finite transition rates – we allow for actions that induce infinite rates, which in the literature is often referred as *impulsive controls*. In contrast with finite rates that yield transition after an exponentially distributed random time, an impulse control induces an instantaneous transition. Impulsive controls have historically received less attention than the *continuous control case*. However, they do appear naturally in several application domains such as epidemiology, reliability theory, inventory control, and congestion control in communication networks. We refer the interested reader to Dufour and Piunovskiy [2016] for a brief survey on MDPs with impulsive control. In order to calculate Whittle's index in the case of infinite rates, we reformulate the problem with impulse cost to a problem with only cost per unit of time. A similar approach has been exploited in the past in the analysis of continuous time MDPs with impulsive controls, see for example Piunovskiy and Zhang [2020] for the case of a constrained MDP.

In this paper, we focus on the average performance criterion (as presented in the original paper by Whittle). We allow both finite and infinite rates at which a bandit changes state. In Section 2 we present the model and explain how a problem with infinite rates can be reformulated as an equivalent problem with only finite rates. In Section 3 we present a general algorithm that, although computationally demanding, determines whether the problem is indexable and calculates the index for any RMABP with multi-dimensional state space. For a one-dimensional bandit, we then derive sufficient conditions for a problem to be indexable and for the optimal solution to a RMABP to be of threshold type. In Section 4, we show our main result that if threshold policies are optimal and if a certain monotonicity condition holds, Whittle's index can be expressed as a function of the steady-state distribution. In Sections 5 and 6, we apply our unifying characterization to several problems considered in the literature to show that they are indexable, that threshold policies are optimal, and finally to retrieve Whittle's index as a direct application of our unifying analytical expression. In particular, in Section 6.1 we consider the problem of transmitting data over the Internet, as an example of important application in which impulse controls arise naturally, see Avrachenkov et al. [2018].

1.1 Related literature

A classical reference for MDPs is Puterman [2014], and a comprehensive coverage for MABP and RMABP is given in Gittins et al. [2011]. Book length treatments of restless bandits can be found in Jacko [2010] and Ruiz-Hernandez [2008]. A discussion on the nearly-optimal performance of Whittle's index is given in Niño-Mora [2007], and asymptotic optimality of Whittle's index as the number of bandits grows is shown in Weber and Weiss [1990], Verloop [2016]. Further restless bandit formulation has been used in diverse domains (see James et al. [2016], Abbou and Makis [2019]).

For average cost criterion and with a countable state space, Niño-Mora [2006] gives sufficient conditions for an RMABP to be indexable and provides an analytical expression (same as ours) for the Whittle index. One of the conditions consists in showing that the so-called marginal workloads are strictly positive for any set of states. This ensures that optimal policies are of threshold type. Instead, in this paper, we provide an algorithm that does not require threshold optimality, as might be of interest when a bandit lives in a multi-dimensional state space and threshold optimality might be impossible to establish. In addition, in the case when threshold optimality can be established independently, we show that a weaker assumption on the marginal workloads is sufficient for the results to hold.

Even though Whittle's seminal work introduced Whittle's index within the context of average cost criterion, a large body of work has focused on tackling an RMABP under the total discounted cost criterion. For the discounted cost criterion and finite state space, Niño-Mora [2007] provides a thorough analysis and efficient algorithms (based on linear programming) to establish indexability and to give an expression for Whittle's Index. The same approach was undertaken to obtain the Whittle's index for a general RMABP in Niño-Mora [2006].

As explained in the introduction, the analytical computation of the index has mostly been carried out on a problem-to-problem basis. The main idea to calculate them is to sweep the state space, by recursively identifying and calculating the states with higher Whittle's indices. This can be done by iterative schemes as for example in Borkar and Pattathil [2017] for a processor sharing queue, Borkar et al. [2017b] in a problem of cloud computing, Borkar et al. [2017a] for a scheduling problem in a wireless setting, and Pattathil et al. [2017] in the context of content delivery networks. And in some particular cases analytically, see for example Argon et al. [2009] for a load balancing problem with dedicated arrivals, Opp et al. [2005] for outsourcing warranty repairs, Ayer et al. [2019] for Hepatitis C Treatment in US Prisons, and Larrañaga et al. [2016] for restless bandits that are of birth and death type. Another popular approach in the literature to calculate the index for the average cost criterion has been to calculate first the index for the discounted

case, and then let the discounting factor tend to one. This is the approach undertaken in e.g. Glazebrook et al. [2005] for the machine repairman problem, Ansell et al. [2003] for a multi-class queue with convex holding costs, and Nino-Mora [2002] for a queue with admission control. A feature that renders the discounted problem more amenable is that the dynamic programming equation of the MDP has only one unknown, the value function, whereas for the average cost criterion, the dynamic programming equation has two unknowns, one being the value function and the other one the average performance (Puterman [2014]).

In this paper, we take a direct approach and work directly with the average cost criterion. This will allow us to obtain a unifying framework to write an analytical expression for the Whittle's index. As we will explain in Sections 5 and 6, all applications mentioned above for which Whittle's index was found are particular instances of our unifying approach. Note that this includes the examples for which Whittle's index was so far only calculated by iterative numerical scheme.

The large majority of the literature on Whittle index has focused on problems without impulsive control. A notable exception is Ford et al. [2019] which studies the dynamic allocation of assets under a total cost criteria. In order to derive Whittle's index, Ford et al. [2019] first invokes results from Dufour and Piunovskiy [2015] in order to reformulate the problem in which impulse controls are replaced by actions that do not lead to an immediate transition.

2 Model description

We consider an RMABP with K ongoing projects or bandits. At any moment in time, bandit k, $k = 1, ..., K$, is in a certain state $n_k \in \mathcal{N}^d$, with $d \in \mathcal{N}^+$. Decision epochs are defined as the moments when one of the bandits changes its state. At each decision epoch, the controller decides for each bandit to either make the bandit passive, action $a = 0$, or to make the bandit active, action $a = 1$. The latter needs to be done such that some constraint on the number of bandits that are activated is satisfied, as will be discussed below.

Throughout this paper, we consider bandits that are modeled as a continuous-time Markov chain, that is, when bandit k is in state n_k, it changes the state after an exponentially distributed amount of time. Transition rates for bandit k, which can either be finite or infinite, depend only on the bandits' state n_k and the action chosen for this bandit. Let $\mathcal{I}_k(n_k, a)$ be an indicator function for the event of an infinite (impulse) transition for bandit k at state n_k under action a, i.e.,

$$\mathcal{I}_k(n_k, a) := \begin{cases} 1 \text{ if state changes instantaneously,} \\ 0 \text{ otherwise.} \end{cases}$$

If $\mathcal{I}_k(n_k, a) = 1$, let $p_k^a(n_k, m_k)$ be the probability of making an immediate transition to state m_k from state n_k. If $\mathcal{I}_k(n_k, a) = 0$, let $q_k^a(n_k, m_k)$ be the finite transition rate of going from state n_k to m_k under action $a \in \{0, 1\}$. Note that the state of a bandit can evolve both when being active and passive, and this does not depend on the states/actions of other bandits. Hence, given the action taken, the dynamics of each bandit is independent of the others.

A policy ϕ decides which bandits are made active. Because of the Markovian property, we focus on policies which base their decision only on the current states of the bandits. For policy ϕ, $N_k^\phi(t)$ denotes the state of bandit k at time t, and $\vec{N}^\phi(t) = (N_1^\phi(t), \ldots, N_K^\phi(t))$. Let $S_k^\phi(\vec{N}^\phi(t)) \in \{0, 1\}$ represent whether or not bandit k is made active at time t, that is, it equals 1 if the bandit is activated, and 0 otherwise.

For bandit k, let $f_k(n_k, a)$ be a function of state n_k and action a and assume it is bounded by a polynomial in n_k. Further assume that if $\mathcal{I}_k(n_k, a) = 1$, then $f_k(n_k, a) = 0$. A policy ϕ is called feasible if it satisfies the following constraint:

$$\sum_{k=1}^{K} f_k(N_k^\phi(t), S_k^\phi(\vec{N}(t))) \leq M, \tag{1}$$

where M is some given constant. That is, (1) gives a constraint on the number of activated/passive bandits resulting in finite transition rates. We denote by $\mathcal{U}$ the set of policies that satisfy the constraint (1) and make the system ergodic. We assume that such a policy exists, hence $\mathcal{U}$ is non-empty.

Since $f_k(n_k, a) = 0$ if $\mathcal{I}_k(n_k, a) = 1$, it is obvious that there is no constraint on the number of impulses at a given time t (we refer to Remark 2.1 for a discussion on the control of impulses). For non-impulsive control, setting $f_k(N_k^\phi, S_k^\phi(\vec{N})) = S_k^\phi(\vec{N})$, the constraint (1) reduces to $\sum_{k=1}^{K} S_k^\phi(\vec{N}) \leq M$, that is, at most M out of K bandits can be made active. This is a slight variation of the classical restless bandit problem, where exactly M bandits need to be activated at any time t[1]. Another interesting function is when $f_k(N_k^\phi, S_k^\phi(\vec{N}))$ represents the expected capacity

[1]This can be shown by introducing so-called dummy bandits with zero cost and fixed state, see Verloop [2016].

occupation (volume), in which case the constraint (1) reduces to the family of sample-path knapsack capacity allocation constraints as recently explored in Jacko [2016], Graczová and Jacko [2014].

Let $C_k(n_k, a)$ denote the cost per unit of time when bandit k is in state n_k and is either passive ($a = 0$) or active ($a = 1$). Let $L_k^{\infty}(n_k, m_k, a)$ be the lump-sum cost when bandit k under action a immediately changes state from n_k to m_k. We assume that both $C_k(n_k, a)$ and $L_k^{\infty}(n_k, m_k, a)$ can be bounded by a polynomial in n_k and m_k. Instead of a lump-sum cost $L_k^{\infty}(n_k, m_k, a)$, we will work with $C_k^{\infty,\phi}(\vec{n}, a)$, which is defined as the cost per unit of time due to impulse transitions when policy ϕ is implemented, we are in state $\vec{n}$ and action a is taken. This is given by

$$C_k^{\infty,\phi}(\vec{n}, a) := \sum_{\tilde{n}_k} \sum_{m_k} \Big(q_k^a(n_k, \tilde{n}_k) \times \mathcal{I}_k(\tilde{n}_k, S_k^{\phi}(\vec{M}_k(\vec{n}, \tilde{n}_k))) \times p_k^{S_k^{\phi}(\vec{M}_k(\vec{n}, \tilde{n}_k))}(\tilde{n}_k, m_k) \times L_k^{\infty}(\tilde{n}_k, m_k, S_k^{\phi}(\vec{M}_k(\vec{n}, \tilde{n}_k))) \Big),$$

where $\vec{M}_k(\vec{n}, \tilde{n}_k)$ equals the state $\vec{n}$ in which the kth component of $\vec{n}$ is replaced by $\tilde{n}_k$.

This can be seen as follows. The cost per unit of time due to impulse transitions consists of the transition rate at which you go to an impulse situation, times its corresponding lump-sum cost. Regarding the first, we have such a transition for some bandit k, which has a transition to some state $\tilde{n}_k$, hence the new state is $\vec{M}_k(\vec{n}, \tilde{n}_k)$, in which it experiences an impulse. Hence, it concerns a transition rate $q_k^a(n_k, \tilde{n}_k)$ if and only if $\mathcal{I}_k(\tilde{n}_k, S_k^{\phi}(\vec{M}_k(\vec{n}, \tilde{n}_k))) = 1$. Now, regarding the corresponding lump-sum cost, this cost depends on the state m_k where bandit k ends up after the impulse. That is, we need to multiply probability $p_k^{S_k^{\phi}(\vec{M}_k(\vec{n}, \tilde{n}_k))}(\tilde{n}_k, m_k)$ with the corresponding lump-sum cost $L_k^{\infty}(\tilde{n}_k, m_k, S_k^{\phi}(\vec{M}_k(\vec{n}, \tilde{n}_k)))$.

The objective is to find a policy $\phi \in \mathcal{U}$ that minimizes the long-run average cost:

$$C^{\phi} := \limsup_{T\to\infty} \sum_{k=1}^{K} \frac{1}{T} \mathbb{E}\left(\int_0^T C_k(N_k^{\phi}(t), S_k^{\phi}(\vec{N}^{\phi}(t))) + C_k^{\infty,\phi}(\vec{N}^{\phi}(t), S_k^{\phi}(\vec{N}^{\phi}(t)))dt \right). \tag{2}$$

The first term is a contribution from holding cost per unit of time and the second term corresponds to the lump-sum cost due to impulses.

Remark 2.1. *There is no sample-path constraint on the impulse control, that is, on the number of impulses allowed at each moment in time, as this will be by default satisfied for any policy (see (1)). Instead, in order to control the number of impulses in the system, one can set the lump-sum cost for infinite transitions, $L_k^{\infty}(n, m, a)$, appropriately. Another method would be to include a constraint on the* time-average *number of impulses. The later is beyond the scope of the paper, as this would imply deriving an index-based heuristic that satisfies a time-average constraint, and, to the best of our knowledge, no index heuristics have been proposed for such settings.*

3 Lagrangian Relaxation and Whittle's index policy

Finding a policy that minimizes the long run average cost (2) under constraint (1) is intractable in general. In fact, it is shown in Papadimitriou and Tsitsiklis [1999] that the restless bandit problems are PSPACE complete, which is much stronger evidence of intractability than NP-hardness. Following Whittle [1988], a very fruitful approach has been to study the relaxed problem in which the constraint (1) is replaced by its time-averaged version, that is,

$$\limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left(\int_0^T \sum_{k=1}^{K} f_k(N_k^{\phi}(t), S_k^{\phi}(\vec{N}^{\phi}(t)))dt \right) \leq M. \tag{3}$$

Let $\mathcal{U}^{REL}$ be the set of stationary policies ϕ that satisfy (3) and for which the Markov chain is ergodic. Note that the set of policies that make the relaxed problem ergodic includes $\mathcal{U}$, i.e., the set of ergodic policies for the original problem. The objective of the relaxed problem is hence to determine a policy that solves (2) under constraint (3). An optimal policy for the relaxed problem then serves as a heuristic for the original optimization problem.

Using the Lagrangian approach, we write the relaxed problem as the following unconstrained problem: find a policy ϕ that minimizes $\mathcal{C}^{\phi}(W) :=$

$$\limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left(\int_0^T \left(\sum_{k=1}^{K} C_k(N_k^{\phi}(t), S_k^{\phi}(\vec{N}^{\phi}(t))) + C_k^{\infty,\phi}(\vec{N}^{\phi}(t), S_k^{\phi}(\vec{N}^{\phi}(t))) - W \left(\sum_{k=1}^{K} f_k(N_k^{\phi}(t), S_k^{\phi}(\vec{N}^{\phi}(t))) - M \right) \right) dt \right), \tag{4}$$

where W is the Lagrange multiplier. The latter can be decomposed into K subproblems, one for each bandit k, that is, minimize

$$\mathcal{C}_k^{\phi} := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \left(C_k(N_k^{\phi}(t), S_k^{\phi}(N_k^{\phi}(t))) + C_k^{\infty,\phi}(N_k^{\phi}(t), S_k^{\phi}(N_k^{\phi}(t))) - W f_k(N_k^{\phi}(t), S_k^{\phi}(N_k^{\phi}(t)))\right) dt\right). \tag{5}$$

With slight abuse of notation, in the above equation, we replaced $\vec{N}^{\phi}(t)$ by $N_k^{\phi}(t)$. We can do so since the stochastic evolution of bandits are independent across each other. Hence, the optimal action for a given bandit (when there is no longer a common constraint on the number of activated bandits) does not depend on the state of other bandits. In other words, in the decomposition, for a given bandit k, the optimal action only depends on its own state.

The solution to (4) is obtained by combining the solution to the K separate optimization problems (5). Under a stationarity assumption, we can invoke ergodicity to show that (5) is equivalent to minimizing

$$\mathbb{E}(C_k(N_k^{\phi}, S_k^{\phi}(N_k^{\phi}))) + \mathbb{E}(C_k^{\infty,\phi}(N_k^{\phi}, S_k^{\phi}(N_k^{\phi}))) - W\mathbb{E}(f_k(N_k^{\phi}, S_k^{\phi}(N_k^{\phi}))), \tag{6}$$

where N_k^{ϕ} is distributed as the stationary distribution of the state of bandit k under policy ϕ. If $f_k(N_k^{\phi}, S_k^{\phi}(N_k^{\phi})) = 1 - S_k^{\phi}(N_k^{\phi})$, the Lagrange multiplier W can be interpreted as subsidy for passivity. For ease of notation, we denote the total expected cost of bandit k under policy ϕ by

$$\mathbb{E}(T_k(N_k^{\phi}, S_k^{\phi}(N_k^{\phi}))) := \mathbb{E}(C_k(N_k^{\phi}, S_k^{\phi}(N_k^{\phi}))) + \mathbb{E}(C_k^{\infty,\phi}(N_k^{\phi}, S_k^{\phi}(N_k^{\phi}))).$$

3.1 Indexability and Whittle's Index

Indexability is a property that allows us to develop a heuristic for the original problem. This property imposes that as the Lagrange multiplier, W, increases, the collection of states in which the optimal action is passive increases.

Definition 3.1. *A bandit is indexable if the set of states in which passive is an optimal action in (6) (denoted by $D_k(W)$) increases in W, that is, $W' < W \Rightarrow D_k(W') \subseteq D_k(W)$.*

Note that in case the set of states $D_k(W)$ is decreasing in W, one can simply switch the role of active and passive for bandit k, and the bandit would be indexable. If the RMABP is indexable, Whittle's index in state N_k is defined as follows:

Definition 3.2. *When a bandit is indexable, Whittle's index in state n_k is defined as the smallest value for the subsidy such that an optimal policy for (6) is indifferent of the action in state n_k. Whittle's index is denoted by $W_k(n_k)$.*

Given that the indexability property holds, Whittle [1988] established that the solution to the relaxed control problem (4) will be to activate all bandits that are in a state n_k such that their Whittle's index exceeds the Lagrange multiplier, i.e., $W_k(n_k) > W$. This optimal solution to the relaxed problem might be infeasible for the original model where the constraint (1) has to be satisfied at each decision epoch. Hence, we have the following index-based heuristic for the original problem with the general resource constraint (1).

Definition 3.3 (Index-based policy)**.** *Assume at time t we are in state $\vec{N}(t) = \vec{n}$. The index-based heuristic will activate bandits in a greedy manner: activate the bandit with the highest non-negative Whittle's index, $W_k(n_k)$, until either the constraint is met or there are no more bandits with non-negative index to activate.*

In our definition of the index-based policy, bandits with a negative Whittle's index are not activated. This is a straight consequence of the relaxed optimization problem: when the Whittle's index is negative in state m, i.e., $W_k(m) < 0$, bandit k in state m is made active only if $W < W_k(m)$. Since $W_k(m) < 0$, the bandit is activated only if a cost W is paid for being passive. In case $f_k(N_k^{\phi}, S_k^{\phi}(\vec{N}_k^{\phi})) = S_k^{\phi}(N_k^{\phi})$, this definition reduces to the Whittle's index policy as studied in classical RMABPs. In that setting, the sample-path constraint (1) can also have a strict equality sign, in which case one can simply adapt the index-based policy by activating the M bandits with the highest indices, even though these indices can be negative. We further note that in case the function $f_k(\cdot,\cdot)$ represents the expected capacity occupation, the index-based rule as defined above has been numerically shown to perform well in Graczová and Jacko [2014].

Remark 3.4. *In the case of a single bandit, i.e., $K = 1$, with no sample-path constraint on activation, the index-based policy provides an* optimal solution *to the original problem. One can choose a uniformly-bounded constraint function $f(n, a)$ in such a way that the bandit is indexable. When $W = 0$, the objectives of the original and the relaxed problem are the same. Hence, the index-based policy (activate the bandit whenever it is in a state n with $W(n) > 0$) results in an optimal solution for the original problem for the single bandit. In Section 6 this is discussed for a content delivery network.*

3.2 Expression of Whittle's index

In this section we present an algorithm, which checks for indexability, and, if indexable, calculates Whittle's index. This algorithm is computationally expensive. In section 4, we show how this algorithm simplifies in the case some structural results can be shown for the relaxed problem. Since the action set (binary actions) is finite for each state, we can restrict our focus to deterministic stationary policies, see [Puterman, 2014, Theorem 11.4.8].

Let $X^U \subset \mathcal{N}^d$ be the set consisting of all states bandit k can be in. For a given subset $X \subset X^U$, let policy X refer to the policy that keeps the bandit passive in all states $x \in X$. We denote the empty set by ε.

Algorithm 1 Indexability check and Whittle's index computations

Require: $X^U, f_k(.), T_k(.)$

Initialization: Define X_0 as the union of sets that equals $\arg \inf_{X \subset X^U} \mathbb{E}(f_k(N_k^X, S_k^X(N_k^X)))$. Define $X_{0'}$ as the union of sets that equals $\arg \sup_{X \subset X^U} \mathbb{E}(f_k(N_k^X, S_k^X(N_k^X)))$.

Step j: Compute

$$W_j = \inf_{X \subset X^U} \frac{\mathbb{E}(T_k(N_k^X, S_k^X(N_k^X))) - \mathbb{E}(T_k^{X_{j-1}}(N_k^{X_{j-1}}, S_k^{X_{j-1}}(N_k^{X_{j-1}})))}{\mathbb{E}(f_k(N_k^X, S_k^X(N_k^X))) - \mathbb{E}(f_k(N_k^{X_{j-1}}, S_k^{X_{j-1}}(N_k^{X_{j-1}})))}, j \geq 1.$$

where the infimum is taken over all those $X \subset X^U$ such that the denominator is non-zero.

Step j_1: If the infimum is not attained in any finite set, go to step j'. Otherwise, let $\mathcal{X}_j$ denote the collection of sets that reaches the infimum. Define X_j as the union of all sets $arg \max_{Y \in \mathcal{X}_j} \mathbb{E}(f_k(N_k^Y, S_k^Y(N_k^Y)))$.

Step j_2: If $X_{j-1} \subset X_j$, set $W_k(x) := W_j$ for all states $x \in X_j \backslash X_{j-1}$. Else, the system is not indexable and stop.

Step j_3: If $X_j = X^U$, then the system is indexable and stop. Otherwise go to step $j+1$.

Step j': Compute

$$W_{j'} = \sup_{X \subset X^U} \frac{\mathbb{E}(T_k(N_k^X, S_k^X(N_k^X))) - \mathbb{E}(T_k^{X_{j'-1}}(N_k^{X_{j'-1}}, S_k^{X_{j'-1}}(N_k^{X_{j'-1}})))}{\mathbb{E}(f_k(N_k^X, S_k^X(N_k^X))) - \mathbb{E}(f_k(N_k^{X_{j'-1}}, S_k^{X_{j'-1}}(N_k^{X_{j'-1}})))}, j \geq 1.$$

where the supremum is taken over all those $X \subset X^U$ such that the denominator is non-zero.

Step j'_1: If the supremum is not attained in any finite set, Whittle's index is not found. Otherwise, let $\mathcal{X}_{j'}$ denote the collection of sets that reaches the infimum. Define $X_{j'}$ as the union of all sets $arg \min_{Y \in \mathcal{X}_{j'}} \mathbb{E}(f_k(N_k^Y, S_k^Y(N_k^Y)))$.

Step j'_2: If $X_{j'-1} \subset X_{j'}$, set $W_k(x) := W_{j'}$ for all states $x \in X_{j'} \backslash X_{j'-1}$, Else, the system is not indexable and stop.

Step j'_3: If $X_{j'} = X^U$ then the system is indexable, and stop. Otherwise go to step $j'+1$.

Algorithm 1 can be explained as follows. In the relaxed subproblem, the objective is to find a policy for bandit k that minimizes (6). Note that (6) is a linear function in W. Hence, when comparing two policies ϕ_1 and ϕ_2, defined by their sets of passive states $X^{(\phi_1)}$ and $X^{(\phi_2)}$, respectively, for which the slopes of (6) are not equal, there is a value for the subsidy, $W_{1,2}$, for which the costs (6) under both policies are equal. This is given by

$$W_{1,2} := \frac{\mathbb{E}(T_k(N_k^{X^{(\phi_1)}}, S_k^{X^{(\phi_1)}}(N_k^{X^{(\phi_1)}}))) - \mathbb{E}(T_k(N_k^{X^{(\phi_2)}}, S_k^{(\phi_2)}(N_k^{(\phi_2)})))}{\mathbb{E}(f_k(N_k^{X^{(\phi_1)}}, S_k^{X^{(\phi_1)}}(N_k^{X^{(\phi_1)}}))) - \mathbb{E}(f_k(N_k^{(\phi_2)}, S_k^{(\phi_2)}(N_k^{(\phi_2)})))}.$$

When calculating the indices, one needs to find the policies (and their corresponding passive sets) that minimize (6) for each W. As a first step, one can start by searching for the optimal policy when $W = -\infty$ (the set of passive states corresponding to this optimal policy is denoted by X_0) and when $W = \infty$ (corresponding policy denoted by $X_{0'}$). Recursively, one can find the switching point W where the optimal passive set changes. That is, in Step j, given it is known that the policy described by the passive set X_{j-1} is optimal when $W_{j-2} < W \leq W_{j-1}$, find the crossing

point where another policy becomes optimal. Such a crossing point is denoted by W_j and the corresponding optimal passive set by X_j. Now, if $X_{j-1} \subset X_j$, for all j, this implies indexability and the index for any state in $X_j \backslash X_{j-1}$ is given by W_j. A mathematical proof of Algorithm 1 would follow according to similar steps as in Glazebrook et al. [2009], where an algorithm is developed for finding Whittle's index in the context of admission control and routing of impatient customers.

Remark 3.5. *In order to numerically run this algorithm one would need X^U to have finite cardinality, in order to find the infimum/supremum in step j and j'. Hence, when $|X^U| = \infty$, numerical computation through the algorithm would only provide an approximation for Whittle's index, since one would need to limit the search in the inf/sup over a finite number of subsets X.*

If instead $|X^U| < \infty$, the algorithm provides Whittle's index, if indexable. The computational complexity of the algorithm is $\mathcal{O}\left(|X^U|^4 2^{|X^U|}\right)$. This can be seen by noting that (i) the number of possible subsets $X \subset X^U$ is $2^{|X^U|}$, (ii) the computation of the steady-state distribution is $\mathcal{O}\left(\frac{|X^U|^3}{3}\right)$ (using Gauss elimination, see Bolch et al. [2006]), (iii) finding the infimum or supremum requires $2^{|X^U|}$ comparisons, and (iv) the number of steps in the algorithm is of the order $\mathcal{O}(|X^U|)$.

Remark 3.6. *Since we consider the average cost criterion, the calculations in step j and j' in Algorithm 1 are feasible provided the steady-state distributions are known. If instead we had considered the total discounted cost criterion, the latter would depend on the initial state and transient behavior, and hence the calculation as done in Algorithm 1 is not possible.*

Remark 3.7. *In Algorithm 1, an indexability check is included. Note that in case indexability had been proved independently, the algorithm would simplify, as one can replace in step j the "$\inf_{X \subset X_U}$" by "$\inf_{X: X_{j-1} \subset X \subset X_U}$", and similarly in step j'.*
In addition, in Algorithm 1 structural properties of the optimal solution of the relaxed optimization problem are not taken into account. In Section 4 we will describe how a structural property (threshold optimality) can help in the calculation of the index.

4 Threshold policies

For certain one-dimensional problems, it can be established that the structure of the optimal solution of problem (6) is of threshold type. That is, there is a threshold function $n_k(W)$ such that when bandit k is in a state $m_k \leq n_k(W)$, then action a is optimal, and otherwise action a' is optimal, $a,\ a' \in \{0,1\}$ and $a \neq a'$. We let policy $\phi = n$ denote a threshold policy with threshold n, and we refer to it as 0-1 type if $a = 0$ and $a' = 1$, and 1-0 type if $a = 1$ and $a' = 0$. In general it can be hard to verify whether an optimal solution is of threshold type. In this section we provide sufficient conditions for threshold optimality and show how Algorithm 1 simplifies.

The following result characterizes sufficient conditions on the transition rates $q_k^a(\cdot,\cdot)$ and the jump probabilities $p_k^a(\cdot,\cdot)$ such that a threshold policy solves problem (6). For example, condition (i) can be interpreted as follows. If there are no upward jumps under the active action or under an impulse control, and there can be an upward jump of at most one under the passive action, then 0-1 type of threshold policies are optimal. The proof can be found in the Appendix.

Proposition 1. *Assume $N_k(t) \in \mathcal{N}$ If one of the following conditions holds,*

(i) $q_k^1(N, N+i) = 0,\ \forall\ i \geq 1,\ \ q_k^0(N, N+i) = 0,\ \forall\ i \geq 2,$ *and* $p_k^a(N, N+i) = 0,\ \forall\ i \geq 1, a = 0, 1,$

(ii) $q_k^0(N, N-i) = 0,\ \forall\ i \geq 1,\ \ q_k^1(N, N-i) = 0,\ \forall\ i \geq 2,$ *and* $p_k^a(N, N-i) = 0,\ \forall\ i \geq 1, a = 0, 1,$

then there exists an $n_k \in \{-1, 0, 1, ...\}$ such that a 0-1 type of threshold policy with threshold n_k, optimally solves problem (6).

Alternatively, if one of the following conditions holds,

(iii) $q_k^1(N, N-i) = 0,\ \forall\ i \geq 1,\ \ q_k^0(N, N-i) = 0,\ \forall\ i \geq 2$ *and* $p_k^a(N, N-i) = 0,\ \forall\ i \geq 1, a = 0, 1,$

(iv) $q_k^0(N, N+i) = 0,\ \forall\ i \geq 1,\ \ q_k^1(N, N+i) = 0,\ \forall\ i \geq 2,$ *and* $p_k^a(N, N+i) = 0,\ \forall\ i \geq 1, a = 0, 1,$

then, there exists an $n_k \in \{-1, 0, 1, ...\}$ such that a 1-0 type of threshold policy with threshold n_k, optimally solves problem (6).

Optimality of a threshold policy for the relaxed optimization problem has been proved for several RMABPs, several examples can be found in [Gittins et al., 2011, Section 6.5]. All applications as presented in this paper, fit the sufficient conditions, and hence have the threshold structure. We further note that there are models where the optimality of threshold policies has been established, but whose model parameters do not fall within the conditions of Proposition 1, see for example Ansell et al. [2003], Glazebrook et al. [2009], Larrañaga et al. [2015].

When an optimal solution for problem (6) is of threshold type, we have the following sufficient condition for indexability. For the proof we refer to the Appendix.

Proposition 2. *If an optimal solution of (6) is of threshold type, and* $\mathbb{E}(f_k(N_k^n, S_k^n(N_k^n)))$ *is non-negative and strictly increasing in* n*, then problem (6) is indexable.*

Remark 4.1 (Algorithm 1 with threshold-based policies.)**.** *Algorithm 1 simplifies in case it is known beforehand that* (i) *threshold policies are optimal for the relaxed problem, and* (ii) *the bandit is indexable. If it is known that 0-1 type of threshold policies are optimal and that the problem is indexable, from Remark 3.7, one can restrict the search in Step* j *and Step* j' *to sets* X *of the form* $\{m \leq n\}$ *where* $n > n_{j-1}$*, and initialization step reduces to finding* $n_0 = \arg\inf_{n\in\mathbb{N}} \mathbb{E}(f_k(N_k^n, S_k^n(N_k^n)))$ *and* $X_0 = \{m \leq n_0\}$*. Similar for* $X_{0'}$*.*

When threshold policies are optimal, the computational complexity of Algorithm 1 reduces from $\mathcal{O}\left(|X^U|^4 2^{|X^U|}\right)$ *to* $\mathcal{O}\left(|X^U|^5\right)$ *(the latter is the computational complexity of Algorithm 1 as obtained in Remark 3.5). This follows since the infimum/supremum is taken over sets of the form* $\{m \leq n\}$*, of which there are* $|X^U|$*, as opposed to* $2^{|X^U|}$ *subsets in the general setting.*

In case the monotonic nature of the function

$$\frac{\mathbb{E}(T_k^n(N_k^n, S_k^n(N_k^n))) - \mathbb{E}(T_k^{n-1}(N_k^{n-1}, S_k^{n-1}(N_k^{n-1})))}{\mathbb{E}(f_k(N_k^n, S_k^n(N_k^n))) - \mathbb{E}(f_k(N_k^{n-1}, S_k^{n-1}(N_k^{n-1})))} \tag{7}$$

can be proven, the computation of Whittle's index further simplifies:

Proposition 3. *Assume an optimal solution of (6) is of threshold type and bandit* k *is indexable. If (7) is a monotone function in* n*, then Whittle's index* $W_k(n)$ *is given by (7).*

Further, it follows that Whittle's index is non-decreasing if 0-1 type of threshold policies are optimal and is non-increasing if 1-0 type of threshold policies are optimal.

To illustrate our unifying framework, in the next two sections, we briefly present a few specific applications when the evolution of the bandit is driven by a Markov chain with finite and infinite transition rates. Full details are given in the Appendices.

5 Applications: finite transition rates

There exists a large set of papers that calculate the Whittle's index for RMABPs with finite transition rates and time-average performance objective, see Section 1.1 for further references. We consider one such application (machine repairman problem) to illustrate the applicability of our results. We however emphasize that all indices as derived in previous literature, can be derived directly from our unifying framework.

Machine repairman problem

We consider the classical machine repairman problem with $\mathcal{M}$ non-identical machines and $\mathcal{R}$ repairmen, where $\mathcal{R} \leq \mathcal{M}$. Any number of repairmen may be active at any decision epoch, but no more than one may work on any individual machine at a time.

Glazebrook et al. [2005] modeled the machine repairman problem as an RMABP in *discrete time* and obtained Whittle's index for the average-cost criterion by first considering the discounted-cost criterion and then letting the discounting factor tend to one. In this section, we describe the machine repairman model as an RMABP in *continuous time*, and allow state dependent transition rates with a general cost function. We use Proposition 3 to derive Whittle's index for the average-cost criterion, and as special cases retrieve the indices of Glazebrook et al. [2005]. This generalizes the model in Glazebrook et al. [2005], where the analysis was restricted to constant cost structure (for model 2) and state-independent transition rates.

The state of the system at epoch t is an $\mathcal{M}$-dimensional vector $X(t) = \{X_1(t), X_2(t), ..., X_{\mathcal{M}}(t)\}$, where $X_k(t)$ is the state of machine $k \leq \mathcal{M}$, and $X_k(t) \in \{0, 1, \ldots\}$. The machine state measures the degree of deterioration and is

assumed to evolve independently of the states of the other machines. When one of the $\mathcal{R}$ repairmen is working on machine k, it makes a transition from state n_k to the pristine state 0 at repair rate $r_k(n_k)$. If instead machine k is unattended, its state deteriorates from state n_k to state n_k+1 after an exponential amount of time with deterioration rate $\lambda_k(n_k)$. In addition, if machine k in state n_k is unattended, it experiences a catastrophic breakdown at rate $\psi_k(n_k)$, after which the machine is replaced by a new machine at a considerable lump-sum cost $L_k^b(n_k)$.

Let $L_k^r(n_k)$ be the lump-sum cost of using the repairman in state n_k, where typically $L_k^b(n_k) >> L_k^r(n_k)$. Let $C_k^d(n_k)$ denote the per unit cost of deterioration for an unattended machine k in state n_k. The objective is to deploy the repairmen in order to minimize the total long-run average cost.

This problem can be cast in the RMABP framework as follows. Each machine is a bandit. At each decision epoch, at most $\mathcal{R}$ bandits/machines can be activated (repairman is sent to it). A bandit is active ($a=1$) if a repairman is deployed to this machine, and passive ($a=0$) otherwise. The machine repairman problem is hence characterized by the following transition rates:

$$q_k^1(n_k,0)=r_k(n_k),\ q_k^0(n_k,0)=\psi_k(n_k) \text{ and } q_k^0(n_k,n_k+1)=\lambda_k(n_k), \tag{8}$$

and the cost functions under action $a=0$ and $a=1$ are given by,

$$\begin{aligned} C_k(n_k,0)&=\psi_k(n_k)L_k^b(n_k)+C_k^d(n_k),\\ C_k(n_k,1)&=r_k(n_k)L_k^r(n_k). \end{aligned}$$

We set $f_k(N_k^\phi,S_k^\phi(\vec{N}))=S_k^\phi(\vec{N})$, so that the constraint $\sum_{k=1}^{\mathcal{M}} f_k(N_k^\phi,S_k^\phi(\vec{N}))=\sum_{k=1}^{\mathcal{M}} S_k^\phi(\vec{N})\le\mathcal{R}$, ensures that at most $\mathcal{R}$ repairmen are active at a time.

The transition rates (8) satisfy the sufficient conditions as presented in Proposition 1, hence an optimal policy of the relaxed optimization problem (5) is of threshold type with 0-1 structure. Further, $\mathbb{E}(f_k(N_k^{n_k},S_k^{n_k}(N_k^{n_k})))=\sum_{m=0}^{n_k}\pi_k^{n_k}(m)$, with $\pi_k^{n_k}(m)$ as the stationary probability of being in state m under threshold n_k. In addition, it can be verified that $\sum_{m=0}^{n_k}\pi_k^{n_k}(m)$ is strictly increasing in n_k if $r_k(n_k)\le r_k(n_k+1)$ for all n_k (see Appendix B.2). Thus, the indexability follows from Proposition 2. Together with Proposition 3, we obtain the following closed-form expression for Whittle's index. We refer to Appendix B.3 for the proof.

Proposition 4. *Assume $r_k(n)\le r_k(n+1),\ \forall\ n$ and $r_k(1)>0$. It holds that bandit k is indexable. Consider*

$$\frac{(C_{Sum}(n)+L_k^r(n+1)P_n)\left(P_{Sum}(n-1)+\frac{P_{n-1}}{r_k(n)}\right)-(C_{Sum}(n-1)+L_k^r(n)P_{n-1})\left(P_{Sum}(n)+\frac{P_n}{r_k(n+1)}\right)}{\frac{P_{n-1}}{r_k(n)}\sum_{i=0}^{n}\frac{P_i}{\lambda_k(i)}-\frac{P_n}{r_k(n+1)}\sum_{i=0}^{n-1}\frac{P_i}{\lambda_k(i)}}, \tag{9}$$

where $P_0:=1, P_i:=\prod_{j=1}^{i}p_k(j)$, $p_k(j):=\frac{\lambda_k(j)}{\lambda_k(j)+\psi_k(j)}$, $P_{Sum}(n):=\sum_{i=0}^{n}\frac{P_i}{\lambda_k(i)}$ and $C_{Sum}(n):=\sum_{i=1}^{n}\left[(P_{i-1}-P_i)L_k^b(i)+\frac{P_iC_k^d(i)}{\lambda_k(i)}\right]$. If (9) is monotone in n, then Whittle's index is given by (9).

For the parameter settings such that (9) is monotone, the Whittle's index is given by (9), for the others, the index can be computed by Algorithm 1.

For specific choices of the parameters, Whittle's indices were previously obtained in Glazebrook et al. [2005]. We now show how they follow directly from (9). In the first model considered in Glazebrook et al. [2005], there are no catastrophic breakdowns, only deterioration costs are considered, repair rates and repair cost are state-independent. That is, $r_k(n)=r_k$, $\psi_k(n)=0$, $L_k^b(n)=0$, $L_k^r(n)=L_k^r$, $\forall\ n$. In that case, (9) reduces to

$$W_k(n)=r_k\left[\sum_{i=0}^{n-1}\frac{C_k^d(n)-C_k^d(i)}{\lambda_k(i)}+\frac{C_k^d(n)-r_kL_k^r}{r_k}\right], \tag{10}$$

which is indeed monotone if $C_k^d(n)$ is an increasing sequence in n (see Appendix B.4 for details).

Recall that we consider a continuous-time model, while Glazebrook et al. [2005] consider discrete time set-up. In particular, in the discrete-time model, a machine is repaired in one time slot. Setting $r_k=1$ (so that in our model the expected time to repair a machine equals 1) and interpreting $1/\lambda_k(n)$ as the expected time a machine (not under repair) spends in state n, Equation (10) matches the index as derived in [Glazebrook et al., 2005, Corollary 1].

In the second model considered in Glazebrook et al. [2005], there is no deterioration cost, but there is a lump-sum cost for catastrophic breakdowns. All lump-sum costs and repair rates are state independent. That is, $r_k(n) = r_k$, $C_k^d(n) = 0$, $L_k^r(n) = R_k, L_k^b(n) = B_k$, $\forall\ n$. Now, (9) reduces to

$$W_k(n) = \frac{B_k\left(\frac{1-p_k(n)}{r_k} - \frac{p_k(n)}{\lambda_k(n)} + \sum_{i=0}^{n}\frac{P_i}{\lambda_k(i)} - p_k(n)\sum_{i=0}^{n-1}\frac{P_i}{\lambda_k(i)}\right)}{\frac{1}{r_k}\left(\sum_{i=0}^{n}\frac{P_i}{\lambda_k(i)} - p_k(n)\sum_{i=0}^{n-1}\frac{P_i}{\lambda_k(i)}\right)} - R_k, \tag{11}$$

which is indeed monotone if $\psi_k(n)$ is an increasing sequence in n (see Appendix B.5 for details).

In the discrete-time model of Glazebrook et al. [2005], it is assumed that the time it takes to change state is equal to 1. Setting $\frac{1}{r_k} = 1$ and $\frac{1}{\lambda_k(n)+\psi_k(n)} = 1\ \forall\ n$, (so that in the continuous-time model the expected time to change state equals 1) we retrieve

$$W_k(n) = \frac{B_k\left(\hat{T}(n) - p(n)\hat{T}(n-1) - p(n)\right)}{\left(\hat{T}(n) - p(n)\hat{T}(n-1)\right)} - R_k, \tag{12}$$

where $\hat{T}(n) := \sum_{i=0}^{n}\frac{P_i}{\lambda_k(i)}$ can be interpreted as the expected duration to either reach state n or that a catastrophic breakdown occurs, when starting in state 0. As such, we retrieve the Whittle's index as obtained in [Glazebrook et al., 2005, Corollary 2] for the discrete-time case.

6 Applications: infinite transition rates

To the best of our knowledge, the calculation for Whittle's index has not been earlier explored for impulse control, that is, infinite transition rates, under the average cost criteria. In this section, we illustrate the applicability of our expression for Whittle's index in problems with impulsive control by considering examples in the areas of congestion control in Transmission Control Protocol (TCP) and scheduling in content delivery network.

6.1 Congestion control of TCP flows

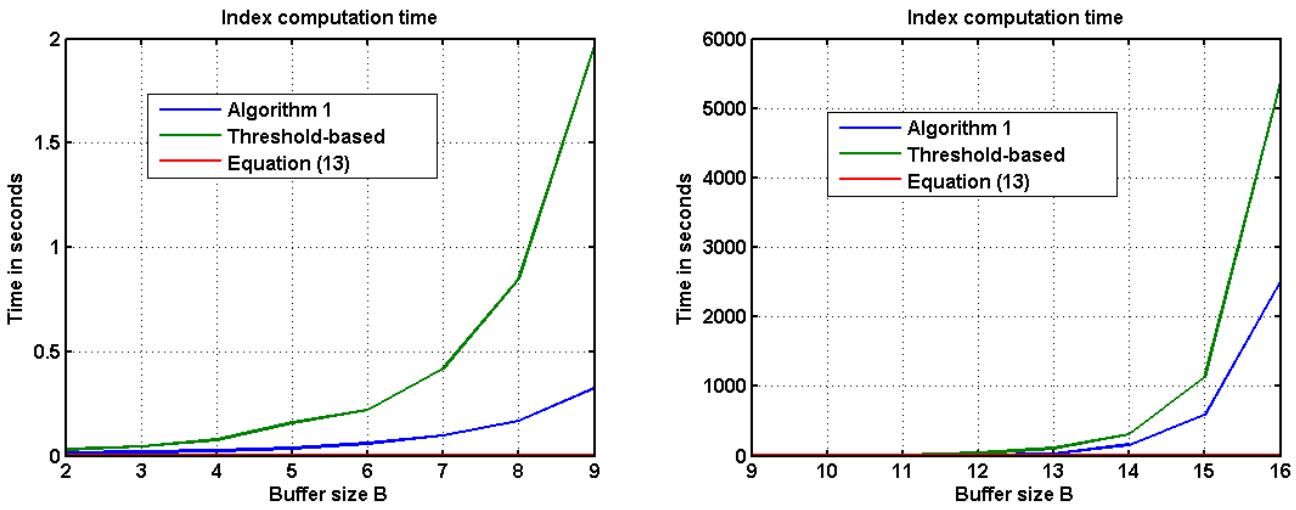
We consider the following congestion control model for a simple version of TCP, which is the algorithm that regulates the congestion on the Internet. There are K flows trying to deliver packets to their destination via a bottleneck router with buffer size B. We assume that flow k has implemented an additive-increase/multiplicative-decrease (AIMD) mechanism as in TCP. The congestion window of each flow is adapted according to received acknowledgements. For each positive acknowledgement (ACK), the congestion window is increased by the reciprocal of its current value, which approximately corresponds to an increase by one packet during a round-trip-time (RTT) without lost packets. We model this with Poisson arrivals of packets for flow k with arrival rate λ_k, which correspond to the time period of one RTT. For each negative acknowledgement (NACK), the congestion window is *immediately* decreased using the formula, congestion window = $\max\{\lfloor\gamma_k \times \text{congestion window}\rfloor, 1\}$, where $0 \leq \gamma_k < 1$ is the multiplicative decrease factor and the function $\lfloor . \rfloor$ denotes the floor function. Whenever the bottleneck router is full, a controller decides which flow will receive a NACK in order to maximize the average reward. We take reward as the generalized α-fairness function $R_k^{(\alpha)}(n)$, which is a function of n, the number of outstanding packets of flow k:

$$R_k^{(\alpha)}(n) = \begin{cases} \frac{(1+n)^{(1-\alpha)}-1}{1-\alpha} & \text{if } \alpha \neq 1, \\ \log(n+1) & \text{if } \alpha = 1. \end{cases}$$

The above reward is earned if the controller admits the packet to the router ($a = 1$ or ACK), and 0 otherwise ($a = 0$ or NACK). The parameterized family of the above generalized $\alpha-$fair reward aims at maximizing the router's total aggregated utility. Note that the parameter α permits to recover a wide variety of utilities such as max-min, maximum throughput and proportional fairness (see Mo and Walrand [2000], Altman et al. [2008]).

We model the above scenario as an RMABP where each flow represents a bandit and the state of the bandit k, n_k, represents the number of outstanding packets. The parameters of bandit k are:

$$\begin{aligned} &\mathcal{I}_k(n_k, 0) = 1 \quad \text{and } p_k^0(n_k, \max\{\lfloor\gamma_k.n_k\rfloor, 1\}) = 1, \\ &\mathcal{I}_k(n_k, 1) = 0 \quad \text{and } q_k^1(n_k, n_k+1) = \lambda_k, \end{aligned}$$



a) b)

Figure 1: Comparison of the time required to calculate the index by Algorithm 1 (all possible subsets), Algorithm 1 (threshold based sets) and , Equation (13)

where action $a = 1$ ($a = 0$) stands for sending ACK (NACK). We set $f_k(N_k^\phi, S_k^\phi(\vec{N})) = N_k^\phi S_k^\phi(\vec{N})$, which ensures that all flows that receive an ACK can send their outstanding packets through the bottleneck router with buffer size B, $\sum_{k=1}^{K} f_k(N_k^\phi, S_k^\phi(\vec{N})) \leq B$. Setting $\mathcal{I}_k(n_k, 0) = 1$ ensures that the congestion window of flow k instantaneously changes when it receives a NACK. We further set $C_k(n, 0) = 0$ and $C_k(n, 1) = -R_k^{(\alpha)}(n)$.

It follows from Proposition 1 that an optimal policy of the relaxed optimization problem (6) is of threshold type with 1-0 structure. It can be verified that $\mathbb{E}(f_k(N_k^{n_k}, S_k^{n_k}(N_k^{n_k}))) = \sum_{m=0}^{n_k} m\pi_k^{n_k}(m)$ is strictly increasing in n_k (see Appendix C.1 for details), so that indexability follows from Proposition 2. Together with Proposition 3, the following result characterizes a closed form expression for Whittle's index (see Appendix C.2 for details).

Lemma 6.1. *Consider*

$$\begin{cases} \dfrac{2\lambda_k\left(\sum_{m=S}^{n-1}(1+m)^{1-\alpha}-(n-S)(1+n)^{1-\alpha}\right)}{(n-S)(n-S+1)(1-\alpha)} & \text{if } \alpha \neq 1, \\ \dfrac{2\lambda_k\left(\sum_{m=S}^{n-1}\log(1+m)-(n-S)\log(1+n)\right)}{(n-S)(n-S+1)} & \text{if } \alpha = 1, \end{cases} \tag{13}$$

with $S = \max\{\lfloor \gamma_k.(n+1) \rfloor, 1\}$. *If* (13) *is monotone in* n, *then Whittle's index is given by* (13).

We could not prove the monotonic nature of (13). However, we observe numerically that this indeed is the case for a wide set of parameter settings. If monotonicity of Whittle's index cannot be established, the index can be computed using Algorithm 1.

The index-based policy (Definition 3.3) would be as follows: activate/ACK all flows with highest Whittle's indices until the buffer is filled. All other flows receive a NACK. Since a NACK implies a multiplicative decrease, it is obvious that if the buffer is full, under the index policy only the flow with the smallest index value (13) will receive a NACK.

The paper Avrachenkov et al. [2013] studied as well the congestion control problem as a multi-armed restless bandit framework. This work however assumes discrete time, and as such, no infinite transition rates could be considered. In addition, for their analysis, they require a bound on the number of outstanding packets per flow. Avrachenkov et al. [2013] numerically verify for indexability and obtain Whittle's indices in closed form only in the case of at most three outstanding packets per flow. Our approach considerably simplifies the analysis, allows to prove indexability and provides an analytical expression for Whittle's index.

In Figure 1 we compare the time required to calculate the Whittle index of a TCP flow on a desktop. Namely, we compare the time required by running (*i*) Algorithm 1 over all possible subsets (see Remark 3.5), (*ii*) Algorithm 1 over threshold-based subsets (see Remark 4.1), and (*iii*) calculating Equation (13). The curves clearly illustrate that running Algorithm 1 (even in the case of threshold based subsets) quickly becomes computationally demanding, even for problems of relatively small size.

6.2 Content delivery network

In this section, we adopt the basic model of a content delivery network, as explored in Larrañaga et al. [2015]. Our primary focus is to illustrate the optimality of Whittle's index policy for a single-armed bandit (Remark 3.4) with impulse control.

In a content delivery network, the bulk of traffic (e.g. software updates, video content etc.) is delay tolerant. Hence, requests can be delayed and grouped, so as to be transmitted in a multi-cast mode through the network. The challenge is to balance the *gains* of grouping requests in order to save transmission capacity against the *risk* of not meeting the deadline of one or more jobs. We assume that all the jobs instantaneously depart upon receiving the service.

Let the state n denote the number of waiting requests. Let jobs arrive according to a Poisson process with a state dependent rate $\lambda(n)$. For each request, after a state dependent exponential expiration time with rate $\theta(n)/n$, the request abandons the system (deadline is passed). Upon activation of the server, all waiting requests are cleared instantaneously. The objective is to minimize the long-run average cost incurred by the waiting jobs, by abandonments, as well as the set-up cost paid upon activation of the server.

Let $C^h(n)$ be the state-dependent cost per unit of time that n requests are held in the queue. Let $L^a(n)$ be the state-dependent penalty (lump-sum cost) for a job abandoning the queue, which may depend on the action a chosen. Let $L_s^\infty(n)$ be the set-up cost (lump-sum cost) of clearing the batch of size n.

We can model this as a single armed restless bandit, where action $a = 1$ ($a = 0$) stands for serving (not serving) all the waiting requests. We drop the subscript k since there is only one bandit. We have the following transitions:

$$\begin{aligned} &\mathcal{I}(n,0) = 0 \quad \text{and } q^0(n,n+1) = \lambda(n),\ q^0(n,n-1) = \theta(n), \\ &\mathcal{I}(n,1) = 1 \quad \text{and } p^1(n,0) = 1. \end{aligned}$$

We further set $f(N^\phi, S^\phi(\vec{N})) = -\mathbf{1}_{\{N^\phi \in \mathcal{M}^\phi\}}$, where $\mathcal{M}^\phi = \{m \in \{0,1,2,...\} : S^\phi(m) = 0, S^\phi(m+1) = 1\}$, in order to find an *optimal* activation policy (see Remark 3.4). Note that we need the function $f(.)$ to depend on policy ϕ, different to what was assumed in Section 2. We did so in order to find a constraint function f for which it holds that $\mathbb{E}(f(N^n, S^n(N^n)))$ is strictly increasing. Taking for example $f(n,a) = 1 - a$, would result in $\mathbb{E}(f(N^n, S^n(N^n))) = 1$, since when looking to the time-average, one is always passive. It can be checked that all results of this paper (and their proofs) hold true for this given choice of function $f(N^\phi, S^\phi(\vec{N})) = -\mathbf{1}_{\{N^\phi \in \mathcal{M}^\phi\}}$.

An optimal policy of the relaxed optimization problem (6) is of threshold type with 0-1 structure. The latter follows directly from Proposition 1. In addition, it can be verified that, if $\lambda(n)$ is non-decreasing, then $E(f(N^n, S^n(\vec{N}))) = -\pi^n(n)$ is strictly increasing in n, see Appendix D. Thus, the indexability follows from Proposition 2. From Proposition 3, we obtain an analytical expression for Whittle's index (see Appendix D for details).

Lemma 6.2. *Assume $\lambda(n)$ is non-decreasing. Consider*

$$\frac{\sum_{i=1}^{n-1} \tilde{C}(i)(\pi^n(i) - \pi^{n-1}(i)) + \tilde{C}(n)\pi^n(n) + L_s^\infty(n+1)\lambda(n)\pi^n(n) - L_s^\infty(n)\lambda(n-1)\pi^{n-1}(n-1)}{\pi^{n-1}(n-1) - \pi^n(n)}, \tag{14}$$

where $\tilde{C}(n) = nC^h(n) + \theta(n)L^a(n)$. If (14) is non-decreasing in n, then Whittle's index is given by (14).

We conclude that an optimal policy in the content delivery problem is to activate the bandit/queue whenever the numerator of (14) is non-negative. This follows directly from Remark 3.4.

If the rates and costs are state-independent, i.e., $\lambda(n) = \lambda$, $\theta(n) = i\theta$, $C^h(n) = C^h$, $L^a(n) = L^a$ and $L_s^\infty(n) = L_s^\infty$, $\forall\, n$, the optimal policy simplifies to activating whenever $\tilde{C}\left(\mathbb{E}(N^n) - \mathbb{E}(N^{n-1})\right) - \lambda L_s^\infty(\pi^{n-1}(n-1) - \pi^n(n)) \geq 0$, with $\tilde{C} = C^h + \theta L^a$. This coincides with the results obtained in [Larrañaga et al., 2015, Proposition 3].

7 Conclusions

In the main contribution of this work, we derive an analytical expression for Whittle's index under the average-cost criterion when each bandit evolves as a continuous-time Markov chain with possible impulse control. The Whittle's index is given in a compact expression, which makes the implementation of the index heuristic easier. We show that with this general formula, we can retrieve Whittle's index for many application, that were previously derived in the literature on a case-by-case basis.

Acknowledgement

We would like to thank Zhang Yi and Alexey Piunovskiy for helpful discussions on optimal impulse control.

This research is partially supported by the French Agence Nationale de la Recherche (ANR) through the project ANR-15-CE25-0004 (ANR JCJC RACON) and by ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02.

References

Abderrahmane Abbou and Viliam Makis. Group maintenance: A restless bandits approach. *INFORMS Journal on Computing*, 2019. doi: 10.1287/ijoc.2018.0863.

Eitan Altman, Konstantin Avrachenkov, and Andrey Garnaev. Generalized α-fair resource allocation in wireless networks. In *2008 47th IEEE Conference on Decision and Control*, pages 2414–2419. IEEE, 2008.

PS Ansell, Kevin D Glazebrook, José Niño-Mora, and M O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57(1):21–39, 2003.

Nilay Tanik Argon, Li Ding, Kevin D Glazebrook, and Serhan Ziya. Dynamic routing of customers with general delay costs in a multiserver queuing system. *Probability in the Engineering and Informational Sciences*, 23(2):175–203, 2009.

K.E. Avrachenkov, A. Piunovskiy, and Y. Zhang. Impulsive control for g-aimd dynamics with relaxed and hard constraints. In *Proceedings of IEEE CDC*, pages 880–887, 2018.

Konstantin Avrachenkov, Urtzi Ayesta, Josu Doncel, and Peter Jacko. Congestion control of TCP flows in internet routers by means of index policy. *Computer Networks*, 57(17):3463–3478, 2013.

Turgay Ayer, Can Zhang, Anthony Bonifonte, Anne C Spaulding, and Jagpreet Chhatwal. Prioritizing Hepatitis C treatment in US prisons. *Operations Research*, 2019.

G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi. *Queueing networks and markov chains*. Wiley, 2006.

Vivek S Borkar and Sarath Pattathil. Whittle indexability in egalitarian processor sharing systems. *Annals of Operations Research*, pages 1–21, 2017.

Vivek S Borkar, Gaurav S Kasbekar, Sarath Pattathil, and Priyesh Shetty. Opportunistic scheduling as restless bandits. *IEEE Transactions on Control of Network Systems*, 2017a.

Vivek S Borkar, K Ravikumar, and Krishnakant Saboo. An index policy for dynamic pricing in cloud computing under price commitments. *Applicationes Mathematicae*, 44:215–245, 2017b.

F. Dufour and A. B. Piunovskiy. Impulsive control for continuous-time markov decision processes. 47(1):106–127, 2015.

F. Dufour and A. B. Piunovskiy. Impulsive control for continuous-time markov decision processes: A linear programming approach. *Applied Mathematics & Optimization*, 74(1):129–161, Aug 2016. ISSN 1432-0606. doi: 10.1007/s00245-015-9310-8. URL https://doi.org/10.1007/s00245-015-9310-8.

Stephen Ford, Michael P. Atkinson, Kevin Glazebrook, and Peter Jacko. On the dynamic allocation of assets subject to failure. *European Journal of Operational Research*, 2019. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2019.12.018. URL http://www.sciencedirect.com/science/article/pii/S0377221719310379.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

Kevin D Glazebrook, HM Mitchell, and PS Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research*, 165(1):267–284, 2005.

Kevin D Glazebrook, Christopher Kirkbride, and Jamal Ouenniche. Index policies for the admission control and routing of impatient customers to heterogeneous service stations. *Operations Research*, 57(4):975–989, 2009.

Darina Graczová and Peter Jacko. Generalized restless bandits and the knapsack problem for perishable inventories. *Operations Research*, 62(3):696–711, 2014.

Peter Jacko. *Dynamic priority allocation in restless bandit models.* Lambert Academic Publishing Saarbrücken, Germany, 2010.

Peter Jacko. Resource capacity allocation to stochastic dynamic competitors: knapsack problem for perishable items and index-knapsack heuristic. *Annals of Operations Research*, 241(1-2):83–107, 2016.

Terry James, Kevin Glazebrook, and Kyle Lin. Developing effective service policies for multiclass queues with abandonment: asymptotic optimality and approximate policy improvement. *INFORMS Journal on Computing*, 28(2): 251–264, 2016.

Maialen Larrañaga, Onno J Boxma, Rudesindo Núñez-Queija, and Mark S Squillante. Efficient content delivery in the presence of impatient jobs. In *Teletraffic Congress (ITC 27), 2015 27th International*, pages 73–81. IEEE, 2015.

Maialen Larrañaga, Urtzi Ayesta, and Ina Maria Verloop. Dynamic control of birth-and-death restless bandits: application to resource-allocation problems. *IEEE/ACM Transactions on Networking*, 24(6):3812–3825, 2016.

Jeonghoon Mo and Jean Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on networking*, (5):556–567, 2000.

José Nino-Mora. Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. *Mathematical Programming*, 93(3):361–413, 2002.

José Niño-Mora. Restless bandit marginal productivity indices, diminishing returns, and optimal control of make-to-order/make-to-stock M/G/1 queues. *Mathematics of Operations Research*, 31(1):50–84, 2006.

José Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *Top*, 15(2):161–198, 2007.

Michelle Opp, Kevin Glazebrook, and Vidyadhar G Kulkarni. Outsourcing warranty repairs: Dynamic allocation. *Naval Research Logistics (NRL)*, 52(5):381–398, 2005.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

Sarath Pattathil, Vivek S Borkar, and Gaurav S Kasbekar. Distributed server allocation for content delivery networks. *arXiv preprint arXiv:1710.11471*, 2017.

Alexey Piunovskiy and Yi Zhang. On reducing a constrained gradual-impulsive control problem for a jump markov model to a model with gradual control only. *SIAM Journal on Control and Optimization*, 58(1):192–214, 2020.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

D Ruiz-Hernandez. *Indexable restless bandits: Index policies for some families of stochastic scheduling and dynamic allocation problems.* VDM Verlag, 2008.

Ina Maria Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *The Annals of Applied Probability*, 26(4):1947–1995, 2016.

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

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A): 287–298, 1988.

A Proof of Propositions

In this section we provide the proofs of different propositions. For ease of notation, we removed the subscript k from all the proofs in this section.

A.1 Proof of Proposition 1

Proof. Since $\mathcal{U}_{REL}$ is non-empty, there exists a stationary optimal policy ϕ^* that optimally solves the subproblem (6) for a bandit. Define $n^* = \min\{m \in \{0,1,...\} : S^{\phi^*}(m) = 1\}$. This implies $S^{\phi^*}(m) = 0 \ \forall \ m < n^*$ and $S^{\phi^*}(n^*) = 1$. From the structure on the transition rates and jump probabilities in (i) of Proposition 1, we have $q_k^1(N, N+i) = 0, \ \forall \ i \geq 1$, $q_k^0(N, N+i) = 0, \ \forall \ i \geq 2$, and $p_k^a(N, N+i) = 0, \ \forall \ i \geq 1, a = 0, 1$. The above transition structure ensures that all the states $m > n^*$ are transient. Hence $\pi^{\phi^*}(m) = 0 \ \forall \ m > n^*$. Thus, the following holds under the optimal policy ϕ^*:

$$\begin{aligned}
\mathbb{E}(C(N^{\phi^*}, S^{\phi}(\vec{N}^{\phi^*}))) &= \sum_{m=0}^{n^*-1} C(m,0)\pi^{\phi^*}(m) + C(n^*,1)\pi^{\phi^*}(n^*), \\
\mathbb{E}(f(N^{\phi^*}, S^{\phi}(\vec{N}^{\phi^*}))) &= \sum_{m=0}^{n^*-1} f(m,0)\pi^{\phi^*}(m) + f(n^*,1)\pi^{\phi^*}(n^*),
\end{aligned}$$

and lump-sum cost under the optimal policy ϕ^* is given by:

$$\begin{aligned}
\mathbb{E}(C^{\infty,\phi^*}(N^{\phi^*}, S^{\phi^*}(\vec{N}^{\phi^*})) = \quad & \sum_{\tilde{n}} \sum_{m} \mathbb{E}\left(q^{S^{\phi^*}(\vec{N}^{\phi^*})}(N^{\phi^*}, \tilde{n}) \times \mathcal{I}(\tilde{n}, S^{\phi^*}(\vec{M}^{\phi^*}(\vec{N}^{\phi^*}, \tilde{n}))) \right. \\
& \left. \times p^{S^{\phi^*}(\vec{M}^{\phi^*}(\vec{N}^{\phi^*}, \tilde{n}))}(\tilde{n}, m) \times L^{\infty}(\tilde{n}, m, S^{\phi^*}(\vec{M}^{\phi^*}(\vec{N}^{\phi^*}, \tilde{n}))) \right),
\end{aligned}$$

From Markov chain theory, the average number of times state y is visited in the next decision epoch under action a given the current state x can be written as:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} 1^a_{\{X_n = x, \ X_{n+1} = y\}} = \pi(x) q^a(x,y).$$

Given that the parameters satisfy (i) of Proposition 1, the lump-sum cost can equivalently be written as;

$$\begin{aligned}
\mathbb{E}(C^{\infty,\phi^*}(N^{\phi^*}, S^{\phi^*}(\vec{N}^{\phi^*})) = & \sum_{n=0}^{n^*-1} \sum_{m=0}^{n} \left(p^0(n,m) L^{\infty}(n,m,0) \left[\sum_{k=0, k \neq n}^{n^*-1} q^0(k,n)\pi^{\phi^*}(k) + q^1(n^*,n)\pi^{\phi^*}(n^*) \right] \right) \\
& + \sum_{m=0}^{n^*} \left(p^1(n^*,m) L^{\infty}(n^*,m,1) \left[\sum_{k=0}^{n^*-1} q^0(k,n^*)\pi^{\phi^*}(k) \right] \right).
\end{aligned}$$

In the above expected lump-sum cost, the first term is the contribution in cost due to transition from states $0, 1, 2, \cdots, n^* - 1$ and the second term is that for the transition from state n^*. Additionally, it exploits the fact that $\pi^{\phi^*}(m) = 0 \ \forall \ m > n^*$. It follows from the expressions of the above expected costs that the long run average cost under the optimal policy ϕ^*,

$$\mathbb{E}(C(N^{\phi^*}, S^{\phi^*}(N^{\phi^*}))) \ + \mathbb{E}(C^{\infty,\phi^*}(N^{\phi^*}, S^{\phi^*}(N^{\phi^*}))) - W\mathbb{E}(f(N^{\phi^*}, S^{\phi^*}(N^{\phi^*}))),$$

is the same as the long run average cost under a 0-1 type threshold policy with threshold n^*,

$$\mathbb{E}(C(N^{n^*}, S^{n^*}(N^{n^*}))) \ + \mathbb{E}(C^{\infty,n^*}(N^{n^*}, S^{n^*}(N^{n^*}))) - W\mathbb{E}(f(N^{n^*}, S^{n^*}(N^{n^*}))).$$

Thus, a 0-1 type of threshold policy with threshold n^* is optimal when (i) is satisfied. The alternate rates (ii) can be proven to result in 0-1 type threshold optimality along the similar lines by considering the set $\max\{m \in \{0,1,...\} : S^{\phi^*}(m) = 0\}$. □

A.2 Proof of Proposition 2

Proof. We will focus on 0-1 type of threshold policies throughout the proof. The case of threshold policies of type 1-0 can be proven similarly. Since an optimal solution of problem (6) is of threshold type for a given subsidy W, the optimal average cost will be $g(W) := \min_n g^{(n)}(W)$ where

$$g^{(n)}(W) = \mathbb{E}(T(N^n, S^n(N^n))) - W\mathbb{E}(f(N^n, S^n(N^n))).$$

We denote the minimizer of $g(W)$ by $n(W)$. Note that the function $g(W)$ is a lower envelope of affine non-increasing functions of W due to the non-negative nature of $\mathbb{E}(f(\cdot))$. It thus follows that $g(W)$ is a concave non-increasing function.

It follows that the right derivative of $g(W)$ in W is given by $-\mathbb{E}(f(N^{n(W)}, S^{n(W)}(N^{n(W)})))$. Since $g(W)$ is concave in W, the right derivative is non-increasing in W. Together with the fact $\mathbb{E}(f(N^n, S^n(N^n)))$ is strictly increasing in n, it hence follows that $n(W)$ is non-decreasing in W. Since an optimal policy is of 0-1 threshold type, the set of states where it is optimal to be passive can be written as $D(W) = \{m : m \leq n(W)\}$. Since $n(W)$ is non-decreasing, by definition this implies that bandit k is indexable.

□

A.3 Proof of Proposition 3

We will focus on 0-1 type of threshold policies throughout the proof. Let $\tilde{W}(n)$ be the value for subsidy such that the average cost under threshold policy n is equal to that under threshold policy $n-1$. By using (6), we have $\mathbb{E}(T(N^n, S^n(N^n))) - \tilde{W}(n)\mathbb{E}(f(N^n, S^n(N^n))) = \mathbb{E}(T(N^{n-1}, S^{n-1}(N^{n-1}))) - \tilde{W}(n)\mathbb{E}(f(N^{n-1}, S^{n-1}(N^{n-1})))$. Hence, $\tilde{W}(n)$ is given by,

$$\frac{\mathbb{E}(T^n(N^n, S^n(N^n))) - \mathbb{E}(T^{n-1}(N^{n-1}, S^{n-1}(N^{n-1})))}{\mathbb{E}(f(N^n, S^n(N^n))) - \mathbb{E}(f(N^{n-1}, S^{n-1}(N^{n-1})))},$$

which is the same as (7). Since $\tilde{W}(n)$ is monotone, it can be verified by exploiting threshold optimality that $g(\tilde{W}(n)) = g^{(n)}(\tilde{W}(n)) = g^{(n-1)}(\tilde{W}(n))$. Similarly, $g(\tilde{W}(n-1)) = g^{(n-1)}(\tilde{W}(n-1)) = g^{(n-2)}(\tilde{W}(n-1))$. Further, monotonicity of $\tilde{W}(n)$ implies the following two possibilities:

1. Non-decreasing nature, i.e., $\tilde{W}(n-1) \leq \tilde{W}(n)$
2. Non-increasing nature, i.e., $\tilde{W}(n-1) \geq \tilde{W}(n)$

But $\tilde{W}(n-1) \geq \tilde{W}(n)$ results in a contradiction from indexability and 0-1 type of threshold optimality. Thus, $\tilde{W}(n)$ has to be non-decreasing, i.e., $\tilde{W}(n-1) \leq \tilde{W}(n)$.

It follows from indexability and 0-1 type of threshold optimality that for all $W \leq \tilde{W}(n)$, the set of states where it is optimal to be passive, $D(W)$, satisfies $D(W) \subseteq \{m : m \leq n-1\}$. Again from indexability in a similar way, $D(W) \supseteq \{m : m \leq n-1\}$ for all $W \geq \tilde{W}(n-1)$. Thus, for $\tilde{W}(n-1) \leq W \leq \tilde{W}(n)$, $\{m : m \leq n-1\} \subseteq D(W) \subseteq \{m : m \leq n-1\}$ which implies that threshold policy $n-1$ is optimal for all $\tilde{W}(n-1) \leq W \leq \tilde{W}(n)$ and hence $g(W) = g^{(n-1)}(W)$ for $\tilde{W}(n-1) \leq W \leq \tilde{W}(n)$. Hence, $\tilde{W}(n)$ is the smallest value of the subsidy such that activating the bandit in state n becomes optimal, that is, Whittle's index is given by $W(n) = \tilde{W}(n)$.

B Machine repairman problem

In this section, we provide the details to obtain the stationary distribution, prove indexability and derive Whittle's index for two specific models of the machine repairman problem of Section 5.

B.1 Stationary distribution

In this section, we determine the stationary distribution under a 0-1 type of threshold policy n. Thus, action $a = 0$ is taken in states $0, 1, 2, \cdots, n$ and action $a = 1$ in states $n+1, n+2, ...$ The transition diagram for the evolution of Markov chain is shown in Figure 2. The balance equations for the stationary distribution under the threshold policy n are given by

$$\begin{aligned} \lambda(0)\pi^n(0) &= \psi(1)\pi^n(1) + \psi(2)\pi^n(2) + \cdots + \psi(n)\pi^n(n) + r(n+1)\pi^n(n+1), \\ \lambda(m)\pi^n(m) &= (\lambda(m+1) + \psi(m+1))\pi^n(m+1) \quad \text{for } m = 0,\ 1,\ 2, ...,\ n-1, \\ \lambda(n)\pi^n(n) &= r(n+1)\pi^n(n+1). \end{aligned} \tag{15}$$

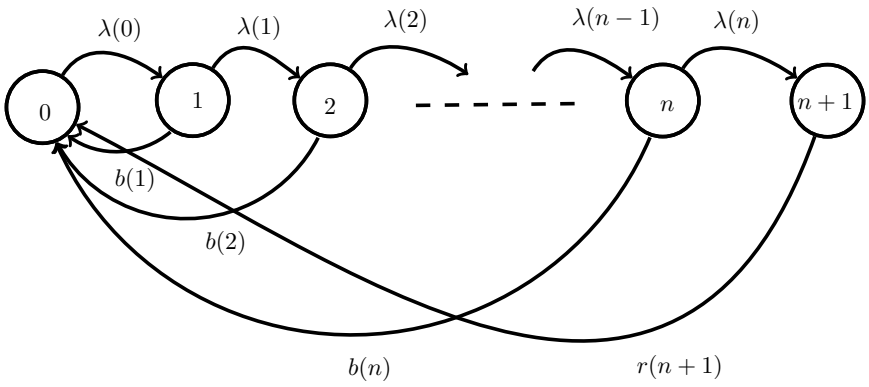


Figure 2: Transition diagram under the threshold policy n for machine repairman problem

Using $\sum_{m=1}^{n+1} \pi^n(m) = 1$, one obtains

$$\begin{aligned}
\pi_k^{n_k}(m_k) &= \frac{P_{m_k}}{\lambda_k(m_k)\left(\sum_{i=0}^{n_k} \frac{P_i}{\lambda_k(i)} + \frac{P_{n_k}}{r_k(n_k+1)}\right)} \ \forall\ m_k = 0, 1, 2, ...n_k, \\
\pi_k^{n_k}(n_k+1) &= \frac{P_{n_k}}{r_k(n_k+1)\left(\sum_{i=0}^{n_k} \frac{P_i}{\lambda_k(i)} + \frac{P_{n_k}}{r_k(n_k+1)}\right)}, \qquad (16) \\
\pi_k^{n_k}(m_k) &= 0 \ \forall\ m_k = n_k+2, \cdots
\end{aligned}$$

where $P_i = \prod_{j=1}^{i} p_k(j)$ and $p_k(j) = \frac{\lambda_k(j)}{\lambda_k(j)+\psi_k(j)}$; $P_0 = 1$.

B.2 Indexability

Lemma B.1. *Machine k is indexable if the repair rates are non-decreasing in their state, i.e., $r_k(n) \leq r_k(n+1)\ \forall\ n$, and $r_k(1) > 0$. In particular, all machines are indexable for state-independent repair rates.*

Proof. From Proposition 3, it follows that machine k is indexable if $\mathbb{E}(f_k(N_k^n, S_k^n(N_k^n)))$ is strictly increasing in n. Recall that $f_k(n, a) = \mathbf{1}_{\{a=0\}}$. Under the 0-1 type of threshold structure policy, with threshold n, we have

$$\mathbb{E}(f_k(N_k^n, S_k^n(N_k^n))) = \sum_{m=0}^{n} \pi_k^n(m).$$

Thus, machine k is indexable if $\sum_{m=0}^{n} \pi_k^n(m)$ is strictly increasing in n. Since $\pi_k^n(m) = 0$ for $m > n+1$, this is equivalent to proving that $\pi_k^n(n+1)$ is strictly decreasing in n. From Equation (16) and some algebra, we obtain that

$$\pi_k^n(n+1) - \pi_k^{n-1}(n) = \frac{\left(\frac{\lambda_k(n)(r_k(n)-r_k(n+1))-\psi_k(n)r_k(n+1)}{\lambda_k(n)+\psi_k(n)}\right)\sum_{i=0}^{n-1} \frac{P_i}{\lambda_k(i)} - r_k(n+1)\frac{P_n}{\lambda_k(n)}}{r_k(n)r_k(n+1)\left(\sum_{i=0}^{n} \frac{P_i}{\lambda_k(i)} + \frac{P_n}{r_k(n+1)}\right)\left(\sum_{i=0}^{n+1} \frac{P_i}{\lambda_k(i)} + \frac{P_{n+1}}{r_k(n+2)}\right)}.$$

Note that the denominator is strictly positive. Since $r_k(n)$ is non-decreasing and $r_k(1) > 0$, the numerator is strictly negative. That is, the result follows. □

B.3 Whittle's index: Proof of Proposition 4

Since a 0-1 type of threshold policy is optimal, using Proposition 3, the Whittle index is given by Equation (7), i.e.,

$$W_k(n) = \frac{\mathbb{E}(C_k(N_k^n, S_k^n(N_k^n))) - \mathbb{E}(C_k(N_k^{n-1}, S_k^{n-1}(N_k^{n-1})))}{\sum_{m=0}^{n} \pi_k^n(m) - \sum_{m=0}^{n-1} \pi_k^{n-1}(m)}, \tag{17}$$

if (17) is non-decreasing. The expected cost under threshold policy n in the nominator is given by

$$\mathbb{E}(C_k(N_k^n, S_k^n(N_k^n))) = \sum_{m=1}^{n} \left[\psi_k(m) L_k^b(m) + C_k^d(m)\right] \pi_k^n(m) + r_k(n+1) L_k^r(n+1) \pi_k^n(n+1).$$

Using the expression for the stationary distribution as derived in Appendix B.1, we obtain that the denominator of (17) simplifies to

$$\sum_{i=0}^{n} \pi_k^n(i) - \sum_{i=0}^{n-1} \pi_k^{n-1}(i) = \frac{\frac{P_{n-1}}{r_k(n)} \sum_{i=0}^{n} \frac{P_i}{\lambda_k(i)} - \frac{P_n}{r_k(n+1)} \sum_{i=0}^{n-1} \frac{P_i}{\lambda_k(i)}}{\left(\sum_{i=0}^{n} \frac{P_i}{\lambda_k(i)} + \frac{P_n}{r_k(n+1)}\right)\left(\sum_{i=0}^{n-1} \frac{P_i}{\lambda_k(i)} + \frac{P_{n-1}}{r_k(n)}\right)},$$

where $P_i = \prod_{j=1}^{i} p_k(j)$ and $p_k(j) = \frac{\lambda_k(j)}{\lambda_k(j)+\psi_k(j)}$; $P_0 = 1$. After some algebra, we obtain that (17) simplifies to the one stated in Proposition 4.

B.4 Model 1: Deterioration cost per unit

We consider now a particular case when there are no breakdowns. Thus, $\psi_k(n_k) = 0$ and $L_k^b(n_k) = 0$. This simplifies since $p_k(j) = 1$ and $P_i = 1$, and hence the expression in Proposition 4 simplyfies to

$$\frac{\left(\sum_{i=1}^{n} \frac{C_k^d(i)}{\lambda_k(i)} + L_k^r(n+1)\right)\left(\sum_{i=0}^{n-1} \frac{1}{\lambda_k(i)} + \frac{1}{r_k(n)}\right) - \left(\sum_{i=1}^{n-1} \frac{C_k^d(i)}{\lambda_k(i)} + L_k^r(n)\right)\left(\sum_{i=0}^{n} \frac{1}{\lambda_k(i)} + \frac{1}{r_k(n+1)}\right)}{\frac{1}{r_k(n)} \sum_{i=0}^{n} \frac{1}{\lambda_k(i)} - \frac{1}{r_k(n+1)} \sum_{i=0}^{n-1} \frac{1}{\lambda_k(i)}}. \tag{18}$$

If in addition $r_k(n) = r_k$ for all n, we obtain (after some algebra) from Equation (18) that

$$W_k(n) = r_k \left[\sum_{i=0}^{n-1} \frac{C_k^d(n) - C_k^d(i)}{\lambda_k(i)} + \frac{C_k^d(n) - r_k L_k^r}{r_k}\right]. \tag{19}$$

In addition,

$$W_k(n) - W_k(n+1) = r_k \left[(C_k^d(n) - C_k^d(n+1)) \left(\sum_{i=0}^{n} \frac{1}{\lambda_k(i)} + \frac{1}{r_k}\right)\right],$$

which is negative when the $C_k^d(n)$ is non-decreasing.

B.5 Model 2: lump-sum cost for breakdown

Here, we assume that $C_k^d(n_k) = 0$, $r_k(n) = r_k(n+1) = r_k$, $L_k^r(n) = R_k, L_k^b(n) = B_k \ \forall\ n$, and $\psi_k(n)$ is an increasing sequence. From Proposition 4, Whittle's index simplifies to (9). Hence, $W_k(n) - W_k(n+1)$ simplifies to:

$$W_k(n) - W_k(n+1) = \frac{r_k B_k \left(\frac{P_n}{r_k} + \sum_{i=0}^{n} \frac{P_i}{\lambda_k(i)}\right)\left(\frac{1}{\psi(n+1)} - \frac{1}{\psi(n)}\right)}{\left((1-p_k(n)) \sum_{i=0}^{n} \frac{P_i}{\lambda_k(i)} + \frac{p_k(n) P_n}{\lambda_k(n)}\right)\left((1-p_k(n+1)) \sum_{i=0}^{n+1} \frac{P_i}{\lambda_k(i)} + \frac{p_k(n+1) P_{n+1}}{\lambda_k(n+1)}\right)},$$

which is negative under the increasing breakdown rates assumption.

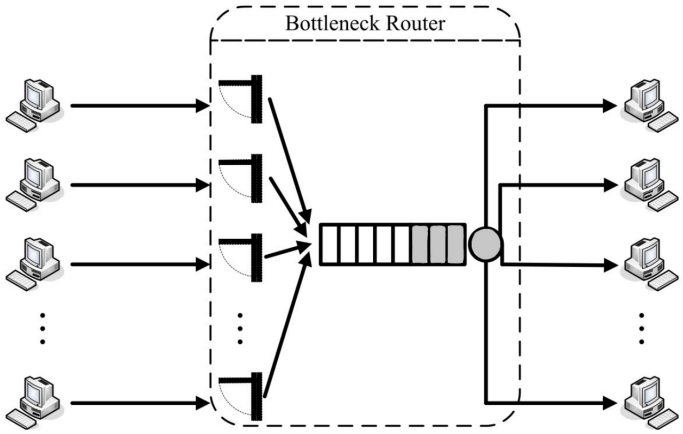


Figure 3: A bottleneck router in TCP with multiple flows [Avrachenkov et al., 2013].

C Congestion control in TCP

In Section 6.1 we described a TCP model, where multiple users (flows) are trying to transmit packets through a bottleneck router as shown in Figure 3.

C.1 Stationary distribution

Under a 1-0 type threshold policy n, action $a = 1$ is taken in states $0, 1, 2, \cdots, n$ and action $a = 0$ in states $n+1, n+2, ...$ When action $a = 0$ is taken at state $n + 1$, the state instantaneously changes to $S := \max\{\lfloor \gamma.(n + 1)\rfloor, 1\}$. Figure 4 shows the rates and its stationary distribution is given by

$$\begin{aligned} \pi^n(m) &= 0;\ m = 0, 1, 2, ..., S-1, \\ \pi^n(m) &= \frac{1}{n-S+1};\ m = S, S+1, ..., n. \end{aligned}$$

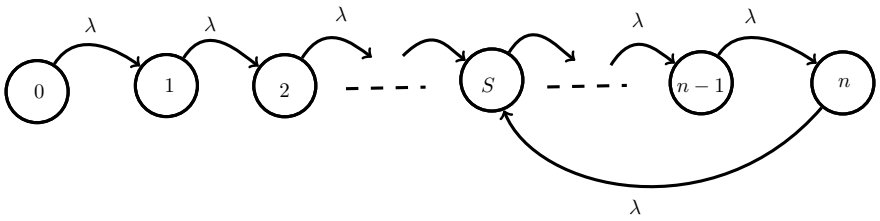


Figure 4: Transition diagram under the threshold policy 'n' for TCP congestion control problem

We then obtain

$$\mathbb{E}(f_k(N_k^n, S_k^n(N_k^n))) = \sum_{m=0}^{n} m\pi_k^n(m) = \frac{n^2 + n - S(S-1)}{2(n-S+1)},$$

where $S = \max\{\lfloor \gamma_k.(n + 1)\rfloor, 1\}$. It can be easily argued that

$$\mathbb{E}(f_k(N_k^n, S_k^n(N_k^n))) - \mathbb{E}(f_k(N_k^{n-1}, S_k^{n-1}(N_k^{n-1}))) = 1/2 > 0. \tag{20}$$

Thus, $\mathbb{E}(f_k(N_k^n, S_k^n(N_k^n)))$ is strictly increasing in n and the result follows.

C.2 Expression of Whittle's index: Proof of Lemma 6.1

Since 1-0 type of threshold policies are optimal, using Proposition 3, the Whittle index is given by

$$W_k(n) = \frac{\mathbb{E}(T_k^n(N_k^n, S_k^n(N_k^n))) - \mathbb{E}(T_k^{n-1}(N_k^{n-1}, S_k^{n-1}(N_k^{n-1})))}{\mathbb{E}(f_k(N_k^n, S_k^n(N_k^n))) - \mathbb{E}(f_k(N_k^{n-1}, S_k^{n-1}(N_k^{n-1})))}, \tag{21}$$

if Equation (21) is non-increasing in n.

We have

$$\mathbb{E}(T_k^n(N_k^n, S_k^n(N_k^n))) = \sum_{m=0}^{n} C_k(m,1)\lambda_k \pi_k^n(m) + C_k(n+1,0)\lambda_k \pi_k^n(n+1),$$

which simplifies to

$$\mathbb{E}(T_k^n(N_k^n, S_k^n(N_k^n))) = \begin{cases} \frac{\lambda_k \sum_{m=S}^{n} (1-(1+m)^{1-\alpha})}{(n-S+1)(1-\alpha)} & \text{if } \alpha \neq 1, \\ \frac{-\lambda_k \sum_{m=S}^{n} \log(1+m)}{(n-S+1)} & \text{if } \alpha = 1; \end{cases}$$

Together with (20) and Equation (21), results in the Whittle index as stated in Lemma 6.1.

D Content delivery network

We consider here the content delivery network as described in Section 6.2, see also Figure 5

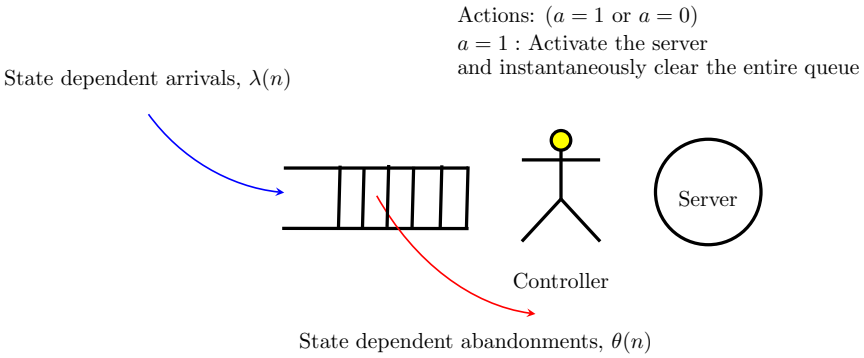


Figure 5: Optimal clearing framework as single-armed restless bandit

D.1 Stationary distribution

Under a 0-1 type of threshold policy n, action $a = 0$ is taken in states $0, 1, 2..., n$ and action $a = 1$ in states $n+1, n+2, ...$ The transition diagram is shown in Figure 6.

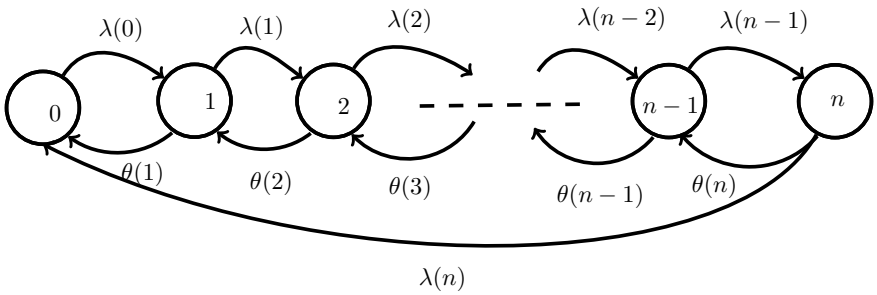


Figure 6: Transition diagram under the threshold policy n in the content delivery network

The balance equations under this chain are

$$\pi^n(0)\lambda(0) = \pi^n(1)\theta(1) + \lambda(n)\pi^n(n), \tag{22}$$

$$(\lambda(k) + \theta(k))\pi^n(k) = \lambda(k-1)\pi^n(k-1) + \theta(k+1)\pi^n(k+1) \quad \text{for } k = 1,\ 2, ..., n-1, \tag{23}$$

$$\lambda(n-1)\pi^n(n-1) = \theta(n)\pi^n(n) + \lambda(n)\pi^n(n), \tag{24}$$

which together with the normalization condition $\sum_{i=0}^{n} \pi^n(i) = 1$ results in the following stationary distribution:

$$\begin{aligned}
\pi^n(m) &= \frac{\pi^n(n)\lambda(n)}{\lambda(m)} \left[1 + \sum_{i=1}^{n-m} p(m+1, m+i)\right] \ \forall\ m = 0, 1, 2, ...n-1, \\
\pi^n(n) &= \left(1 + \sum_{k=0}^{n-1} \frac{\lambda(n)}{\lambda(k)} \left[1 + \sum_{i=1}^{n-k} p(k+1, k+i)\right]\right)^{-1}, \qquad (25) \\
\pi^n(m) &= 0 \ \forall\ m = n+1, \cdots, \qquad (26)
\end{aligned}$$

where $p(k+1, k+i) = \dfrac{\theta(k+1)\theta(k+2)...\theta(k+i)}{\lambda(k+1)\lambda(k+2)...\lambda(k+i)}\ \forall\ i \geq 1$.

The summation term in denominator of (25) is strictly increasing in n if $\lambda(n)$ is non-decreasing. Thus, $\pi^n(n)$ will be strictly decreasing in n under non-decreasing assumption on $\lambda(n)$.

D.2 Whittle's index: Proof of Lemma 6.2

The expected cost under threshold policy n is given by

$$\mathbb{E}(T^n(N^n, S^n(N^n))) = \sum_{i=1}^{n} (iC^h(i) + \theta(i)L^a(i))\pi^n(i) + \lambda(n)L_s^{\infty}(n+1)\pi^n(n).$$

Similarly, for the threshold policy $n-1$

$$\mathbb{E}(T^{n-1}(N^{n-1}, S^{n-1}(N^{n-1}))) = \sum_{i=1}^{n-1} (iC^h(i) + \theta(i)L^a(i))\pi^{n-1}(i) + \lambda(n-1)L_s^{\infty}(n)\pi^{n-1}(n-1).$$

From Proposition 3, we get the expression as stated in Lemma 6.2.