Improving Upon the generalized $c\mu$ rule: a Whittle approach

Zhouzi Li [*]   Keerthana Gurushankar   Mor Harchol-Balter
Alan Scheller-Wolf
Carnegie Mellon University

April 14, 2025

Abstract

Scheduling a stream of jobs whose holding cost changes over time is a classic and practical problem. Specifically, each job is associated with a holding cost (penalty), where a job's instantaneous holding cost is some increasing function of its class and current age (the time it has spent in the system since its arrival). The goal is to schedule the jobs to minimize the time-average total holding cost across all jobs.

The seminal paper on this problem, by Van Mieghem in 1995 [24], introduced the generalized $c\mu$ rule for scheduling jobs. Since then, this problem has attracted significant interest but remains challenging due to the absence of a finite-dimensional state space formulation. Consequently, subsequent works focus on more tractable versions of this problem.

This paper returns to the original problem, deriving a heuristic that empirically improves upon the generalized $c\mu$ rule and all existing heuristics. Our approach is to first translate the holding cost minimization problem to a novel Restless Multi-Armed Bandit (R-MAB) problem with a *finite* number of arms. Based on our R-MAB, we derive a novel Whittle Index policy, which is both elegant and intuitive.

1 Introduction

Since the seminal paper by Van Mieghem in 1995 [24], the problem of scheduling jobs with Time-Varying Holding Cost (the TVHC problem) has been an important topic in the operations literature: In a single-server multi-class system with k classes of jobs, each job incurs a (time-varying) holding cost for every unit of time it remains in the system, and the goal of the scheduling policy is to minimize the time-average total holding cost across all jobs. Specifically, define the *age* of a job to be the time it has spent in the system since it arrived. For a job of class i, let $c_i(t)$ be the instantaneous holding cost when the job's age is t. We allow different classes to have different holding-cost functions (see Figure 1), but assume these functions are non-decreasing. Note that non-decreasing *instantaneous holding cost* is equivalent to convex *accumulated holding cost*, which is assumed in [24] and all its follow-on works. Also, throughout this paper, we assume that job sizes are *exponentially* distributed unless otherwise specified, and we assume that the job arrival process is Poisson. Let λ_i and μ_i, respectively, denote the arrival rate and the completion rate of class i jobs.

In the special case when the holding cost of each class is *a constant function* (where $c_i(t) = c_i$), the optimal policy is the famous $c\mu$ rule, which always (preemptively) runs the job with the highest

---

[*]zhouzil@andrew.cmu.edu

product $c_i \cdot \mu_i$ [9, 8]. However, in the general case, this problem is much more complicated and the optimal policy is not known. In [24], an analogue of the $c\mu$ rule is proposed, which is now commonly referred to as the *generalized $c\mu$ rule*: The priority of a job is given by the product of its instantaneous holding cost, $c_i(t)$, and its instantaneous failure rate $\mu_i(t)$. While the generalized $c\mu$ rule is defined for general job sizes in [24], if the job sizes are exponentially distributed, the failure rate for class i is just μ_i. We discuss the generalized $c\mu$ rule in Section 1.1.

Note that both the generalized $c\mu$ rule and the $c\mu$ rule are *index policies*: Each job has an index (in both cases the value $c_i(t)\mu_i(t)$) which is a function of only the job's state, and the policy always serves that job with the highest index. Index policies are both simple and powerful. This paper aims to solve the TVHC problem within the class of preemptive *index policies*, yielding a heuristic solution to the problem in general. We now review the literature, and motivate our approach.

1.1 The generalized $c\mu$ rule

The generalized $c\mu$-rule has been shown to be asymptotically optimal for M/G/1 queues in the diffusion limit regime, where both the arrival and service rates go to infinity (and the total load goes to 1) [24]. Under the same diffusion limit, its analogues have also been shown to be asymptotically optimal for more complicated settings such as the multi-server setting and systems with abandonment ([18, 4, 17]). However, outside of the diffusion limit regime, it is known that the generalized $c\mu$-rule can perform poorly. Here we give an example to illustrate this.

Consider a system with two classes of jobs: one with deadlines and one without. For deadline-based jobs, the holding cost is zero before their deadline but becomes significantly higher once the deadline is passed, while jobs without deadlines have a constant holding cost (see Figure 2). Assume that both classes have job sizes following the same $\text{Exp}(\mu)$ distribution. Under the generalized $c\mu$ rule, the system grants priority to deadline-based jobs only *after* they have missed their deadlines. Intuitively, however, it might make more sense to prioritize these jobs *before* their deadline is reached. Simulation results show that by prioritizing the deadline-based jobs before their deadline, the overall holding cost can be substantially reduced in normal-traffic scenarios (see Figure 7b).

Thus, the generalized $c\mu$ rule can be highly suboptimal under normal traffic (even with exponential job sizes), underscoring the need for a better scheduling policy.
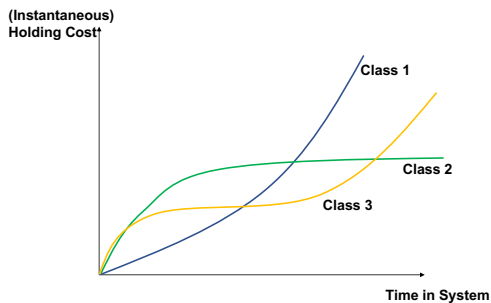

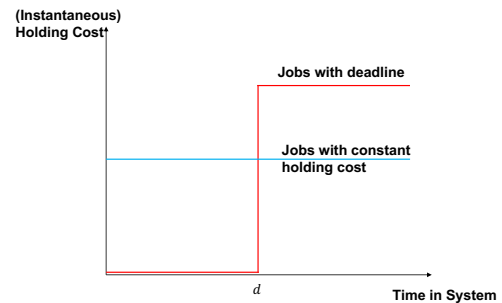Figure 1: Classes with different holding costs.

Figure 2: Example: generalized cμ rule is suboptimal.

1.2 Prior attempts to improve upon the generalized $c\mu$ rule

One of the difficulties in solving the TVHC problem lies in the fact that a job's holding cost depends on its age. Hence, we need to track the ages of all jobs in the system, which requires an *infinite-dimensional* state space. As a result, the existing literature has considered more tractable versions of the TVHC problem that have a *finite-dimensional* state space. Such a finite-dimensional state space allows the authors to represent their problem as a finite-dimensional Markov Decision Problem (MDP), or, more specifically, a Restless Multi-Armed Bandit problem (R-MAB) with a finite number of arms. We provide a brief tutorial on R-MABs in Section 2.

One way to create a finite-dimensional state space for the TVHC problem is to assume a static setting, where all n jobs are present at time 0 and there are no new arrivals. This is the approach taken in [2, 1]. By limiting the number of jobs, the problem can now be translated to an n-arm R-MAB. From this R-MAB the authors then derive a Whittle index, which determines which job to run at every moment in time (see Appendix A for a tutorial on the Whittle index). Figure 3 shows a road-map of the solution. The drawback of the static version setting is that the arrival rate (and consequently the load) of each class cannot be incorporated in the policy. This is unfortunate, because, for example, in Figure 2, it is reasonable to expect that if load is higher, we might want to start working on the deadline-oriented jobs sooner than we would under lower load.

Static Version $\longrightarrow$ **R-MAB** $\longrightarrow$ **The Whittle Index**

Translate to   Derive

Figure 3: *A road-map to derive the Whittle Index policy for the static version of our problem: The static version is first translated to a R-MAB problem. Then the Whittle Index is derived based on the R-MAB problem.*

Another way to create a finite-dimensional state space is to change the TVHC problem so that individual jobs no longer have a holding cost. Instead, the total holding cost for class i is a function of the *number* of class i jobs present, see [13, 5, 7, 16, 3, 15, 12]. With this change, one only needs to track the number of jobs within each class, enabling the problem to be translated into a finite-dimensional state space MDP or a k-arm R-MAB. Several papers, [3, 15, 12], next follow the road-map in Figure 4 where they use the k-arm R-MAB to derive a Whittle index policy. This queue-length holding cost setting is complementary to our age-based holding cost setting, as the policy derived for one setting cannot be translated into a policy in the other setting.

Queue-Length Version $\longrightarrow$ **R-MAB** $\longrightarrow$ **The Whittle Index**

Translate to   Derive

Figure 4: *A road-map used to derive the Whittle Index policy for the queue-length holding cost setting.*

1.3 Our approach and contributions

The Whittle Index policy is known to be a good heuristic for R-MAB problems (see [19] for a discussion). Like the prior work of Figures 3 and 4, we adopt a similar road-map to derive the Whittle Index for our age-based holding cost minimization problem. Unfortunately, in our problem we have infinite-dimensional state space which makes things much harder.[1] First, while

---

[1]In fact, both papers on the static version [2, 1] note the analytical *intractability* of the dynamic age-based TVHC setting.

3

the translation to a tractable R-MAB is straightforward in prior works, it is hard to reduce our problem to a finite-arm R-MAB. Second, as in all the prior works, even after reducing to an R-MAB problem, we still need to derive the Whittle Index from the R-MAB, which requires overcoming two obstacles: *(i)* establishing *indexability*, a key property of the R-MAB ensuring that the Whittle Index is well-defined; and *(ii)* performing the actual derivation of the Whittle index, which can be intricate.

In overcoming these difficulties, our paper makes two primary contributions, outlined in Figure 5: First, Theorem 1 (see Section 3) translates the TVHC problem to a novel R-MAB problem. Second, Theorem 2 (see Section 4) proves indexability and derives the Whittle Index for our R-MAB problem, thus yielding a Whittle-based heuristic policy for our problem.
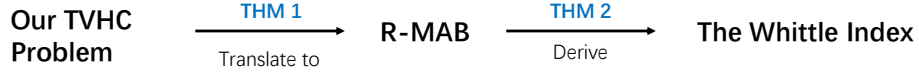
$$
\textbf{Our TVHC Problem} \quad \xrightarrow[\text{Translate to}]{\text{THM 1}} \quad \textbf{R-MAB} \quad \xrightarrow[\text{Derive}]{\text{THM 2}} \quad \textbf{The Whittle Index}
$$

Figure 5: Using Theorem 1 and Theorem 2, we follow this road-map to derive the Whittle Index policy.

Our final result is a Whittle-based policy which always preemptively runs the job with the highest Whittle Index. The Whittle Index has an elegant and intuitive formulation given in Corollary 1 (see Section 4) and repeated here: Our Whittle Index for a class i job with age t, $W_i(t)$, is given by

$$
W_i(t) = \mu_i \cdot \mathbf{E}\left[c_i(t+X)\right], \quad \text{where } X \sim \text{Exp}(\mu_i - \lambda_i). \tag{1}
$$

Intuitively, while the generalized $c\mu$ rule focuses on the current holding cost $c_i(t)$ and schedules according to the index $\mu_i c_i(t)$, our Whittle index looks a little further into the future to age $t + X$. Note that if the load is high (λ_i is close to μ_i), then we look further into the future. This aligns with the motivating example in Figure 2, where it is preferable to prioritize a job before its deadline, particularly when there are likely to be many jobs in the system.

It is also interesting to contrast our policy with the heuristic given in [1], which is the Whittle index in the static version. Under exponential job sizes, their index has the form $\mu_i \cdot \mathbf{E}\left[c_i(t+Y)\right]$, where $Y \sim \text{Exp}(\mu_i)$. Note that our index given in (1) matches theirs when $\lambda_i = 0$, where our setting degenerates to the static setting. In Section 5 we compare our policy with all the existing heuristics (including [1]). Our simulations show that our policy outperforms all the other heuristics.

2 Tutorial on (Markov) R-MAB Problem

In this paper, when we say MAB, we always refer to the *Markov* MAB: In a Markov MAB problem, each arm corresponds to a Markov process with states that evolve based on an underlying state transition model. The agent chooses to pull an arm at each time step, incurring a cost determined by the states of the arms. The objective is to minimize the cumulative cost over time.

Mathematically, let k denote the number of arms. Each arm i is associated with a Markov Decision Process (MDP), which is defined by a state space, the action space $\{active, passive\}$, the transition probabilities given states and actions, and a cost function $c_i(s)$ representing the cost incurred when arm i is in state s. At each time step t, the agent observes the current states of all arms and selects at most one arm to pull (pulling means choosing the action *active*). Cost is incurred for each arm,

and the state of each arm evolves according to its respective MDP and actions. The objective is to minimize the long-run average cost over an infinite time horizon.

A *discounted* MAB is defined exactly the same except for the objective. Instead of the long-run average cost, a discounted MAB aims to minimize the cumulative discounted cost, where the cost at time t is multiplied by $e^{-\alpha t}$, and $\alpha > 0$ is called the discount factor.

A MAB is called *restful* if the state of an arm does not evolve when the arm is not pulled (the action is *passive*). In contrast, An MAB is called *restless* if the state evolves whether the arm is pulled or not. For a restful MAB, the optimal policy is the Gittins index policy ([11, 22]), while the optimal policy for restless MAB (R-MAB) is open. The TVHC problem is intrinsically restless.

3 Translation to a Discounted R-MAB

The main goal of this section is to complete the first step in the road-map (Figure 5), which is the translation from the TVHC problem to a discounted R-MAB problem. First in Section 3.1, we restate the TVHC problem setting, and prove that one should serve jobs within each class in FCFS order. This fact allows us to only consider index functions that enforce FCFS within each class, which we will see is essential for the translation to an R-MAB problem. Second in Section 3.2, we prove our theorem which translates the TVHC problem to an R-MAB problem. Finally in Section 3.3, a discount factor is introduced to the R-MAB problem. The discounted R-MAB problem serves as the starting point of the typical Whittle Index approach. The Whittle Index for the R-MAB without discounting is obtained by taking the limit on the discount factor (see Section 3.3).

3.1 The TVHC problem

3.1.1 Problem setting

We briefly restate the TVHC problem setting as follows: In a single-server system, there are k types of jobs. For each type i, we assume a Poisson arrival process with rate λ_i. We also assume that the job size distribution is exponentially distributed with rate μ_i. For a type i job of age t, its instantaneous holding cost is $c_i(t)$, which is a non-decreasing and smooth[2] function. The objective of the scheduler is to schedule the jobs in order to minimize the time-average mean holding cost. Specifically, Let $c_{total}(s)$ denote the sum of the holding costs of all jobs currently in the system at time s. Then the objective is to minimize $\mathbf{E}$ [Holding Cost], where

$$\mathbf{E}\,[\text{Holding Cost}] = \lim_{t \to \infty} \frac{1}{t} \int_{s=0}^{t} c_{total}(s)ds.$$

To make the problem well-defined, we assume that there exists a policy to make the time-average mean holding cost converge.

---

[2]We need smoothness to simplify our derivation. However, our policy applies to any holding cost function that can be arbitrarily closely approximated by a smooth function (e.g., a deadline function, or any continuous function).

3.1.2   Index functions

We only focus on index policies in this paper. These are policies where a job's index (priority level) only depends on its own state, not on the state of other jobs in the system. Specifically, since the job sizes are exponential, a job's index only depends on the job's type and age and does not depend on its attained service. Thus an index policy is specified by a set of functions $\{V_i(\cdot)\}_{i=1}^k$, where $V_i(t)$ is the index of a job of type i with age t.

3.1.3   FCFS within each class

Intuitively, since the holding cost is increasing and the job sizes are exponentially distributed, to minimize mean holding cost, we should schedule jobs within each type in FCFS (First Come First Serve) order , since earlier arriving jobs have higher cost. Mathematically, we have the following lemma:

Lemma 3.1 (FCFS within each type is optimal). *The optimal policy must serve jobs within each type in FCFS order.*

Proof. See Appendix B. ☐

Motivated by Lemma 3.1, throughout this paper, we focus on index policies that enforce FCFS order within each class. This is equivalent to having a non-decreasing index function for each class with a FCFS tie-breaking rule. In Section 3.2, we will leverage the fact that the index functions enforce FCFS within each class to substantially reduce the dimensionality of the state space.

3.2   Translation to an R-MAB problem

In this section, we translate the TVHC problem into an R-MAB problem. Intuitively, our key idea is to let each arm of the R-MAB problem track the age of the oldest job within each class in the TVHC problem. At first this seems insufficent, because we're not capturing the state of all the other jobs within each class. However we will show how the FCFS ordering within each class ensures that the distribution of the ages of younger jobs can be effectively captured through the stochastic behavior of the Poisson arrival process (a generally similar, yet distinct, idea is applied to a different problem in [23]). Our key theorem is as follows:

Theorem 1 (Translation to an R-MAB). *There exists an R-MAB problem, such that for any set of non-decreasing index functions $\{V_i(\cdot)\}_{i=1}^k$, the corresponding index policies (breaking ties by FCFS) in the TVHC problem and the R-MAB problem incur the same cost.*

Proof Sketch. We construct a continuous R-MAB as follows.

- The R-MAB has k arms, each representing a class. There are two actions for each arm (active or passive), where arm i is active means the oldest class i job is served, and passive means the job served at this moment is of some other type.

- *Arm state $T_i(t)$*: The ith arm's state is $T_i(t) \in \mathbb{R}$. The arm state can be interpreted in the TVHC problem as the age of the oldest type i job at time t. If there is no type i job in the system, $T_i(t)$ is negative, which means the next type i arrival happens $-T_i(t)$ time later. Note that for arm i, the action active is only allowed at time t when $T_i(t)$ is positive.

- *Transition Probability:* If the action for arm i is passive, the ith arm state grows with rate 1. Otherwise if arm i is active, the ith arm state may drop an $\text{Exp}(\lambda_i)$ amount according to a Poisson process (when completions happen). Mathematically, the transition function is

$$\text{passive: } dT_i(t) = dt, \quad \text{active: } dT_i(t) = dt - X \cdot dN_{\mu_i}(t),$$

  where $X \sim \text{Exp}(\lambda_i)$, and N_{μ_i} is a Poisson counting process with rate μ_i.

  This transition function can be interpreted in the TVHC problem as follows: A passive action means that the oldest class i job is not in service and its age grows with rate 1. If the action is active, then in the next dt time period, the oldest class i job is served. There is a probability of $\mu_i dt$ that the job is completed and leaves the system, in which case the oldest class i job in the system becomes the previously second-oldest class i job. Since the inter-arrival time follows the distribution $\text{Exp}(\lambda_i)$, the age of the oldest job drops by an $\text{Exp}(\lambda_i)$ amount.

- *Constraint:* The number of active arms at any time is at most 1.

- *Cost Function $r_i(s)$:* Arm i incurs a cost of $r_i(s)$ at state s, where $r_i(s)$ is defined to be

$$r_i(s) := c_i(s) + \mathbf{E}\left[\sum_{j=1}^{N_i} c_i(Y_j)\right], \tag{2}$$

  where $Y_j = \sum_{m=1}^{j} X_m$, $X_m \sim \text{Exp}(\lambda_i)$, and N_i is the random variable such that $Y_{N_i} < s$ and $Y_{N_i+1} \geqslant s$. For $s < 0$, define $r_i(s) = 0$.

  To interpret (2), observe that $r_i(s)$ represents the expected total instantaneous holding cost of all class i jobs given that the age of the oldest class i job is s: Since the index policy serves class i jobs in FCFS order, no "young" class i jobs (class i jobs younger than the current oldest one) have been completed. Since the inter-arrival times are distributed as $\text{Exp}(\lambda_i)$, their ages are distributed as

$$s - X_1, \ s - X_1 - X_2, \ \ldots, \ s - \sum_{m=1}^{N_i} X_m.$$

  Thus the expected total instantaneous cost of all the young class i jobs is $\mathbf{E}\left[\sum_{j=1}^{N_i} c_i(s - Y_j)\right]$, which is equal to $\mathbf{E}\left[\sum_{j=1}^{N_i} c_i(Y_j)\right]$ as the N_i arrivals are distributed as uniform order statistics in $(0, s)$.

- *Objective:* The objective is to minimize the long-run expected cost. Mathematically,

$$\text{Cost} = \mathbf{E}\left[\lim_{x \to \infty} \frac{1}{x} \int_0^x \sum_{i=1}^{k} r_i(T_i(t)) dt \ \middle| \ T_i(0) = 0\right].$$

Through the construction and the corresponding interpretation in the TVHC problem, it is straightforward to see a coupling between the ith arm state and the age of the oldest class i job in the TVHC problem. However, although intuitive, it is delicate to rigorously prove that characterizing the holding cost of all young jobs by expectation does not change the expected long-run mean cost.

We refer to Appendix C for the rigorous proof, where we use a coupling argument over sample paths to build equivalence between the two problems. □

3.3 Introducing the discount factor

Finally, we introduce a discount factor into our R-MAB problem. It is a typical step before applying the Whittle Index approach. The discounted objective is defined to be:

$$\text{Cost}(\mathbf{s_0}, \alpha) = \mathbf{E}\left[\int_0^\infty \sum_{i=1}^k r_i(T_i(t)) \cdot \alpha e^{-\alpha t} dt \;\middle|\; T_i(0) = \mathbf{s_0}(i)\right], \tag{3}$$

where $\mathbf{s_0}$ is the vector of the initial states and $\mathbf{s_0}(i)$ is the initial state of arm i.

By standard results in dynamic programming (e.g., [3, 21]), also known as the Wiener's Tauberian theorem, we have the following lemma:

Lemma 3.2. *For any initial state* $\mathbf{s_0}$,

$$\lim_{\alpha \to 0} Cost(\mathbf{s_0}, \alpha) = Cost.$$

This lemma indicates that to get a good index for our R-MAB problem (and thus the TVHC problem), it suffices to find a good index policy to optimize the discounted cost $\text{Cost}(\mathbf{s_0}, \alpha)$, and then take the limit $\alpha \to 0$ on the index.

4 Derivation of the Whittle Index

Given the discounted R-MAB problem, we now perform the second part of the road-map of Figure 5: the derivation of the Whittle Index. Our main theorem is Theorem 2. In Corollary 1 we obtain the Whittle Index for our TVHC problem by taking the limit on the discount factor in Theorem 2.

The section is organized as follows: First in Section 4.1, we define the single-arm bandit formulation and define the Whittle Index. Then in Section 4.2 we propose a guess of the Whittle Index. The closed form of the guessed Whittle Index is derived in Section 4.3. Finally, we verify indexability and that this guess of the Whittle Index is true in Section 4.4, hence proving Theorem 2.

In most parts of this section, we focus on the single-arm bandit problem, and we drop the subscript i for simplicity. However, we should remember that there is an independent single-arm bandit problem for each arm in our R-MAB problem, and the Whittle Index obtained from each single-arm bandit problem serves as the index for each arm in the R-MAB problem.

4.1 Set up the Whittle Index approach

In this section, we set up the problem following the Whittle approach (see Appendix A for a tutorial). We first define the single-arm bandit for class i where not pulling the arm is rewarded with some compensation. Based on this single-arm bandit formulation, the Whittle Index is defined.

Formally, we drop the subscript i and redefine our problem as follows.

- *State and Action:* As before, the state is denoted by $T(t)$. The action set is {active, passive}, and "active" is only allowed at time t if $T(t) > 0$.

- *Transition Probability:*

$$\text{passive: } dT(t) = dt, \quad \text{active: } dT(t) = dt - X \cdot dN_\mu(t), \tag{4}$$

  where $X \sim \text{Exp}(\lambda)$, and N_μ is a Poisson counting process with rate μ.

- *Passive Compensation ℓ:* We define ℓ to be the compensation for a passive action. It is further explained in the cost function below.

- *Cost Function $\theta(s, a)$:* We define the cost incurred at state s and action a to be

$$\theta(s, a) := \begin{cases} \alpha \cdot r(s), & \text{if } a = \text{active,} \\ \alpha \cdot r(s) - \ell, & \text{if } a = \text{passive,} \end{cases}$$

  where $r(s)$ is defined as before:

$$r(s) := c(s) + \mathbf{E}\left[\sum_{j=1}^{N} c(Y_j)\right], \tag{5}$$

  where $Y_j = \sum_{m=1}^{j} X_m$, $X_m \sim \text{Exp}(\lambda)$, and N is the random variable such that $Y_N < s$ and $Y_{N+1} \geqslant s$. For $s < 0$, define $r(s) = 0$.

- *Objective:* The objective is to minimize the discounted total holding cost. Mathematically, let s_0 denote the initial state. Then the discounted total holding cost is

$$\text{Cost}(s_0, \alpha) = \mathbf{E}\left[\int_0^\infty \theta(T(t)) \cdot e^{-\alpha t} dt \ \middle| \ T(0) = s_0\right].$$

A policy π takes a state T and outputs an action a, where a is active or passive. Thus we denote a policy by a function $\pi : (-\infty, \infty) \to \{0, 1\}$ where 1 means active. Note that any policy receiving a negative state must return the passive action.

For any value of the compensation ℓ, consider the optimal policy of the single-arm bandit problem. Define $\Pi(\ell)$ to be the set of states where the passive action is optimal. Note that for any compensation ℓ, any non-positive state is in $\Pi(\ell)$.

Definition 4.1 (Π)**.** *For any compensation ℓ, define*

$$\Pi(\ell) := \{t_0 \mid \text{passive action at state } t_0 \text{ is optimal}\}.$$

Throughout, we omit writing the discount factor α, to simplify notation.

Intuitively, the larger the compensation, the more likely it is that a passive action should be taken. This property is called "indexability."

Definition 4.2 (Indexability)**.** *A problem is indexable if for any $\ell_1 < \ell_2$, we have $\Pi(\ell_1) \subseteq \Pi(\ell_2)$.*

The Whittle Index of a state is defined to be the smallest compensation such that the passive action is optimal at this state. It can be intuitively understood as the value of compensation where active and passive actions are both optimal.

Definition 4.3 (Whittle's Index). *The Whittle Index of a state t with discount factor α is defined to be*

$$W(t, \alpha) := \inf_{\ell}\{t \in \Pi(\ell)\}.$$

Notice that both indexability and the form of the Whittle Index involves understanding $\Pi(\ell)$, which requires us to understand the optimal policy of this single-arm bandit formulation. We follow these steps to establish indexability and derive the Whittle Index: *(i)* We propose a guess of the optimal policy, and accordingly define the guessed Whittle Index (Section 4.2); *(ii)* We compute the guessed Whittle Index (Section 4.3); *(iii)* We verify indexability and that the guessed Whittle Index is the true Whittle Index (Section 4.4).

4.2 A guess of the Whittle Index

Intuitively, it is reasonable to guess that the optimal policy is a threshold policy that chooses the passive action for states smaller than the threshold and chooses active otherwise. Mathematically, we define a threshold policy $Threshold(x)$ below:

Definition 4.4 ($Threshold(x)$). *A threshold policy, $Threshold(x)$, parameterized by state x, selects the passive action iff state $t < x$.*

For any threshold policy, denote its cost by the following notation:

Definition 4.5 (Cost of $Threshold(x)$). *For policy $Threshold(x)$, define the discounted total holding cost given discount factor α, initial state s_0 and compensation ℓ to be $Cost^x(s_0, \alpha, \ell)$.*

With the guess that a threshold policy is optimal, the following "guessed Whittle Index" is proposed. Both guesses will be proven true in Section 4.4.

Definition 4.6 (Guessed Whittle's Index). *The guessed Whittle Index of state t_0, $\widehat{W}(t_0, \alpha)$, is the value of the compensation ℓ such that the policy $Threshold(t_0)$ is the optimal threshold policy for initial state t_0. Mathematically, define $\widehat{W}(0, \alpha) = 0$ and for any $t_0 > 0$, define $\widehat{W}(t_0, \alpha)$ to be the value of ℓ that satisfies* [3]

$$\lim_{\delta \to 0} \frac{Cost^{t_0}(t_0, \alpha, \ell) - Cost^{t_0+\delta}(t_0, \alpha, \ell)}{\delta} = 0. \tag{6}$$

Unfortunately, it is not easy to solve the guessed Whittle Index from (6). The whole next subsection (Section 4.3) is devoted to solving (6).

---

[3]Intuitively, the optimal Threshold policy is the value of x that minimizes $Cost^x(t_0, \alpha, \ell)$. Differentiating yields (6).

4.3 The Guessed Whittle Index

In this section we will solve (6) for the guessed Whittle Index, resulting in Lemma 4.20. First we introduce some notation. Then we analyze the cost of a threshold policy (Lemma 4.11) and give a (complicated) expression for the guessed Whittle Index (Lemma 4.12). Finally, we show how to simplify the expression to the elegant and simple formula in Lemma 4.20.

4.3.1 Notation

A short summary of the notation used is given in Table 1.

<div align="center">

Table 1: Additional Notation Table

Notation	Meaning
$\overline{Cost}(t_0, \alpha)$	The discounted cost incurred during an M/M/1 busy period
T_1	$T_1 \sim BP^{M/M/1}$, the length of an M/M/1 busy period
T_2	$T_2 \sim \text{Exp}(\lambda)$
$\gamma_i, \ i = 1, 2$	$\mathbf{E}\left[e^{-\alpha T_i}\right]$
$\Gamma(t_0, \alpha)$	$\mathbf{E}\left[\int_0^{T_2} \alpha r(t_0 - T_2 + t)e^{-\alpha t}dt\right]$
X	$X \sim \text{Exp}\left(\frac{\alpha}{1-\gamma_1}\right)$

</div>

Definition 4.7 ($\overline{Cost}$). *In an M/M/1 queue, suppose a busy period starts at time 0, and the age of the oldest job in the system is defined to be $A(t)$. Define T_1 to be the first time such that $A(t) < 0$ (which means there is no job in the system and the busy period ends). Define $\overline{Cost}(t_0, \alpha)$ to be the expected discounted cost incurred during the busy period where the cost at time t is $r(t_0 + A(t))$:*

$$\overline{Cost}(t_0, \alpha) := \mathbf{E}\left[\int_0^{T_1} r(t_0 + A(t)) \cdot \alpha e^{-\alpha t}dt\right]. \tag{7}$$

Definition 4.8 ($\Gamma(t_0, \alpha)$). *Let $T_2 \sim Exp(\lambda)$. We define $\Gamma(t_0, \alpha)$ to be the expected discounted cost during the interval $[t_0 - T_2, t_0]$:*

$$\Gamma(t_0, \alpha) := \mathbf{E}\left[\int_0^{T_2} r(t_0 - T_2 + t) \cdot \alpha e^{-\alpha t}dt\right]. \tag{8}$$

Definition 4.9 (γ_1, γ_2). *Define $\gamma_1 := \mathbf{E}\left[e^{-\alpha T_1}\right]$ (where $T_1 \sim BP^{M/M/1}$ is the length of an M/M/1 busy period) and $\gamma_2 := \mathbf{E}\left[e^{-\alpha T_2}\right]$ (where $T_2 \sim Exp(\lambda)$).*

We have the following lemma characterizing γ_1 and γ_2.

Lemma 4.10. *γ_1 and γ_2 satisfy the following equations:*

$$\mu = \frac{(\alpha + \lambda - \lambda\gamma_1)\gamma_1}{1 - \gamma_1}, \qquad \gamma_2 = \frac{\lambda}{\lambda + \alpha}. \tag{9}$$

Proof. Since $\gamma_1 = \mathbf{E}\left[e^{-\alpha T_1}\right] = \widetilde{T_1}(\alpha)$, where T_1 is the length of an M/M/1 busy period, we have

the following equation from classic queueing theory (e.g., see Section 27.2 in [14]):

$$\gamma_1 = \frac{\mu}{\mu + \alpha + \lambda - \lambda\gamma_1},$$

yielding the first equation. Likewise, the second equation comes from:

$$\gamma_2 = \mathbf{E}\left[e^{-\alpha T_2}\right] = \widetilde{T_2}(\alpha) = \frac{\lambda}{\lambda + \alpha}.$$

$\square$

4.3.2 An expression for the guessed Whittle Index

Now we work on solving (6) for the guessed Whittle Index. In this subsection, through analyzing the cost of a threshold policy (Lemma 4.11), a complicated expression for the guessed Whittle Index is derived (Lemma 4.12).

We first analyze the cost incurred by the $Threshold(t_0+\delta)$ policy. Recall that the cost of a threshold policy is defined in Definition 4.5.

Lemma 4.11. *For any $\delta \geqslant 0$, we have that*

$$Cost^{t_0+\delta}(t_0, \alpha, \ell) = \frac{1}{1 - \gamma_1\gamma_2}\Bigg(\delta \cdot (\alpha \cdot r(t_0) - \ell) + e^{-\alpha\delta} \cdot \overline{Cost}(t_0 + \delta, \alpha)$$

$$+ e^{-\alpha\delta}\gamma_1 \cdot \left(\Gamma(t_0 + \delta, \alpha) - \delta\alpha\gamma_2 r(t_0) - \ell \cdot \frac{1}{\alpha}(1 - \gamma_2 \cdot e^{\alpha\delta})\right)\Bigg) + o(\delta). \tag{10}$$

Proof. We consider the state transition process of policy $Threshold(t_0 + \delta)$ from the initial state t_0 until the next time the state returns to t_0. As shown in Figure 6, there are three phases:

Phase 1 (the $[0, \delta]$ time period): The policy stays passive, and the cost incurred at every moment is $\alpha r(t) - \ell$. Thus the incurred discounted cost during phase 1 is

$$\delta \cdot (\alpha \cdot r(t_0) - \ell) + o(\delta). \tag{11}$$

Phase 2 (the time from δ until the next time that the state is below $t_0 + \delta$): Define T_1 to be the length of phase 2. Since the policy stays active from time δ, we know that the dynamics of the state is exactly the same as that of the oldest age in an M/M/1 queue. Thus the time period from δ until there is no job with age more than $t_0 + \delta$ is the length of a busy period in an M/M/1, i.e.,

$$T_1 \sim BP^{M/M/1}. \tag{12}$$

Moreover, we know that the cost incurred during this time period is $\overline{Cost}(t_0 + \delta, \alpha)$. Note that there is a discount factor of $e^{-\alpha\delta}$ after phase 1. Thus the expected total discounted cost incurred during phase 2 is

$$e^{-\alpha\delta} \cdot \overline{Cost}(t_0 + \delta, \alpha). \tag{13}$$

Phase 3 (time until the state returns to t_0): At time $T_1 + \delta$, the state drops from being above $t_0 + \delta$ to below. Since the drop is exponential with rate λ (see (4)), we know that the state

12

at time $T_1 + \delta$ is lower than $t_0 + \delta$ by an exponential overshoot, which is $\mathrm{Exp}(\lambda)$. We define the state at $T_1 + \delta$ to be $t_0 + \delta - T_2$, where $T_2 \sim \mathrm{Exp}(\lambda)$.

Now since the policy $Threshold(t_0 + \delta)$ stays passive when the state is smaller than $t_0 + \delta$, phase 3 lasts for $T_2 - \delta$ time before the state returns to t_0 (i.e., the age grows to t_0). Thus the total discounted cost incurred during phase 3 is

$$e^{-\alpha(\delta + T_1)} \cdot \int_0^{T_2 - \delta} (\alpha r(t_0 + \delta - T_2 + t) - \ell)e^{-\alpha t}dt. \tag{14}$$

State

Time

$t_0 \quad t_0 + \delta \qquad\qquad\qquad t_0 + \delta - T_2 \qquad\qquad t_0$

$0 \quad \delta \qquad\qquad\qquad\qquad T_1 + \delta \qquad\qquad T_1 + T_2$
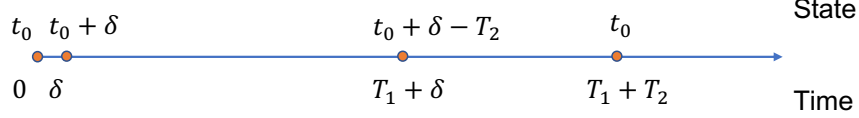
Figure 6: Illustration state transition process.

After three phases, the state returns to t_0. Thus summing up (11), (13), (14) and taking the expectation gives the following equation:

$$\mathrm{Cost}^{t_0+\delta}(t_0, \alpha, \ell) = \delta \cdot (\alpha \cdot r(t_0) - \ell) + o(\delta) \tag{15}$$
$$+ e^{-\alpha\delta} \cdot \overline{Cost}(t_0 + \delta, \alpha) \tag{16}$$
$$+ e^{-\alpha\delta}\mathbf{E}\left[e^{-\alpha T_1}\right] \cdot \mathbf{E}\left[\int_0^{T_2-\delta} (\alpha r(t_0 + \delta - T_2 + t) - \ell)e^{\alpha t}dt\right] \tag{17}$$
$$+ \mathbf{E}\left[e^{-\alpha T_1}\right]\mathbf{E}\left[e^{-\alpha T_2}\right]\mathrm{Cost}^{t_0+\delta}(t_0, \alpha, \ell). \tag{18}$$

We can use the notation from Table 1 to simplify the terms. We can simplify line (17) as follows:

$$\mathbf{E}\left[\int_0^{T_2-\delta} (\alpha r(t_0 + \delta - T_2 + t) - \ell)e^{\alpha t}dt\right]$$
$$= \mathbf{E}\left[\int_0^{T_2} (\alpha r(t_0 + \delta - T_2 + t))e^{-\alpha t}\right] - \mathbf{E}\left[\int_{T_2-\delta}^{T_2} (\alpha r(t_0 + \delta - T_2 + t))e^{-\alpha t}\right] - \ell \cdot \mathbf{E}\left[\int_0^{T_2-\delta} e^{-\alpha t}dt\right]$$
$$= \Gamma(t_0 + \delta, \alpha) - \mathbf{E}\left[\delta\alpha r(t_0)e^{-\alpha T_2} + o(\delta)\right] - \ell \cdot \frac{1}{\alpha}(1 - \mathbf{E}\left[e^{-\alpha(T_2-\delta)}\right])$$
$$= \Gamma(t_0 + \delta, \alpha) - \delta\alpha r(t_0)\gamma_2 + o(\delta) - \ell \cdot \frac{1}{\alpha}(1 - \gamma_2 \cdot e^{\alpha\delta}).$$

Replacing (17) with the above equation, and rearranging terms, gives the proof. $\qquad\square$

We now use Lemma 4.11 to solve equation (6) to get the guessed Whittle Index.

Lemma 4.12. *The guessed Whittle index with discount factor α is given by:*

$$\widehat{W}(t_0, \alpha) = \frac{1}{1 - \gamma_1}\left(\alpha(1 - \gamma_1\gamma_2)r(t_0) + \overline{Cost}'(t_0, \alpha) - \alpha\overline{Cost}(t_0, \alpha) + \gamma_1\Gamma'(t_0, \alpha) - \alpha\gamma_1\Gamma(t_0, \alpha)\right). \tag{19}$$

13

Proof. Note that (6) is equivalent to

$$\left.\frac{d\text{Cost}^{t_0+\delta}(t_0, \alpha, \ell)}{d\delta}\right|_{\delta=0} = 0$$

The left hand side is the derivative of (10) with respect to δ. For simplicity, we omit the $o(\delta)$ term and define

$$\overline{Cost}'(t_0, \alpha) := \left.\frac{d\overline{Cost}(t, \alpha)}{dt}\right|_{t=t_0}, \quad \text{and} \quad \Gamma'(t_0, \alpha) := \left.\frac{d\Gamma(t, \alpha)}{dt}\right|_{t=t_0}.$$

Now the derivative of (10) has the form:

$$\begin{aligned}
\left.\frac{d\text{Cost}^{t_0+\delta}(t_0, \alpha)}{d\delta}\right|_{\delta=0} = {} & \frac{1}{1 - \gamma_1\gamma_2}\Bigg(\alpha r(t_0) - \ell + \overline{Cost}'(t_0, \alpha) - \alpha\overline{Cost}(t_0, \alpha) \\
& + \gamma_1\left(\Gamma'(t_0, \alpha) - \alpha\gamma_2 r(t_0) + \ell \cdot \gamma_2\right) \\
& - \alpha\gamma_1\left(\Gamma(t_0, \alpha) - \ell\frac{1 - \gamma_2}{\alpha}\right)\Bigg).
\end{aligned}$$

Thus, (6) is equivalent to:

$$\alpha(1 - \gamma_1\gamma_2)r(t_0) + \overline{Cost}'(t_0, \alpha) - \alpha\overline{Cost}(t_0, \alpha) + \gamma_1\Gamma'(t_0, \alpha) - \alpha\gamma_1\Gamma(t_0, \alpha) = (1 - \gamma_1) \cdot \ell,$$

which yields:

$$\ell = \frac{1}{1 - \gamma_1}\left(\alpha(1 - \gamma_1\gamma_2)r(t_0) + \overline{Cost}'(t_0, \alpha) - \alpha\overline{Cost}(t_0, \alpha) + \gamma_1\Gamma'(t_0, \alpha) - \alpha\gamma_1\Gamma(t_0, \alpha)\right).$$

Since $\widehat{W}(t_0, \alpha)$ is defined to be the value of ℓ which satisfies (6), this gives the proof. $\qquad\square$

4.3.3 Simplifying the guessed Whittle Index

Lemma 4.12 gives a complicated expression for the guessed Whittle Index. In this subsection, we work on simplifying Equation (19). The final result is an elegant formula for the guessed Whittle Index (see Lemma 4.20). Since the proof is long and intricate, for the readers' convenience, we give a table summarizing each lemma below (Table 2).

The following lemma is a basic formula in ODE. We present it without proof.

Lemma 4.13 (First Order ODE). *For the first order ODE, $y' + P(x)y = Q(x)$, the general solution has the form*

$$y = e^{-\int P(x)\, dx}\left(\int e^{\int P(x)\, dx}Q(x)\, dx + C\right).$$

We also give a basic formula on exponential variables.

Lemma 4.14. *For any smooth function f and exponential variable X, we have that*

$$\mathbf{E}\left[f(x_0 + X)\right] - f(x_0) = \mathbf{E}\left[X\right]\mathbf{E}\left[f'(x_0 + X)\right].$$

14

$$f(x_0) - \mathbf{E}\left[f(x_0 - X)\right] = \mathbf{E}\left[X\right]\mathbf{E}\left[f'(x_0 - X)\right].$$

Proof. Suppose $X \sim \text{Exp}(\theta)$. Then we have that

$$
\begin{aligned}
\mathbf{E}\left[f(x_0 + X)\right] - f(x_0) = \int_0^\infty f'(x_0 + t)\mathbb{P}\left[X > t\right]dt &= \int_0^\infty f'(x_0 + t)e^{-\theta t}dt \\
&= \frac{1}{\theta}\int_0^\infty f'(x_0 + t)\theta e^{-\theta t}dt \\
&= \mathbf{E}\left[X\right]\mathbf{E}\left[f'(x_0 + X)\right].
\end{aligned}
$$

The other equation is derived similarly. $\square$

The next lemma is a necessary condition for the TVHC problem to be well-defined. It states that the instantaneous holding cost $c(t)$ cannot grow too fast. This provides the basis for the proof of Lemma 4.17. We present this lemma with the subscripts i, but will later drop all subscripts.

Lemma 4.15. *If there exists a policy such that the long-run mean holding cost of the TVHC problem converges, then for any type i, we have that*

$$\lim_{t \to \infty} \frac{\int_0^t c_i(x)dx}{e^{(\mu_i - \lambda_i)t}} = 0.$$

Proof. If the long-run mean holding cost converges, then it must also be the case that the long-run mean holding cost converges if there are only type i jobs in the system, and all jobs are served in FCFS order. In this case, the response time of each type i job is the response time in M/M/1, which is distributed $\text{Exp}(\mu_i - \lambda_i)$.

Thus we have that the mean holding cost, which can be computed by the product of the arrival rate and the expected accumulated holding cost per job, has the formula

$$\lambda_i \mathbf{E}\left[\int_0^A c_i(x)dx\right],$$

where $A \sim \text{Exp}(\mu_i - \lambda_i)$. If $\mathbf{E}\left[\int_0^A c_i(x)dx\right]$ converges, it is equal to:

$$\int_0^\infty (\mu_i - \lambda_i)\frac{\int_0^t c_i(x)dx}{e^{(\mu_i - \lambda_i)t}}dt. \tag{20}$$

Thus we know that $\lim_{t\to\infty} \frac{\int_0^t c_i(x)dx}{e^{(\mu_i-\lambda_i)t}} = 0$, otherwise expression (20) does not converge. $\quad\square$

We now return to the single-arm bandit problem. We first need to characterize the cost function r which is defined in (5).

Lemma 4.16 (r')**.** *The derivative of r is given by*

$$r'(t) = c'(t) + \lambda c(t).$$

Proof. Note that in the definition of $r(s)$ (Equation (5)), $\{Y_j\}$ can be viewed as the set of Poisson arrivals with rate λ from time 0 to time s.

A Poisson arrival process from time 0 to $t + \delta$ can be divided into two independent Poisson arrival processes: the one from time 0 to t and another one from time t to $t + \delta$. Thus we have that

$$r(t + \delta) = c(t + \delta) + \mathbf{E}\left[\sum_j c(Y_j)\right] + \mathbf{E}\left[\sum_j c(Y_j')\right], \tag{21}$$

where $\{Y_j\}$ is the Poisson arrivals during $[0, t]$, and $\{Y_j'\}$ is the Poisson arrivals during $[t, t + \delta]$. The expected number of arrivals during $[t, t+\delta]$ is $\lambda\delta$, and the instantaneous holding cost of those jobs is approximately $c(t_0)$. Thus $\mathbf{E}\left[\sum_j c(Y_j')\right] = c(t_0) \cdot \lambda\delta + o(\delta)$. This together with (21) yields:

$$r'(t_0) = \frac{r(t_0 + \delta) - r(t_0)}{\delta} = c'(t_0) + \lambda c(t_0).$$

$\quad\square$

We now derive the expression of $\overline{Cost}(t, \alpha)$, which is defined in Definition 4.7.

Lemma 4.17 (Formula for $\overline{Cost}$)**.** *For any t and $\alpha > 0$,*

$$\overline{Cost}(t, \alpha) = (1 - \gamma_1)\mathbf{E}\left[r(t + X)\right], \qquad where\ X \sim Exp\left(\frac{\alpha}{1 - \gamma_1}\right) \tag{22}$$

Proof. Recall that $\overline{Cost}(t, \alpha)$ is defined to be

$$\overline{Cost}(t, \alpha) := \mathbf{E}\left[\int_0^{T_1} \alpha r(t + A(s))e^{-\alpha s}ds\right],$$

where $A(s)$ is the random variable denoting the age of the oldest job in an M/M/1 queue at time s.

During the first δ time period, a job completes with probability $\mu\delta$. If this happens, the state $A(s)$ drops by an exponential amount ($A(\delta) = \delta - Exp(\lambda)$). There are two cases: (1) with probability $e^{-\lambda\delta}$, the decrement is more than δ, which means $A(\delta) < 0$ and the busy period ends; (2) with probability $1 - e^{-\lambda\delta}$, the dropping amount is smaller than δ and the busy period continues. In this case, suppose $A(\delta) = \Delta < \delta$. Thus, conditioning on a drop happening in the first δ time, we have that $\overline{Cost}$ is, with a slight abuse of notation:

$$\left[\overline{Cost}(t, \alpha)\ |\ \text{a drop happens in the first } \delta \text{ time}\right] = \alpha r(t)\delta + \left(1 - e^{-\lambda\delta}\right)e^{-\alpha\delta}\overline{Cost}(t + \Delta, \alpha) + o(\delta). \tag{23}$$

16

Otherwise, with probability $1 - \mu\delta$, no drop happens during the first δ time. Then $A(s)$ starts at δ, and experiences an M/M/1 busy period like process until it drops below δ. During this time, $\overline{Cost}(t + \delta, \alpha)$ is incurred. At the point that $A(s)$ drops below δ, we know that $A(s)$ is distributed as $\delta - \text{Exp}(\lambda)$ because this is an exponential overshoot. Thus with probability $e^{-\lambda\delta}$, $A(s)$ is below 0, which means the busy period ends. Otherwise with probability $1 - e^{-\lambda\delta}$, $A(s)$ is still larger than 0 (again, denoted by Δ) and the busy period continues. Thus the conditional value of $\overline{Cost}$ in this case is

$$\left[\overline{Cost}(t, \alpha) \mid \text{no drops happen in the first } \delta \text{ time}\right]$$
$$= \alpha r(t)\delta + e^{-\alpha\delta}\left(\overline{Cost}(t + \delta, \alpha) + \mathbf{E}\left[e^{-\alpha T_1}\right](1 - e^{-\lambda\delta})\overline{Cost}(t + \Delta, \alpha)\right) + o(\delta). \quad (24)$$

Combining (23) and (24), we have that

$$\overline{Cost}(t, \alpha) = \alpha r(t)\delta + \mu\delta\left(\left(1 - e^{-\lambda\delta}\right)e^{-\alpha\delta}\overline{Cost}(t + \Delta, \alpha)\right) + o(\delta)$$
$$+ (1 - \mu\delta)\left(e^{-\alpha\delta}\left(\overline{Cost}(t + \delta, \alpha) + \gamma_1(1 - e^{-\lambda\delta})\overline{Cost}(t + \Delta, \alpha)\right)\right).$$

Thus we have that

$$\lim_{\delta\to 0}\frac{\overline{Cost}(t + \delta, \alpha) - \overline{Cost}(t, \alpha)}{\delta}$$
$$= -\alpha r(t) - \lim_{\delta\to 0}\mu\left(\left(1 - e^{-\lambda\delta}\right)e^{-\alpha\delta}\overline{Cost}(t + \Delta, \alpha)\right)$$
$$+ \lim_{\delta\to 0}\frac{1 - (1 - \mu\delta)e^{-\alpha\delta}}{\delta}\overline{Cost}(t + \delta, \alpha) - \lim_{\delta\to 0}(1 - \mu\delta)e^{-\alpha\delta}\gamma_1\frac{1 - e^{-\lambda\delta}}{\delta}\overline{Cost}(t + \Delta, \alpha)$$
$$= -\alpha r(t) + (\mu + \alpha)\overline{Cost}(t, \alpha) - \lambda\gamma_1\overline{Cost}(t, \alpha)$$
$$= -\alpha r(t) + (\mu + \alpha - \lambda\gamma_1)\overline{Cost}(t, \alpha).$$

Note that this is a first-order ODE of the function $\overline{Cost}(t, \alpha)$ with respect to t. By Lemma 4.13,

$$\overline{Cost}(t, \alpha) = e^{(\mu+\alpha-\lambda\gamma_1)t}\left(\int -e^{-(\mu+\alpha-\lambda\gamma_1)s}\alpha r(s)ds + C\right). \quad (25)$$

Note that by (9),

$$\alpha + \mu - \lambda\gamma_1 = \alpha - \lambda\gamma_1 + \frac{(\alpha - \lambda\gamma_1 + \lambda)\gamma_1}{1 - \gamma_1} = \frac{\alpha}{1 - \gamma_1}.$$

Thus (25) can be rewritten as

$$\overline{Cost}(t, \alpha) = e^{\frac{\alpha}{1-\gamma_1}t}\left(\int_t^\infty e^{-\frac{\alpha}{1-\gamma_1}s}\alpha r(s)ds + C\right)$$
$$= \int_0^\infty e^{-\frac{\alpha}{1-\gamma_1}s}\alpha r(t + s)ds + Ce^{\frac{\alpha}{1-\gamma_1}t}$$
$$= (1 - \gamma_1)\mathbf{E}\left[r(t + X)\right] + Ce^{\frac{\alpha}{1-\gamma_1}t}, \quad (26)$$

where $X \sim \text{Exp}(\frac{\alpha}{1-\gamma_1})$. Next we prove that $C = 0$.

By Lemma 4.16, we have that

$$r(t) = c(t) + \lambda \int_0^t c(x)dx + C_1.$$

Together with Lemma 4.15 (dropping the subscripts i), we have that

$$\lim_{t \to 0} \frac{r(t)}{e^{(\mu-\lambda)t}} = 0.$$

From (26) we have that for any t,

$$|C| \leqslant \overline{Cost}(t,\alpha)e^{-\frac{\alpha}{1-\gamma_1}t} + (1 - \gamma_1)\mathbf{E}\left[r(t+X)\right]e^{-\frac{\alpha}{1-\gamma_1}t}.$$

However, both terms on the right hand side go to zero as $t \to \infty$ (see Lemma D.1 in Appendix D). Thus we have proven that $C = 0$. By Equation (26) we have the proof. $\square$

We now characterize another important term in (19), the function Γ.

Lemma 4.18 (Formula for Γ).

$$\Gamma(t,\alpha) = \frac{\alpha}{\alpha + \lambda}\mathbf{E}\left[r(t - T_2)\right]. \tag{27}$$

Proof. The proof is relatively straightforward:

$$\begin{aligned}
\Gamma(t,\alpha) := \mathbf{E}\left[\int_0^{T_2} \alpha r(t - T_2 + s)e^{-\alpha s}ds\right] &= \int_0^\infty \int_0^x \alpha r(t - x + s)e^{-\alpha s}ds\ \lambda e^{-\lambda x}dx \\
&= \int_0^\infty \int_s^\infty \alpha r(t - x + s)e^{-\alpha s}\lambda e^{-\lambda x}dxds \\
&= \int_0^\infty \int_0^\infty \alpha r(t - y)e^{-\alpha s}\lambda e^{-\lambda(y+s)}dyds \\
&= \int_0^\infty r(t - y)\lambda e^{-\lambda y}dy \cdot \int_0^\infty \alpha e^{-(\alpha+\lambda)s}ds \\
&= \frac{\alpha}{\alpha + \lambda}\mathbf{E}\left[r(t - T_2)\right].
\end{aligned}$$

$\square$

Finally, we use the following lemma to simplify (19).

Lemma 4.19. *Let $T_2 \sim \text{Exp}(\lambda)$, $X \sim \text{Exp}(\frac{\alpha}{1-\gamma_1})$. Then we have the following equations:*

$$\mathbf{E}\left[r'(t - T_2)\right] = \lambda c(t). \tag{28}$$

$$\mathbf{E}\left[r(t - T_2)\right] = r(t) - c(t). \tag{29}$$

$$\mathbf{E}\left[r'(t + X)\right] = \frac{\mu}{\gamma_1}\mathbf{E}\left[c(t + X)\right] - \frac{\alpha}{1 - \gamma_1}c(t). \tag{30}$$

$$\mathbf{E}\left[r(t + X)\right] = r(t) - c(t) + \frac{\mu(1 - \gamma_1)}{\gamma_1 \alpha}\mathbf{E}\left[c(t + X)\right]. \tag{31}$$

18

Proof. The equations follow from Lemmas 4.14 and 4.16. See Lemma D.2 in Appendix D. □

Now substituting (22), (27), (28), (29), (30), (31) into the formula given in Lemma 4.12 gives the desired simple guessed Whittle Index.

Lemma 4.20. *The guessed Whittle Index given discount factor α is:*

$$\widehat{W}(t_0, \alpha) = \mu \mathbf{E}\left[c(t_0 + X)\right], \quad \text{where } X \sim Exp(\frac{\alpha}{1 - \gamma_1}). \tag{32}$$

4.4 Indexability and the Whittle Index

We now use the guessed Whittle Index given in Lemma 4.20 to prove indexability and the correctness of the guess. We only present the proof sketch here due to space limitations and defer the proof to the appendix. However, we need to point out that, though deferred, the proof is also involved.

Theorem 2 (Indexability). *The discounted single-arm bandit problem is indexable with the Whittle Index $W(t_0, \alpha) = \widehat{W}(t_0, \alpha)$, where $\widehat{W}$ is the guessed Whittle Index given in Lemma 4.20.*

Proof. The main lemma for this theorem is to prove that when the compensation $\ell = \widehat{W}(t_0, \alpha)$, the policy $Threshold(t_0)$ is the optimal policy. This is proved by verifying the Hamilton–Jacobi–Bellman equation. See Appendix E for the detailed proof. □

Our final result is the Whittle Index for the scheduling problem. It is a straightforward corollary of Theorem 2 by taking the limit on the discount factor and adding back the subscript i.

Corollary 1. *The Whittle Index for the scheduling problem is*

$$W_i(t_0) = \mu \mathbf{E}\left[c_i(t_0 + X)\right], \quad \text{where } X \sim Exp(\mu_i - \lambda_i).$$

Proof. The formula is given by $W(t_0) = \lim_{\alpha \to 0} \widehat{W}(t_0, \alpha)$:

$$\lim_{\alpha \to 0} \frac{\alpha}{1 - \gamma_1} = \lim_{\alpha \to 0} \frac{\alpha}{1 - \widetilde{T_1}(\alpha)} = \frac{1}{-\widetilde{T_1}'(0)} = \frac{1}{\mathbf{E}\left[T_1\right]} = \mu - \lambda.$$

Adding back the subscript i yields the Whittle Index. □

5 Simulations

We conduct simulations to evaluate the performance of our proposed policy (from Corollary 1), comparing it against other policies in the literature. In our evaluations, we experiment with different numbers of classes, different holding cost functions, and a range of system loads. We present the following main findings:

- Across all experiments, including cases with complex interleaving holding cost functions, our policy consistently matches the lowest time-average holding cost among all candidate policies we considered.

- In certain problem settings, our policy significantly outperforms each candidate policy.

5.1 Policies evaluated

Throughout, when we talk about "our policy," we will refer to the policy from Corollary 1 where the priority of a class i job of age t is

$$W_i(t) = \mu_i \mathbf{E}\left[c_i(t+X)\right], \qquad \text{where } X \sim \text{Exp}(\mu_i - \lambda_i)$$

We compare our policy with a broad spectrum of alternative policies:

FCFS. This policy always serves the job that arrived earliest. FCFS is a very simple policy and we compare against it as a baseline.

Strict Preemptive Priority. This policy assigns a fixed priority to each job class, where jobs from a higher-priority class have preemptive priority over those from a lower-priority class. Jobs within a class are run in FCFS order. We choose a highly optimistic version of this policy where the priority ordering can change at each value of system load. Thus we might run Prio(1;2), where class 1 has priority over class 2 when load is low, but Prio(2;1) when load is high.

Generalized $c\mu$ [24]. This policy always serves the job with highest index $c_i(t) \cdot \mu_i$, where t is the age of the class i job. This policy is known to be optimal in the diffusion limit.

Aalto's Index [1]. In this policy a class i job of age t is given index $V_i(t)$, where[4]

$$V_i(t) = \mu_i \mathbf{E}\left[c_i(t+S)\right], \qquad \text{where } S \sim Exp(\mu_i).$$

Like our policy, this is also a Whittle-index based heuristic, but is motivated by the static setting (no arrivals), and hence does not incorporate the arrival rate λ_i.

5.2 Experimental Results

We have conducted hundreds of experiments comparing the scheduling policies from Section 5.1 under different holding cost functions, job sizes, and arrival rates. In all of these, our policy was either the best policy or matched the best policy. Due to lack of space, we present only four diverse experiments which well-illustrate some important behaviors.

Each experiment shown is represented by *(a)* a set of holding cost functions (not drawn to scale) and *(b)* the corresponding simulation results. We experiment with small and large jobs, where holding cost functions drawn in red or orange correspond to a class with shorter jobs, while those drawn in blue correspond to a class with larger jobs. We experiment with different arrival rates, shown as a proportion of the total arrival rate, λ, across all classes, where λ is specified by the load.
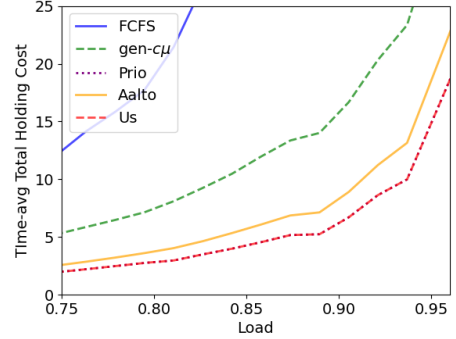
Figure 7a considers an experiment where class 1 jobs (the shorter ones) only incur a holding cost after a deadline is passed, while class 2 jobs (the longer ones) incur a steady (but low) holding cost. In this figure, the arrival rate of class 1 is much higher than that of class 2. Figure 7b shows the corresponding results. We see that FCFS is by far the worst policy. The generalized $c\mu$ rule improves upon FCFS, and Aalto improves upon that. Our policy (shown in red) is significantly better than all the others, except for Preemptive Priority, which is equal to our policy here.

---

[4]This is a simplification of Aalto's policy to the case of exponential job sizes.

To understand what's going on, we first observe that we should be prioritizing class 1 jobs ahead of their deadline, given that the cost of missing the deadline is so high. The $c\mu$ rule only prioritizes class 1 at the deadline point. The Aalto policy improves upon the $c\mu$ rule, by prioritizing class 1 jobs in advance of the deadline, but it doesn't do so early enough (because Aalto doesn't consider load). Our policy improves upon the Aalto policy by prioritizing class 1 jobs even earlier – in fact a better time to start prioritizing class 1 jobs is at age 0, which is what the Preemptive Priority policy does.
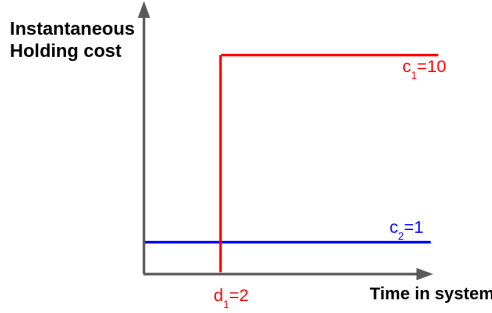
(a) Holding cost functions.
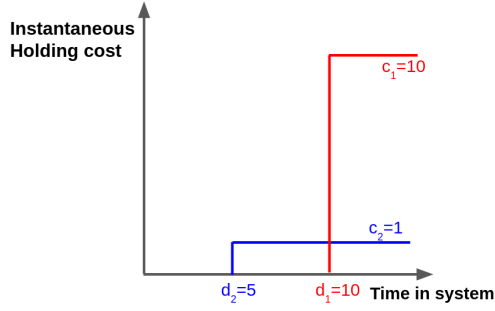
(b) Performance of policies.

Figure 7: Comparison of policies on holding cost functions with one deadline. We fix $\mu_1 = 3, \mu_2 = 1, \lambda_1 = 0.9\lambda$.

Figure 8a considers an experiment where both class 1 jobs (the short ones) and class 2 jobs (the long ones) only incur holding costs after their respective deadlines are passed. Here class 1 jobs incur a high penalty after a late deadline, while class 2 jobs incur a low penalty after an early deadline. The two classes have equal arrival rates. Figure 8b shows the corresponding results. We see that our policy is noticeably better than all the others, except for Aalto's policy, which we only negligibly dominate. Our policy noticeably dominates Preemptive-Priority, which noticeably dominates generalized $c\mu$ which significantly dominates FCFS.

To understand what's going on, observe that it is again useful to prioritize jobs *ahead* of their deadlines, which explains why we outperform FCFS and generalized $c\mu$. Strict priority is no longer effective here, because the deadlines are further out, so always prioritizing the class 1 jobs can be suboptimal. Our policy's performance is similar to Aalto for two reasons: First, because the arrival rates are balanced, we find that when load is not too high, λ_i is small compared to μ_i, and our policy is similar to Aalto's policy. Second, while our policy looks further into the future, it does so for both classes, and hence the benefit in some sense "cancels out," leaving us with a policy similar to Aalto.

Figure 9a considers an experiment where jobs have polynomial holding cost functions. Specifically, class 1 jobs (the shorter ones) have a linear holding cost, while class 2 jobs (the longer ones) have a quadratic holding cost. We set the arrival rates such that the loads from both classes are balanced. Figure 9b shows the corresponding results. We see that our policy significantly outperforms FCFS and Preemptive-Priority; however, our policy only negligibly beats generalized $c\mu$ and Aalto.

To understand why the three dominant policies (our policy, generalized $c\mu$ and Aalto) are similar, we note that, for the polynomial holding cost functions, these three policies have index functions which are polynomials with the same leading coefficient. Consequently, their behaviors are similar.
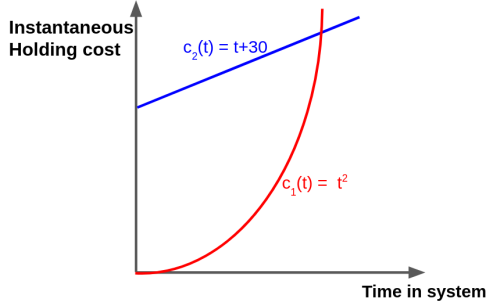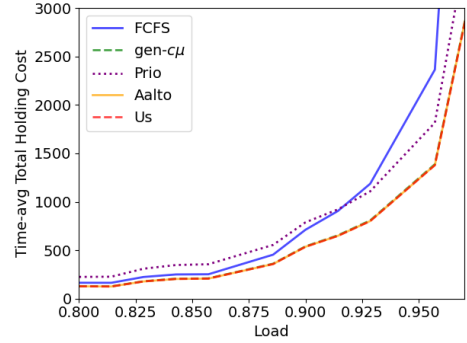
(a) Holding cost functions.
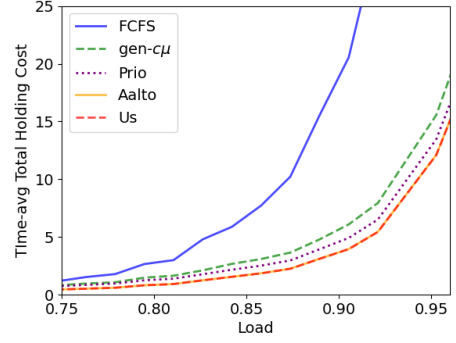
(b) Performance of policies.

Figure 8: Comparison of policies on holding cost functions with deadlines. We fix $\mu_1 = 3, \mu_2 = 1, \lambda_1 = 0.5\lambda$.

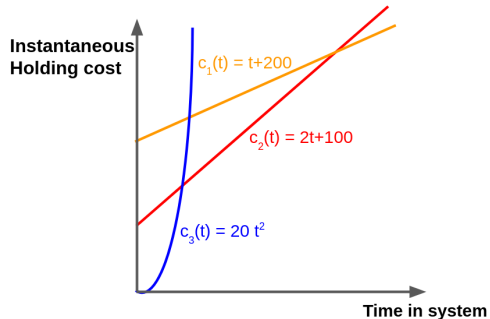(a) Holding cost functions.

(b) Performance of policies.

Figure 9: Comparison of policies on polynomial holding cost functions. We fix $\mu_1 = 3, \mu_2 = 1, \lambda_1 = 0.75\lambda$.

We often find that these three policies perform similarly and vastly improve upon the other policies.

Lastly, Figure 10a considers the case with three job classes. Specifically, class 1 and class 2 jobs (shorter jobs) have linear interleaving holding cost functions, while class 3 jobs (longer jobs) have explosive quadratic holding cost. Figure 10b shows the corresponding results. We see that our policy noticeably improves upon Aalto, which noticeably improves upon generalized $c\mu$, which noticeably improves upon Preemptive-Priority, which significantly improves upon FCFS.

The strict ordering of policies, depicted in Figure 10b is typical of what we see in many experiments, regardless of the number of classes. This strict separation would also be more obvious in Figure 9a if we set both holding costs to be higher. When the costs are higher, the small differences between the policies are more amplified.

Another policy which we experimented with is the Accumulated Priority policy from [23, 10]. However we found that this policy never improved upon our policy in our experiments and was often significantly worse than our policy, so we chose to omit it to keep the graphs cleaner.

(a) Holding cost functions.
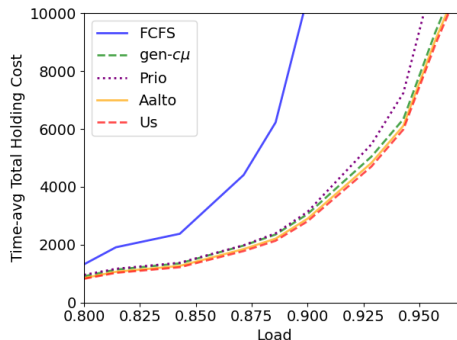
(b) Performance of policies.

Figure 10: Comparison of policies on 3 classes. We fix $\mu_1 = \mu_2 = 3, \mu_3 = 1, \lambda_1 = \lambda_2 = \lambda_3$.

6 Discussion on Optimality

Our policy replicates the diffusion limit optimality of the generalized $c\mu$ rule [24]. The diffusion limit regime in [24] is equivalent to $\lambda_i = \Theta(n), \mu_i = \Theta(n), \frac{\lambda_i}{\mu_i} = 1 - \Theta(\frac{1}{\sqrt{n}})$ where $n \to \infty$. Thus we have that $\mu_i - \lambda_i \to \infty$. This indicates that our policy degenerates to the generalized $c\mu$ rule, and hence is optimal as well.

We have seen that our policy performs well in simulation. However, like all the other Whittle-based heuristics we've discussed, our policy is not always optimal. We now give a counter example.

Assume that there are two classes of jobs, both with *the same* instantaneous holding cost function $c(t) = t$. Both classes have the same completion rate, but the arrival rates are different. Since the holding cost functions and the completion rates are the same in both classes, the optimal policy is FCFS, which is clearly not our policy. Thus, while our Whittle Index policy performs really well compared with other existing policies, there is still a need for further work on optimality.

7 Conclusion

This paper studies the classical TVHC problem: Jobs of different classes arrive over time, where each class of jobs is associated with a holding cost that increases with the job's age. The objective is to schedule the jobs so as to minimize the expected time-average total holding cost.

Various papers have provided heuristics for the TVHC problem. The generalized $c\mu$ rule [24] provides a simple index policy which favors jobs with currently high holding cost and high failure rate; this policy is asymptotically optimal in the diffusion limit. More recent works by Aalto [1], consider the more tractable static setting (no arrivals).

Our work, while also heuristic, takes a principled approach to the TVHC problem with arrivals: (i) We derive the first representation of our problem as an R-MAB with a *finite* number of arms, and (ii) We derive a novel Whittle index policy for the resulting R-MAB. While the analysis is involved, the resulting policy is extremely simple and elegant and incorporates class load. In simulation, our policy improves upon all the other known heuristics.

This story is in no way finished. First, our policy might be generalizable beyond exponential job

size distributions, to those with increasing hazard rate, using our existing R-MAB as a foundation (although this would require a more complex state space). Second, there is much more work needed to find an *optimal* policy.

References

[1] S. Aalto. Whittle index approach to the multi-class queueing systems with convex holding costs and IHR service times. *Mathematical Methods of Operations Research*, 100(3):603–634, 2024.

[2] A. Anand and G. de Veciana. A Whittle's index based approach for QoE optimization in wireless networks. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):1–39, 2018.

[3] P. Ansell, K. D. Glazebrook, J. Nino-Mora, and M. O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57:21–39, 2003.

[4] R. Atar, C. Giat, and N. Shimkin. The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research*, 58(5):1427–1439, 2010.

[5] R. Atar, A. Mandelbaum, and M. I. Reiman. Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 14(3), Aug. 2004.

[6] U. Ayesta, M. K. Gupta, and I. M. Verloop. On the computation of Whittle's index for Markovian restless bandits. *Mathematical Methods of Operations Research*, 93:179–208, 2021.

[7] C. F. Bispo. The single-server scheduling problem with convex costs. *Queueing Systems*, 73:261–294, 2013.

[8] C. Buyukkoc, P. Varaiya, and J. Walrand. The c-mu rule revisited. *Advances in Applied Probability*, 17(1):237–238, 1985.

[9] D. R. Cox. *Queues*. Chapman and Hall/CRC, 2020.

[10] V. A. Fajardo and S. Drekic. Waiting time distributions in the preemptive accumulating priority queue. *Methodology and Computing in Applied Probability*, 19:255–284, 2017.

[11] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):148–164, 1979.

[12] K. D. Glazebrook, R. Lumley, and P. Ansell. Index heuristics for multiclass M/G/1 systems with nonpreemptive service and convex holding costs. *Queueing Systems*, 45:81–111, 2003.

[13] I. Gurvich and W. Whitt. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management*, 11(2):237–253, 2009.

[14] M. Harchol-Balter. *Performance modeling and design of computer systems: queueing theory in action*. Cambridge University Press, 2013.

[15] M. Larrañaga, U. Ayesta, and I. M. Verloop. Index policies for a multi-class queue with convex holding cost and abandonments. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 125–137, 2014.

[16] 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, 2016.

[17] Z. Long, N. Shimkin, H. Zhang, and J. Zhang. Dynamic scheduling of multiclass many-server queues with abandonment: The generalized $c\mu/h$ rule. *Operations Research*, 68(4):1218–1230, 2020.

[18] A. Mandelbaum and A. L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research*, 52(6):836–855, 2004.

[19] J. Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *Top*, 15:161–198, 2007.

[20] J. Niño-Mora. Markovian restless bandits and index policies: A review. *Mathematics*, 11(7):1639, 2023.

[21] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[22] Z. Scully and M. Harchol-Balter. The gittins policy in the M/G/1 queue, 2021.

[23] D. A. Stanford, P. Taylor, and I. Ziedins. Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77:297–330, 2014.

[24] J. A. Van Mieghem. Dynamic scheduling with convex delay costs: The generalized c— mu rule. *The Annals of Applied Probability*, pages 809–833, 1995.

[25] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

A Tutorial on The Whittle Index

The Whittle Index approach was introduced by Whittle in 1988 to develop a heuristic for R-MAB problems ([25]), especially for discounted R-MAB problems. There are recent works proving asymptotic optimality of the Whittle Index (e.g., [6]); even when not optimal, the Whittle Index has long been shown to be a surprisingly good heuristic (see [20] for a recent review).

The standard Whittle approach (for discounted R-MAB problems) consists of three steps:

1. First, the hard constraint that only one arm can be pulled at a time is relaxed to a time-average constraint. Mathematically, suppose the action for arm i at time step t is $a_i(t)$, where $a_i(t) = 1$ if the action is active and $a_i(t) = 0$ otherwise. The hard constraint is

$$\sum_i a_i(t) \leqslant 1, \ \forall t.$$

Whittle relaxes the hard constraint to the following time-average constraint:

$$\sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{k} a_i(t) \leqslant \frac{1}{1-\alpha},$$

where α is the discount factor. Through this relaxation, the optimization problem becomes more tractable. The relaxed R-MAB is studied to obtain a heuristic for the original R-MAB.

2. Next, a Lagrange method is applied to the relaxed R-MAB. Crucially the Lagrange dual function can be decomposed into k independent parts, and accordingly the relaxed R-MAB problem can be decomposed to k independent *single-arm restless bandit problems*, where not pulling the arm earns a compensation of ℓ: ℓ serves as the Lagrange multiplier. This step effectively puts a "price" on choosing to keep an arm active, yielding k single-arm optimization problems, each parameterized by ℓ.

3. Then, one solves each single-arm problem to figure out the critical value of ℓ for each state of each problem, which is the exact point where one is indifferent between being active and passive. This critical value of ℓ measures the "value" of pulling this arm at each state, and serves as the Whittle Index.

4. Finally, the Whittle Index policy pulls the arm with the largest Whittle Index at every moment of time.

In the literature, people usually skip the first two steps and directly analyze the single-arm bandit problem (e.g., [2, 1].) In this paper, we will likewise omit the first two steps and directly provide the single-arm bandit formulation in Section 4.1.

A property of the single-arm bandit problem called "indexability" needs to be established for the Whittle Index to be well-defined. Intuitively, indexability states that as the compensation ℓ grows larger, the optimal action at a certain state can only change from active to passive. This property ensures that the Whittle Index is unique and well-defined. Although indexability is intuitive, it must be proven [25].

B Proof for Lemma 3.1

Lemma B.1 (FCFS within each type is optimal)**.** *The optimal policy must serve jobs within each type in FCFS order.*

Proof. For any policy π, we define a policy π_{FCFS} where π_{FCFS} serves a type i job if and only if π serves a type i job, but π_{FCFS} always serves FCFS within each type. We will show that the time-average total holding cost under π_{FCFS} is no higher than that of π on every sample path.

We now show how to define the sample path that allows us to compare between π and π_{FCFS}. It is important to observe that the result of this lemma does not hold for the obvious sample path, which is based on job arrival times and sizes (think about SRPT within each class, which can improve upon FCFS within each class). Thus we need to define our sample path to leverage the fact that job sizes are memoryless.

We define the sample path as 2 sequences of exponential random variables for each of the k classes. The first sequence for class i represents inter-arrival times of class i jobs. These are clearly distributed as $\text{Exp}(\lambda_i)$. The second sequence for class i represents inter-departure times of class i jobs, where we only look at the system when a class i job is being served. One can imagine that there is an $\text{Exp}(\mu_i)$ timer for class i which is *paused* when the system is working on a job of some type other than class i, and *resumes* when the system returns to working on a class i job. Whenever the timer goes off, the class i job currently being served completes. Note that there might be multiple type i jobs running during the time before the timer goes off.

We now consider the behavior of policies π and π_{FCFS} under a fixed sample path that consists of the $2k$ exponential sequences. Note that π_{FCFS} works on a type i job if and only if π works on a type i job. Thus the exponential timers go off at the same time under both policies. Suppose the timer for class i goes off at times $\{d_1 < d_2 < ...\}$, at which times a type i job completes and departs. Also let the jth class i job's arrival time be denoted by a_j.

Given that π_{FCFS} serves type i jobs in FCFS order, we have that the jth type i job leaves at time d_j. Define $p(j)$ to be a permutation such that $d_{p(j)}$ is the departure time of the jth type i job under π. Now we look at the total holding cost incurred by policies π and π_{FCFS}.

For any type i, the total holding cost of π_{FCFS} incurred by the first N type i jobs is

$$\sum_{j=1}^{N} \int_0^{d_j - a_j} c_i(t)dt, \tag{33}$$

while that of π is

$$\sum_{j=1}^{N} \int_0^{d_{p(j)} - a_j} c_i(t)dt. \tag{34}$$

Observe that (33) is smaller than (34) by the monotonocity and convexity of the function $\int_0^x c_i(t)dt$ with respect to x. This completes the proof. □

C   Proof for Theorem 1

Theorem C.1 (Equivalence of TVHC and R-MAB Problems). *For any set of non-decreasing index functions $\{V_i(\cdot)\}_{i=1}^k$, the corresponding index policies (breaking ties by FCFS) in the TVHC problem and the R-MAB problem incur the same cost.*

Proof. In this appendix, we construct a coupling between the TVHC problem and the R-MAB problem on a certain sample path.

Define a sample path to be $2k$ sequences of exponential variables: $\{a_{i0}, a_{i1}, a_{i2}, ...\}$ where $a_{ij} \sim \text{Exp}(\lambda_i)$ and $\{s_{i1}, s_{i2}, ...\}$ where $s_{ij} \sim \text{Exp}(\mu_i)$. Under this sample path, the index policies $\{V_i\}$ run as follows:

1. *The TVHC Problem:* The first type i job enters at time a_{i0}. The inter-arrival time of type i jobs are a_{ij}. The jth type i job has size s_{ij}.

2. *The R-MAB Problem:* The arm i is initialized at state $-a_{i0}$. After the arm is active for s_{ij} time, it triggers the jth state drop, and the dropping amount is a_{ij}.

Under a fixed sample path, define $T_i(t)$ to be the state of the arm i in the bandit problem, and define $A_i(t)$ to be the age of the oldest type i job in the system in the scheduling problem (if there is no type i job, $A_i(t) < 0$ and $-A_i(t)$ is the time until the next arrival of a type i job). We next prove that $T_i(t) = A_i(t)$ for all t.

To prove this fact, we define $\xi_i(t)$ in the bandit problem to be the total time that arm i is active since the last time the state of arm i dropped. We also define a counter, $count_i^B(t)$, to be the number of state drops of arm i by time t. We define $\tau_i(t)$ in the scheduling problem to be the total service time that the oldest type i has received at time t. We also define a counter, $count_i^S(t)$, to be the number of type i job completions by time t.

Then in the bandit problem, the triple $(T_i(t), \xi_i(t), count_i^B(t))$ is initialized to be $(-a_{i0}, 0, 0)$ and

has the following transition function at time t:

$$\text{If } i = \arg\max_j \{V_j(T_j(t))\} \text{ and } T_i(t) \geq 0:$$

$$\text{If } \xi_i(t) < s_{i,count_i^B(t)}:$$
$$T_i(t+\delta) = T_i(t) + \delta$$
$$\xi_i(t+\delta) = \xi_i(t) + \delta$$
$$count_i^B(t+\delta) = count_i^B(t)$$
$$\text{Else:}$$
$$T_i(t+\delta) = T_i(t) + \delta - a_{i,count_i^B(t)}$$
$$\xi_i(t+\delta) = 0$$
$$count_i^B(t+\delta) = count_i^B(t) + 1$$
$$\text{Else:}$$
$$T_i(t+\delta) = T_i(t) + \delta$$
$$\xi_i(t+\delta) = \xi_i(t)$$
$$count_i^B(t+\delta) = count_i^B(t)$$

$$(35)$$

While we're in the TVHC problem, the triple $(A_i(t), \tau_i(t), count_i^S(t))$ follows the same exact transition function. In this way we prove that $A_i(t) = T_i(t)$ under a fixed sample path.

Now we prove that the expected cost under the two problems is the same. For the TVHC problem, the long-run average holding cost of type i is the expected total type i holding cost at a uniformly random time t^*, which is

$$\sum_{j:\text{ type } i \text{ job in the system at } t^*} c_i(a(j)),$$

where $a(j)$ is the age of the job j. Denote the oldest age among all type i jobs in the system by $A_i(t^*)$. Since the policy runs FCFS within each type, the other type i jobs' inter-arrival times are independent of $A_i(t^*)$, which means $\{a(j) : j \text{ is a type } i \text{ job in the system}\}$ is distributed as $Pois_{\lambda_i}(A_i(t^*))$. Thus we have that

$$\mathbf{E}\left[\sum_{j:\text{ type } i \text{ job in the system at } t^*} c_i(a(j)) \right] = r_i(A_i(t^*)).$$

Note that A_i has the same distribution as T_i due to the coupling sample path argument above. Therefore we have that the expected cost in the two problems are the same. $\qquad \square$

D   Deferred proofs in Section 4.3.3

Lemma D.1. *When $t \to \infty$, $\lim_{t\to\infty} \overline{Cost}(t,\alpha)e^{-\frac{\alpha}{1-\gamma_1}t} = 0$ and $\lim_{t\to\infty} \mathbf{E}\left[r(t+X)\right]e^{-\frac{\alpha}{1-\gamma_1}t} = 0$.*

Proof.

$$\lim_{t\to\infty} \overline{Cost}(t,\alpha)e^{-\frac{\alpha}{1-\gamma_1}t} \quad = \quad \lim_{t\to\infty} \mathbf{E}\left[\int_0^{T_1} \alpha e^{-\alpha s} r(t+A(s))ds\right]e^{-\frac{\alpha}{1-\gamma_1}t}$$

$$\leqslant \quad \lim_{t\to\infty} \mathbf{E}\left[\int_0^{T_1} \alpha r(t+A(s))ds\right]e^{-\frac{\alpha}{1-\gamma_1}t}$$

$$= \quad \lim_{t\to\infty} \alpha\frac{1}{\mu-\lambda}\mathbf{E}\left[r(t+A)\right]e^{-\frac{\alpha}{1-\gamma_1}t} \qquad \text{where } A \sim \text{Exp}(\mu-\lambda)^5$$

$$\overset{(*)}{\leqslant} \quad \lim_{t\to\infty} \frac{\alpha}{\mu-\lambda}\mathbf{E}\left[r(t+A)\right]e^{-(\mu-\lambda)t}$$

$$= \quad 0.$$

$$\lim_{t\to\infty} \mathbf{E}\left[r(t+X)\right]e^{-\frac{\alpha}{1-\gamma_1}t} \quad = \quad \lim_{t\to\infty} \int_0^\infty r(t+x)e^{-\frac{\alpha}{1-\gamma_1}t}\frac{\alpha}{1-\gamma_1}e^{\frac{\alpha}{1-\gamma_1}x}dx$$

$$= \quad \lim_{t\to\infty} \int_0^\infty r(t+x)e^{-\frac{\alpha}{1-\gamma_1}(t+x)}\frac{\alpha}{1-\gamma_1}e^{\frac{\alpha}{1-\gamma_1}2x}dx$$

$$\overset{(*)}{=} \quad \lim_{t\to\infty} \frac{1}{2}\int_0^\infty r(t+x)e^{-(\mu-\lambda)(t+x)}\frac{2\alpha}{1-\gamma_1}e^{\frac{\alpha}{1-\gamma_1}2x}dx$$

$$= \quad \lim_{t\to\infty} e^{-(\mu-\lambda)t}\frac{1}{2}\int_0^\infty r(t+x)e^{-(\mu-\lambda)x}\frac{2\alpha}{1-\gamma_1}e^{\frac{\alpha}{1-\gamma_1}2x}dx$$

$$= \quad 0.$$

where the steps $(*)$ are obtained by using

$$\frac{\alpha}{1-\gamma_1} = \frac{\alpha}{1-\mathbf{E}\left[e^{-\alpha T_1}\right]} \geqslant \frac{1}{\mathbf{E}\left[T_1\right]} = \mu-\lambda,$$

and the inequality is obtained by $1-e^{-\alpha x} \leqslant \alpha x$. $\qquad\qquad\square$

Lemma D.2. *Let $T_2 \sim Exp(\lambda)$, $X \sim Exp(\frac{\alpha}{1-\gamma_1})$. Then we have the following equations:*

$$\mathbf{E}\left[r'(t-T_2)\right] = \lambda c(t). \tag{36}$$

$$\mathbf{E}\left[r(t-T_2)\right] = r(t) - c(t). \tag{37}$$

$$\mathbf{E}\left[r'(t+X)\right] = \frac{\mu}{\gamma_1}\mathbf{E}\left[c(t+X)\right] - \frac{\alpha}{1-\gamma_1}c(t). \tag{38}$$

$$\mathbf{E}\left[r(t+X)\right] = r(t) - c(t) + \frac{\mu(1-\gamma_1)}{\gamma_1\alpha}\mathbf{E}\left[c(t+X)\right]. \tag{39}$$

Proof. We prove the four equations by using Lemma 4.14 and Lemma 4.16:

$$\mathbf{E}\left[r'(t-T_2)\right] = \mathbf{E}\left[c'(t-T_2)\right] + \lambda\mathbf{E}\left[c(t-T_2)\right] \qquad\qquad \textit{Lemma 4.16}$$

$$= \lambda(c(t) - \mathbf{E}\left[c(t-T_2)\right]) + \lambda\mathbf{E}\left[c(t-T_2)\right] \qquad \textit{Lemma 4.14}$$

$$= \lambda c(t).$$

---

[5] A denotes the age of the oldest job in an M/M/1 queue, and $A \sim \text{Exp}(\mu-\lambda)$ given $A > 0$.

$$\mathbf{E}\left[r(t-T_2)\right] = r(t) - \frac{1}{\lambda}\mathbf{E}\left[r'(t-T_2)\right] = r(t) - c(t).$$

$$\begin{aligned}
\mathbf{E}\left[r'(t+X)\right] &= \mathbf{E}\left[c'(t+X)\right] + \lambda\mathbf{E}\left[c(t+X)\right] && \textit{Lemma } 4.16 \\
&= \frac{\alpha}{1-\gamma_1}\left(\mathbf{E}\left[c(t+X)\right] - c(t)\right) + \lambda\mathbf{E}\left[c(t+X)\right] && \textit{Lemma } 4.14 \\
&= \frac{\mu}{\gamma_1}\mathbf{E}\left[c(t+X)\right] - \frac{\alpha}{1-\gamma_1}c(t) && \textit{Equation } (9).
\end{aligned}$$

$$\mathbf{E}\left[r(t+X)\right] = r(t) + \frac{1-\gamma_1}{\alpha}\mathbf{E}\left[r'(t+X)\right] = r(t) - c(t) + \frac{\mu(1-\gamma_1)}{\gamma_1\alpha}\mathbf{E}\left[c(t+X)\right].$$

$\square$

E   Proof for indexability

The goal of this appendix section is to prove Theorem E.3. We need several lemmas for the proof.

We first characterize the optimal policy. Specifically, we show that the policy $Threshold(t_0)$ is optimal when the compensation $\ell = \widehat{W}(t_0, \alpha)$. This is proved by showing the Hamilton–Jacobi–Bellman equation is satisfied.

Lemma E.1. *The optimal policy under compensation $\widehat{W}(t_0, \alpha)$ is the policy $Threshold(t_0)$.*

Proof. It suffices to prove that the policy $Threshold(t_0)$ satisfies the Hamilton–Jacobi–Bellman equation. Let $V(t)$ denote the value function of the policy at state t.

At state t, if during the next δ time the action is passive, the total cost is

$$passive(t) := (\alpha r(t) - \ell)\delta + e^{-\alpha\delta}V(t+\delta). \tag{40}$$

Moreover, if the action is active, the total cost is

$$active(t) := (\alpha r(t))\delta + e^{-\alpha\delta}\left((1-\mu\delta)V(t+\delta) + \mu\delta\mathbf{E}\left[V(t+\delta-T_2)\right]\right), \tag{41}$$

where $T_2 \sim \text{Exp}(\lambda)$. Thus the Hamilton–Jacobi–Bellman equation is equivalent to

$$t \leqslant t_0 \Rightarrow \lim_{\delta\to 0}\frac{passive(t) - active(t)}{\delta} \leqslant 0. \tag{42}$$

$$t \geqslant t_0 \Rightarrow \lim_{\delta\to 0}\frac{passive(t) - active(t)}{\delta} \geqslant 0. \tag{43}$$

By (40) and (41), it is further equivalent to

$$t \leqslant t_0 \Rightarrow \ell \geqslant \mu\left(V(t) - \mathbf{E}\left[V(t-T_2)\right]\right). \tag{44}$$

$$t \geqslant t_0 \Rightarrow \ell \leqslant \mu\left(V(t) - \mathbf{E}\left[V(t-T_2)\right]\right). \tag{45}$$

Case 1: $t \leqslant t_0$. In this case, we want to prove (44).

Note that the policy stays passive when the state is smaller then t_0. Thus we have that

$$
\begin{aligned}
V(t) &= \int_0^{t_0-t} (\alpha r(t+s) - \ell)e^{-\alpha s}ds + e^{-\alpha(t_0-t)}V(t_0) \\
&= \int_0^{t_0-t} \alpha r(t+s)e^{-\alpha s}ds + e^{-\alpha(t_0-t)}V(t_0) - (1 - e^{-\alpha(t_0-t)})\frac{\ell}{\alpha}.
\end{aligned}
\tag{46}
$$

$$
\begin{aligned}
\mathbf{E}\left[V(t - T_2)\right] &= \mathbf{E}\left[\int_0^{T_2} (\alpha r(t - T_2 + s) - \ell)e^{-\alpha s}ds\right] + \mathbf{E}\left[e^{-\alpha T_2}\right]V(t) \\
&= \Gamma(t, \alpha) - \frac{\ell}{\alpha}(1 - \gamma_2) + \gamma_2 V(t).
\end{aligned}
\tag{47}
$$

Thus, using (32), (47),(46), we have that (44) is equivalent to

$$
\mathbf{E}\left[c(t_0 + X)\right] \geqslant (1 - \gamma_2)\left(\int_0^{t_0-t} \alpha r(t+s)e^{-\alpha s}ds + e^{-\alpha(t_0-t)}(V(t_0) + \frac{\ell}{\alpha})\right) - \Gamma(t, \alpha).
\tag{48}
$$

Using (10), (22), (27), (29), (30) and (32), we have that

$$
\begin{aligned}
V(t_0) &= \frac{1}{1 - \gamma_1\gamma_2}\left(\overline{Cost}(t_0, \alpha) + \gamma_1\Gamma(t_0, \alpha) - \frac{\ell}{\alpha}\gamma_1(1 - \gamma_2)\right) &&\text{(10)} \\
&= \frac{1}{1 - \gamma_1\gamma_2}\left((1 - \gamma_1)\mathbf{E}\left[r(t_0 + X)\right] + \gamma_1(1 - \gamma_2)\mathbf{E}\left[r(t_0 - T_2)\right] - \frac{\ell}{\alpha}\gamma_1(1 - \gamma_2)\right) &&\text{by } (22), (27) \\
&= \frac{1}{1 - \gamma_1\gamma_2}\left((1 - \gamma_1)\left(r(t_0) - c(t_0) + \frac{\mu(1 - \gamma_1)}{\gamma_1\alpha}\mathbf{E}\left[c(t_0 + X)\right]\right)\right. &&\text{by } (31) \\
&\quad \left. + \gamma_1(1 - \gamma_2)(r(t_0) - c(t_0)) - \frac{\ell}{\alpha}\gamma_1(1 - \gamma_2)\right) &&\text{by } (29) \\
&= \frac{1}{1 - \gamma_1\gamma_2}\left((1 - \gamma_1)\left(r(t_0) - c(t_0) + \frac{1 - \gamma_1}{\gamma_1\alpha}\ell\right)\right. &&\text{by } (32) \\
&\quad \left. + \gamma_1(1 - \gamma_2)(r(t_0) - c(t_0)) - \frac{\ell}{\alpha}\gamma_1(1 - \gamma_2)\right) \\
&= r(t_0) - c(t_0) + \frac{1}{1 - \gamma_1\gamma_2}\left(\frac{(1 - \gamma_1)^2}{\gamma_1} - \gamma_1(1 - \gamma_2)\right)\frac{\ell}{\alpha} \\
&= r(t_0) - c(t_0) - \frac{\ell}{\alpha} + \frac{1 - \gamma_1}{\gamma_1(1 - \gamma_1\gamma_2)}\frac{\ell}{\alpha} \\
&= r(t_0) - c(t_0) - \frac{\ell}{\alpha} + \frac{\lambda + \alpha}{\mu}\frac{\ell}{\alpha}. &&\text{by } (9)
\end{aligned}
\tag{49}
$$

Thus again using (9), we have that (48) is equivalent to

$$\mathbf{E}\left[c(t_0 + X)\right] \geqslant (1 - \gamma_2) \left(\int_0^{t_0 - t} \alpha r(t + s) e^{-\alpha s} ds + e^{-\alpha(t_0 - t)} \left(r(t_0) - c(t_0) + \frac{1}{1 - \gamma_2} \frac{\ell}{\mu} \right) \right) - \Gamma(t, \alpha).$$

Using (27), (29), (9) and reordering the terms, the inequality is further equivalent to

$$\left(1 - e^{-\alpha(t_0 - t)} \right) \mathbf{E}\left[c(t_0 + X)\right] + (1 - \gamma_2) \left(-e^{-\alpha(t_0 - t)}(r(t_0) - c(t_0)) + r(t) - c(t) \right)$$
$$\geqslant (1 - \gamma_2) \int_0^{t_0 - t} \alpha r(t + s) e^{-\alpha s} ds. \tag{50}$$

Note that the right hand side of (50) can be simplified to:

$$(1 - \gamma_2) \int_0^{t_0 - t} \alpha r(t + s) e^{-\alpha s} ds$$

$$= (1 - \gamma_2) \left(-r(t + s) e^{-\alpha s} \Big|_0^{t_0 - t} + \int_0^{t_0 - t} r'(t + s) e^{-\alpha s} ds \right)$$

$$= (1 - \gamma_2) \left(r(t) - e^{-\alpha(t_0 - t)} r(t_0) + \int_0^{t_0 - t} (c'(t + s) + \lambda c(t + s)) e^{-\alpha s} ds \right) \qquad \textit{Lemma 4.16}$$

$$\leqslant (1 - \gamma_2) \left(r(t) - e^{-\alpha(t_0 - t)} r(t_0) + \int_0^{t_0 - t} c'(t + s) ds + \int_0^{t_0 - t} \lambda c(t_0) e^{-\alpha s} ds \right)$$

$$= (1 - \gamma_2) \left(r(t) - e^{-\alpha(t_0 - t)} r(t_0) + c(t_0) - c(t) + \frac{\lambda}{\alpha} (1 - e^{-\alpha(t_0 - t)}) c(t_0) \right).$$

Thus to prove (50), we only need to prove that

$$\frac{1}{1 - \gamma_2} \left(1 - e^{-\alpha(t_0 - t)} \right) \mathbf{E}\left[c(t_0 + X)\right] + \left(-e^{-\alpha(t_0 - t)}(r(t_0) - c(t_0)) + r(t) - c(t) \right)$$
$$\geqslant r(t) - e^{-\alpha(t_0 - t)} r(t_0) + c(t_0) - c(t) + \frac{\lambda}{\alpha} (1 - e^{-\alpha(t_0 - t)}) c(t_0),$$

which is equivalent to

$$\mathbf{E}\left[c(t_0 + X)\right] \geqslant c(t_0).$$

This holds by monotonocity of c.

Case 2: $t > t_0$. In this case, we want to prove

$$\ell \leqslant \mu \left(V(t) - \mathbf{E}\left[V(t - T_2)\right] \right). \tag{51}$$

Note that starting from state t, the policy stays active until the state drops back to t, during which time a $\overline{Cost}(t, \alpha)$ cost is incurred. After the state drops below t, since each time the state drops by an exponential amount, the state is $t - T_2$. Thus we have that

$$V(t) = \overline{Cost}(t, \alpha) + \gamma_1 \mathbf{E}\left[V(t - T_2)\right]. \tag{52}$$

Substituting (52) into (51) and using (32), we only need to prove that

$$\mathbf{E}\left[c(t_0 + X)\right] \leqslant -\frac{1 - \gamma_1}{\gamma_1} V(t) + \frac{1}{\gamma_1} \overline{Cost}(t, \alpha). \tag{53}$$

Now we characterize $V(t)$. Starting from state t, the policy stays active until the state drops below t. During this time (T_1, which follows the Busy period in M/M/1 distribution), a cost of $\overline{Cost}(t, \alpha)$ is incurred. After the state drops below t, the amount it is below t follows on exponential distribution with rate λ. Suppose it is $y_1 \sim t - \mathrm{Exp}(\lambda)$. If y_1 is still larger than t_0, the policy stays active until the state drops below y_1, incurring another $\overline{Cost}(y_1, \alpha)$ cost. This process continues until the state drops below t_0. Note that each time a $\overline{Cost}(y_i, \alpha)$ is incurred, the state drops by an exponential amount, thus the total number of iterations follow a Poisson distribution $Pois(\lambda(t - t_0))$. After the final iteration, the state is $\mathrm{Exp}(\lambda)$ below t_0, and the policy stays passive until the state grows back to t_0.

Thus we have that

$$V(t) = \overline{Cost}(t, \alpha) + \gamma_1 \mathbf{E}\left[\sum_{i=1}^{M} \gamma_1^{i-1} \overline{Cost}(y_i, \alpha)\right] + \mathbf{E}\left[\gamma_1^{M+1}\right]\left(\Gamma(t_0, \alpha) - \frac{\ell}{\alpha}(1 - \gamma_2) + \gamma_2 V(t_0)\right), \tag{54}$$

where $y_1 = t - \mathrm{Exp}(\lambda), y_{i+1} = y_i - \mathrm{Exp}(\lambda)$, and M is the random variable such that $y_M \geqslant t_0, y_{M+1} < t_0$, with $M \sim Pois(\lambda(t - t_0))$. By the probability generating function of a Poisson variable, we have that

$$\mathbf{E}\left[\gamma_1^{M+1}\right] = \gamma_1 PGF(\gamma_1) = \gamma_1 e^{-\lambda(t - t_0)(1 - \gamma_1)}. \tag{55}$$

Define $\Omega(t, t_0) := \mathbf{E}\left[\sum_{i=1}^{M} \gamma_1^{i-1} \overline{Cost}(y_i, \alpha)\right]$. We treat t_0 as constant and vary t by δ. Note that $\{y_i\}$ is distributed as Poisson arrivals in time (t_0, t), we discuss different cases based on the number of y_i in the interval $(t, t + \delta)$, denoted by N. Note that N follows the distribution $Pois(\lambda \delta)$.

If $N = 0$ (with probability $e^{-\lambda \delta}$), $\Omega(t + \delta, t_0) = \Omega(t, t_0)$.

If $N = 1$ (with probability $\lambda \delta e^{-\lambda \delta}$), we have that

$$\Omega(t + \delta, t_0) = \overline{Cost}(t + \Delta, t_0) + \gamma_1 \Omega(t, t_0),$$

where $0 < \Delta < \delta$.

The probability that $N > 1$ is $O(\delta^2)$.

Thus we have that

$$\Omega(t + \delta, t_0) = e^{-\lambda \delta} \Omega(t, t_0) + \lambda \delta e^{-\lambda \delta}\left(\overline{Cost}(t + \Delta, \alpha) + \gamma_1 \Omega(t, t_0)\right) + O(\delta^2). \tag{56}$$

Therefore,

$$\begin{aligned}
\Omega'(t, t_0) &:= \lim_{\delta \to 0} \frac{\Omega(t + \delta, t_0) - \Omega(t, t_0)}{\delta} \\
&= \lim_{\delta \to 0} \frac{e^{-\lambda \delta} + \lambda \delta e^{-\lambda \delta} \gamma_1 - 1}{\delta} \Omega(t, t_0) + \lambda e^{-\lambda \delta} \overline{Cost}(t + \Delta, \alpha) \\
&= (\lambda \gamma_1 - \lambda) \Omega(t, t_0) + \lambda \overline{Cost}(t, \alpha). \tag{57}
\end{aligned}$$

33

Note that this is a first-order ODE with respect to t. Using Lemma 4.13 and treating $y = \Omega(t, t_0), P(t) = \lambda(1 - \gamma_1), Q(t) = \lambda\overline{Cost}(t, \alpha)$, we have that

$$\Omega(t, t_0) = e^{-\lambda(1-\gamma_1)t}\left(\int e^{\lambda(1-\gamma_1)t}\lambda\overline{Cost}(t, \alpha)dt + C\right).$$

Noting that $\Omega(t_0, t_0) = 0$, we have that

$$\Omega(t, t_0) = e^{-\lambda(1-\gamma_1)t}\int_{t_0}^{t} e^{\lambda(1-\gamma_1)s}\lambda\overline{Cost}(s, \alpha)ds$$

$$= \int_{t_0}^{t} e^{-\lambda(1-\gamma_1)(t-s)}\lambda\overline{Cost}(s, \alpha)ds. \tag{58}$$

Now substituting (58) and (55) into (54), we have that

$$V(t) = \overline{Cost}(t, \alpha) + \gamma_1\Omega(t, t_0) + \mathbf{E}\left[\gamma_1^{M+1}\right]\left(\Gamma(t_0, \alpha) - \frac{\ell}{\alpha}(1 - \gamma_2) + \gamma_2 V(t_0)\right)$$

$$= \overline{Cost}(t, \alpha) + \gamma_1\int_{t_0}^{t} e^{-\lambda(1-\gamma_1)(t-s)}\lambda\overline{Cost}(s, \alpha)ds$$

$$+ \gamma_1 e^{-\lambda(t-t_0)(1-\gamma_1)}\left(\Gamma(t_0, \alpha) - \frac{\ell}{\alpha}(1 - \gamma_2) + \gamma_2 V(t_0)\right). \tag{59}$$

Moreover, we have that

$$\int_{t_0}^{t} e^{-\lambda(1-\gamma_1)(t-s)}\lambda\overline{Cost}(s, \alpha)ds = \frac{1}{1 - \gamma_1}e^{-\lambda(1-\gamma_1)(t-s)}\overline{Cost}(s, \alpha)\Big|_{t_0}^{t}$$

$$- \int_{t_0}^{t}\frac{1}{1-\gamma_1}e^{-\lambda(1-\gamma_1)(t-s)}\overline{Cost}'(s, \alpha)ds$$

$$= \frac{1}{1-\gamma_1}\left(\overline{Cost}(t, \alpha) - e^{-\lambda(1-\gamma_1)(t-t_0)}\overline{Cost}(t_0, \alpha)\right)$$

$$- \int_{t_0}^{t}\frac{1}{1-\gamma_1}e^{-\lambda(1-\gamma_1)(t-s)}\overline{Cost}'(s, \alpha)ds. \tag{60}$$

By (9), (27), (29) and (49), we have that

$$\Gamma(t_0, \alpha) - \frac{\ell}{\alpha}(1 - \gamma_2) + \gamma_2 V(t_0)$$

$$= (1 - \gamma_2)(r(t_0) - c(t_0)) - \frac{\ell}{\alpha}(1 - \gamma_2) + \gamma_2\left(r(t_0) - c(t_0) - \frac{\ell}{\alpha} + \frac{\lambda + \alpha}{\mu}\frac{\ell}{\alpha}\right)$$

$$= r(t_0) - c(t_0) - \frac{\ell}{\alpha} + \frac{\lambda}{\alpha}\frac{\ell}{\mu}. \qquad\qquad \text{by (9)} \tag{61}$$

34

Thus substituting (60) and (61) into (59), we have that

$$V(t) = \overline{Cost}(t, \alpha) + \frac{\gamma_1}{1 - \gamma_1} \left(\overline{Cost}(t, \alpha) - e^{-\lambda(1 - \gamma_1)(t - t_0)} \overline{Cost}(t_0, \alpha) \right)$$

$$- \int_{t_0}^{t} \frac{\gamma_1}{1 - \gamma_1} e^{-\lambda(1 - \gamma_1)(t - s)} \overline{Cost}'(s, \alpha) ds$$

$$+ \gamma_1 e^{-\lambda(t - t_0)(1 - \gamma_1)} \left(r(t_0) - c(t_0) - \frac{\ell}{\alpha} + \frac{\lambda}{\alpha} \frac{\ell}{\mu} \right) \tag{62}$$

Note that by (22), (31) and (32),

$$\overline{Cost}(t_0, \alpha) = (1 - \gamma_1)(r(t_0) - c(t_0) + \frac{1 - \gamma_1}{\gamma_1} \frac{\ell}{\alpha}).$$

Thus, we have that

$$V(t) = \frac{1}{1 - \gamma_1} \overline{Cost}(t, \alpha) - \frac{\gamma_1}{1 - \gamma_1} e^{-\lambda(1 - \gamma_1)(t - t_0)} \overline{Cost}(t_0, \alpha)$$

$$- \int_{t_0}^{t} \frac{\gamma_1}{1 - \gamma_1} e^{-\lambda(1 - \gamma_1)(t - s)} \overline{Cost}'(s, \alpha) ds$$

$$+ \gamma_1 e^{-\lambda(t - t_0)(1 - \gamma_1)} \left(r(t_0) - c(t_0) - \frac{\ell}{\alpha} + \frac{\lambda}{\alpha} \frac{\ell}{\mu} \right)$$

$$= \frac{1}{1 - \gamma_1} \overline{Cost}(t, \alpha) - \gamma_1 e^{-\lambda(1 - \gamma_1)(t - t_0)} \left(r(t_0) - c(t_0) + \frac{1 - \gamma_1}{\gamma_1} \frac{\ell}{\alpha} \right)$$

$$- \int_{t_0}^{t} \frac{\gamma_1}{1 - \gamma_1} e^{-\lambda(1 - \gamma_1)(t - s)} \overline{Cost}'(s, \alpha) ds$$

$$+ \gamma_1 e^{-\lambda(t - t_0)(1 - \gamma_1)} \left(r(t_0) - c(t_0) - \frac{\ell}{\alpha} + \frac{\lambda}{\alpha} \frac{\ell}{\mu} \right)$$

$$= \frac{1}{1 - \gamma_1} \overline{Cost}(t, \alpha) + \gamma_1 e^{-\lambda(1 - \gamma_1)(t - t_0)} \left(\frac{\lambda}{\mu} - \frac{1}{\gamma_1} \right) \frac{\ell}{\alpha}$$

$$- \int_{t_0}^{t} \frac{\gamma_1}{1 - \gamma_1} e^{-\lambda(1 - \gamma_1)(t - s)} \overline{Cost}'(s, \alpha) ds \tag{63}$$

Now look back to our goal, which is proving (53):

$$\mathbf{E}\left[c(t_0 + X) \right] \leqslant -\frac{1 - \gamma_1}{\gamma_1} V(t) + \frac{1}{\gamma_1} \overline{Cost}(t, \alpha).$$

This is equivalent to proving

$$V(t) - \frac{1}{1 - \gamma_1} \overline{Cost}(t, \alpha) \leqslant -\frac{\gamma_1}{1 - \gamma_1} \mathbf{E}\left[c(t_0 + X) \right].$$

By (63), our goal is equivalent to proving

$$\gamma_1 e^{-\lambda(1-\gamma_1)(t-t_0)}\left(\frac{\lambda}{\mu}-\frac{1}{\gamma_1}\right)\frac{\ell}{\alpha}-\int_{t_0}^t\frac{\gamma_1}{1-\gamma_1}e^{-\lambda(1-\gamma_1)(t-s)}\overline{Cost}'(s,\alpha)ds\leqslant-\frac{\gamma_1}{1-\gamma_1}\mathbf{E}\left[c(t_0+X)\right]$$

Rearranging the terms and using (32), our goal is equivalent to proving that

$$\mathbf{E}\left[c(t_0+X)\right]\left(\frac{\gamma_1}{1-\gamma_1}+e^{-\lambda(1-\gamma_1)(t-t_0)}\left(\frac{\lambda\gamma_1}{\alpha}-\frac{\mu}{\alpha}\right)\right)\leqslant\int_{t_0}^t\frac{\gamma_1}{1-\gamma_1}e^{-\lambda(1-\gamma_1)(t-s)}\overline{Cost}'(s,\alpha)ds.$$

(64)

Note that by (22) and (31), we have that

$$\overline{Cost}(t,\alpha)=(1-\gamma_1)(r(t)-c(t)+\frac{\mu(1-\gamma_1)}{\gamma_1\alpha}\mathbf{E}\left[c(t+X)\right]).$$

Thus using Lemma 4.16 and (9),

$$\begin{aligned}\overline{Cost}'(t,\alpha)&=(1-\gamma_1)\left(\lambda c(t)+\frac{\mu(1-\gamma_1)}{\gamma_1\alpha}\mathbf{E}\left[c'(t+X)\right]\right)\\&=(1-\gamma_1)\left(\lambda c(t)+\frac{\lambda+\alpha-\lambda\gamma_1}{\alpha}\mathbf{E}\left[c'(t+X)\right]\right).\end{aligned}$$

Moreover, by (9),

$$\frac{\lambda\gamma_1}{\alpha}-\frac{\mu}{\alpha}=\frac{1}{\alpha}(\lambda\gamma_1-\frac{(\lambda+\alpha-\lambda\gamma_1)\gamma_1}{1-\gamma_1})=-\frac{\gamma_1}{1-\gamma_1},$$

Thus we have that our goal, equivalent to (64), is equivalent to showing that

$$\mathbf{E}\left[c(t_0+X)\right]\frac{1}{1-\gamma_1}\left(1-e^{-\lambda(1-\gamma_1)(t-t_0)}\right)\leqslant\int_{t_0}^t e^{-\lambda(1-\gamma_1)(t-s)}\left(\lambda c(s)+\frac{\lambda+\alpha-\lambda\gamma_1}{\alpha}\mathbf{E}\left[c'(s+X)\right]\right)ds.$$

(65)

Finally, we have that for any $s > t_0$,

$$\begin{aligned}&\lambda c(s)+\frac{\lambda+\alpha-\lambda\gamma_1}{\alpha}\mathbf{E}\left[c'(s+X)\right]\\&=\lambda c(s)+\frac{\lambda+\alpha-\lambda\gamma_1}{\alpha}\cdot\frac{\alpha}{1-\gamma_1}\left(\mathbf{E}\left[c(s+X)\right]-c(s)\right)&&Lemma\ 4.14\\&=\left(\lambda-\frac{\lambda+\alpha-\lambda\gamma_1}{1-\gamma_1}\right)c(s)+\frac{\lambda+\alpha-\lambda\gamma_1}{1-\gamma_1}\mathbf{E}\left[c(s+X)\right]\\&=-\frac{\alpha}{1-\gamma_1}c(s)+\frac{\lambda+\alpha-\lambda\gamma_1}{1-\gamma_1}\mathbf{E}\left[c(s+X)\right]\\&\geqslant\left(-\frac{\alpha}{1-\gamma_1}+\frac{\lambda+\alpha-\lambda\gamma_1}{1-\gamma_1}\right)\mathbf{E}\left[c(s+X)\right]\\&=\lambda\mathbf{E}\left[c(s+X)\right]\\&\geqslant\lambda\mathbf{E}\left[c(t_0+X)\right].\end{aligned}$$

36

Thus

$$\int_{t_0}^t e^{-\lambda(1-\gamma_1)(t-s)} \left(\lambda c(s) + \frac{\lambda + \alpha - \lambda\gamma_1}{\alpha} \mathbf{E}\left[c'(s+X) \right] \right) ds$$

$$\geqslant \int_{t_0}^t e^{-\lambda(1-\gamma_1)(t-s)} \lambda \mathbf{E}\left[c(t_0+X) \right] ds$$

$$= \mathbf{E}\left[c(t_0+X) \right] \frac{1}{1-\gamma_1} \left(1 - e^{-\lambda(1-\gamma_1)(t-t_0)} \right).$$

□

To prove indexability, there are actually two other corner cases to prove.

Lemma E.2.   *1. If the compensation $\ell \leqslant \mu\mathbf{E}\left[c(X) \right]$, always active (except at state 0) is optimal.*

2. If $\lim_{t\to\infty} \mu\mathbf{E}\left[c(t+X) \right] = M$ exists and the compensation $\ell \geqslant \mu M$, always passive is optimal.

Proof. To prove the first argument, the argument is the same as what we have in Case 2 of the proof of Lemma E.1. Now we prove the second argument.

Since $\lim_{t\to\infty} \mu\mathbf{E}\left[c(t+X) \right] = M$ exists and c is monotone, we have that $\lim_{t\to\infty} c(t) = M$.

For the policy "always passive", we can derive the value function:

$$V(t) = \int_0^\infty (\alpha r(t+s) - \ell)e^{-\alpha s} ds$$

$$= \int_0^\infty \alpha r(t+s)e^{-\alpha s} ds - \frac{\ell}{\alpha}$$

$$= -r(t+s)e^{-\alpha s}\Big|_0^\infty - \int_0^\infty -r'(t+s)e^{-\alpha s} ds - \frac{\ell}{\alpha}$$

$$= r(t) + \int_0^\infty (c'(t+s) + \lambda c(t+s))e^{-\alpha s} ds - \frac{\ell}{\alpha}. \qquad Lemma\ 4.16 \qquad (66)$$

Now we verify that this value function satisfies the Hamilton–Jacobi–Bellman equation, which is equivalent to showing:

$$\lim_{\delta\to 0} \frac{passive(t) - active(t)}{\delta} \leqslant 0.$$

By the same argument in (40) and (41), our goal is equivalent to showing

$$\ell \geqslant \mu(V(t) - \mathbf{E}\left[V(t-T_2) \right]). \qquad (67)$$

Note that (47) still holds. Thus it suffices to prove that

$$\ell \geqslant \mu\left((1-\gamma_2)(V(t) + \frac{\ell}{\alpha}) - \Gamma(t,\alpha) \right).$$

Using (66), (27) and (29), we only need to prove that

$$\frac{\ell}{\mu} \geqslant (1-\gamma_2)\left(r(t) + \int_0^\infty (c'(t+s) + \lambda c(t+s))e^{-\alpha s} ds - (r(t) - c(t)) \right).$$

Note that the right hand side can be bounded as follows:

$$(1 - \gamma_2) \left(r(t) + \int_0^\infty (c'(t+s) + \lambda c(t+s)) e^{-\alpha s} ds - (r(t) - c(t)) \right)$$

$$\leq (1 - \gamma_2) \left(c(t) + \int_0^\infty c'(t+s) ds + \int_0^\infty \lambda M e^{-\alpha s} ds \right)$$

$$= (1 - \gamma_2) (\lim_{t \to \infty} c(t) + \frac{\lambda}{\alpha} M)$$

$$= (1 - \gamma_2) \frac{\lambda + \alpha}{\alpha} M$$

$$= M.$$

Since $\ell \geq \mu M$, we have the proof. $\qquad\square$

Finally we can prove indexability.

Theorem E.3 (Indexability). *The discounted single-arm bandit problem is indexable with the Whittle Index $W(t_0, \alpha) = \widehat{W}(t_0, \alpha)$.*

Proof. We first prove indexability. Recall that

$$\Pi(\ell) := \{t_0 \mid \text{passive action at state } t_0 \text{ is optimal}\}.$$

Now we prove that for any $\ell_1 < \ell_2$, we have that $\Pi(\ell_1) \subseteq \Pi(\ell_2)$.

If $\ell_1 \leq \mu \mathbf{E}\left[c(X)\right]$, by Lemma E.2, $\Pi(\ell_1) = \{0\}$, and the statement holds immediately.

If $\ell_2 \geq \lim_{t \to \infty} \mu \mathbf{E}\left[c(t+X)\right]$, by Lemma E.2, $\Pi(\ell_2) = \mathbb{R}$, and the statement holds immediately.

Otherwise, we have that for some t_1, t_2, $\ell_1 = \widehat{W}(t_1, \alpha)$, $\ell_2 = \widehat{W}(t_2, \alpha)$. Since $\widehat{W}(t, \alpha)$ is non-decreasing with t, we have that $t_1 < t_2$. By Lemma E.1, $\Pi(\ell_1) = \{t \mid t \leq t_1\}$, and $\Pi(\ell_2) = \{t \mid t \leq t_2\}$. Thus we prove the statement.

The proof that the Whittle Index is $\widehat{W}(t, \alpha)$ is straightforward. Mathematically, we want to show that

$$\widehat{W}(t_0, \alpha) = \inf_\ell \{t \in \Pi(\ell)\}.$$

This follows immediately by Lemma E.2 and Lemma E.1. $\qquad\square$