Dynamic Scheduling with Bayesian Updating of Customer Characteristics

Buyun Li • Xiaoshan Peng • Owen Q. Wu

Kelley School of Business, Indiana University, Bloomington, IN 47405, {libu, xp1, owenwu}@iu.edu

January 26, 2025

In many service industries, decision making about service scheduling often relies on assessing and prioritizing customer needs and value using professional judgment and customer data. Traditional scheduling models assume perfect knowledge of customer service rewards and delay costs, which is unrealistic. This paper considers the optimal scheduling problem in a multi-class queueing system where the system manager learns the reward of serving customers dynamically. We model the scheduling problem as a restless multiarmed bandit (RMAB) problem, with each customer class representing an arm characterized by queue length and the manager's belief about the reward distribution. We derive the Whittle index for each customer class. The resulting Whittle index scheduling policy which prioritizes the class of customers with the highest Whittle index. We prove that the Whittle index offers an optimal solution for a system with two customer classes-one with perfect information and one with unknown parameters-and show that it is near-optimal for more general settings numerically. Our results show that the incentive to serve a class of customers with unknown rewards increases with service rate, higher belief in rewards, arrival rate and length of wait, which contrasts with traditional models. This finding highlights that as queues grow longer, the priority for serving them increases due to extended busy periods. Furthermore, for a fixed product of service rate and reward, we find that customer classes with higher service rates provides higher incentives for learning. By understanding these dynamics, managers can better allocate resources, ensuring that longer queues, which imply greater potential delays and customer dissatisfaction, are addressed more promptly.

Key words: Queueing Scheduling, Bayesian Learning, Restless-Armed Bandit, Whittle Index

1. Introduction

In many service industries, decision-making on service scheduling often relies on assessing and prioritizing customer needs and value by utilizing both professional judgment and data from similar cases. In healthcare practice, nurses and physicians often adjust patient schedules based on assessments of the benefits of medical care, e.g., scheduling for emergency care (Ala and Chen 2022), surgeries (Oliveira et al. 2020), and diagnostic imaging (Déry et al. 2020). These assessments are grounded in healthcare professionals' evolving knowledge and experience with patients' symptoms, conditions, treatment benefits, and costs of delay, which are continuously updated based on treatment outcomes. Similarly, in the hospitality (Roy et al. 2023) and financial services (Coro 2023) industries, high-value customers are prioritized to increase revenue, with practices such as VIP queues and direct access to financial advisors. In such settings, customer valuation is learned through transactional histories and spending characteristics. When information on customer needs and value is limited, optimizing the learning process while making scheduling decisions becomes particularly challenging and important.

Motivated by these industry practices, this paper explores the joint optimization of learning and queueing scheduling decisions. Classical results in the scheduling of multi-class queues (see Section 2 for a detailed review) are typically derived under the assumption that the system manager has perfect knowledge of the service reward (or delay costs) associated with serving each class of customers. Our paper relaxes this assumption. Instead, we assume that the system manager learns about service rewards via Bayesian updating while making dynamic scheduling decisions. We explore the optimal scheduling policy under Bayesian updating and study how the optimal policy depends on the queue length and beliefs about the service rewards. Motivated by industry practices, this paper explores the joint optimization of learning and queue-scheduling decisions. While classical analyses of multi-class queues assume that the system manager has perfect knowledge of service rewards or delay costs (see Section 2), we relax this assumption by allowing the manager to learn the parameter of service rewards via Bayesian updating while simultaneously making scheduling decisions. To maintain analytical tractability and align with managerial practice, we focus learning on service rewards rather than both service rates and rewards. In most business and operational contexts, service tasks are tightly standardized, so service rates are already well

measured and calibrated in standard operating procedures. Empirical evidence supports this modeling choice: call center handling times show consistent stability across daily and seasonal periods (Brown et al. 2005), operating room planners rely on extensive case-duration planning that makes surgical processing times highly predictable at the tactical level (Strum et al. 2000), and fast-service restaurants measure cashier cycle times to within seconds, treating them as fixed for drive-through analysis (Teguh and Setiawan 2012). Consequently, the main source of uncertainty and thus the lever that most influences scheduling performance is the value realized once service is completed, whether expressed as clinical benefit, incremental revenue, or customer retention. By embedding Bayesian learning of service rewards directly within the scheduling model, we place the learning mechanism where it matters most and retain the analytical tractability needed to characterize how queue length and evolving beliefs interact to determine optimal priorities.

We consider a single-server queueing system with K classes of customers. Customers arrive to the system for service and wait in the queues if the server is busy. The system manager determines a preemptive scheduling policy to serve the K classes of customers to maximize the expected total discounted reward over an infinite horizon. The system manager does not have perfect knowledge of the customer service rewards; in particular, the parameters of the reward distributions are unknown to the manager. Thus, upon service completion of each customer from class k ($k = 1, \ldots, K$), the system manager collects a realization from the service reward distribution, and updates their beliefs about the parameters via the Bayes rule. We explore the structure of the optimal scheduling policy under Bayesian updating. Added Text We choose the Beyasian updating framework because the Bayesian framework encapsulates all the information from the past in a fixed dimension belief vector whose conjugate update is analytically tractable; this compact state structure makes the formulation of dynamic programming feasible and leads directly to Whittle-index characterizations. Consequently, Bayesian learning is uniquely suited when the research goal is to derive and analyze the optimal scheduling policy rather than a heuristic bound on long-term regret.

Since the system state includes not only the queue lengths but also the system manager's beliefs about the customers' service rewards of all K classes, it is challenging to solve the optimal scheduling policy. Instead of solving the optimal scheduling problem directly, we formulate the dynamic scheduling problem with Bayesian updating as a restless multi-armed bandit (RMAB)

problem. Each customer class is represented as an arm, with the state of each arm being the queue length and the manager's belief about the reward distribution for that class. Thus, the optimal scheduling problem is equivalent to the decision of which arm to serve next upon customer arrival and departure. We establish the indexability of this RMAB problem and derive the corresponding Whittle index. We obtain an explicit characterization of the Whittle index. Specifically, we show that the Whittle index can be obtained by solving an optimal stopping problem, in which the optimal stopping time is constrained within the busy period for the current queue length. Thus, the index increases with both a higher reward belief and the queue length because a longer queue length reflects a longer busy period. Our analysis demonstrates that the scheduling policy derived from the Whittle index is optimal in a specific scenario with two customer classes: one with perfect reward information and the other with unknown reward parameters. For general settings, the scheduling policy given by the Whittle index provides a near-optimal heuristic that facilitates an efficient solution numerically.

This paper makes several contributions. First, we show the structure of the Whittle index within the framework of a Bayesian learning model, bridging the gap between learning and scheduling in queues. Second, our analysis underscores the impact of queue length on optimal scheduling policy, a departure from the existing queueing literature that typically finds queue length irrelevant in scheduling decisions. This novel insight improves our understanding of how service managers can effectively integrate learning and service delivery across various service sectors. Finally, through numerical studies, we show that the Whittle index scheduling policy has a near-optimal performance under various parameter settings. Our paper makes several contributions. First, we provide an exact characterization of the structure of the Whittle index within a Bayesian learning framework, bridging the gap between dynamic learning and scheduling in queueing systems. The closed-form characterization of the Whittle index provides a more refined analysis compared to regret analysis on different heuristics. Second, our analysis shows that queue length directly influences the incentive to explore, in contrast to the classical queueing literature where queue length is often decoupled from priority rules. This insight highlights a previously underexplored interplay between learning and queue scheduling decisions, providing insights on how service managers can effectively integrate learning with service delivery across diverse sectors. Finally, through extensive

numerical experiments, we demonstrate that the Whittle index scheduling policy achieves near-optimal performance across a wide range of parameter settings, reinforcing its practical value as a scalable and effective heuristic for systems with uncertain service rewards.

2. Literature Review

Our paper builds on three streams of literature: the classical queueing scheduling problem with multiple classes of customers, the application of multi-armed bandit (MAB) problems and their extensions in queueing control, and the integration of learning frameworks (particularly Bayesian learning) within queueing systems. Our work synthesizes these approaches to provide insights on how learning of the customer reward characteristics impacts optimal scheduling decisions.

In classical queueing scheduling problems with multiple classes, it is well-known that the $c\mu$-rule is optimal in wide range of settings. Cox and Smith (1961) is the first to propose the $c\mu$-rule, which optimally solves the classical queueing scheduling problem under an average cost criterion, assuming linear holding costs, exponential customer inter-arrival times, and general service times. Harrison (1975a,b) extend the scope of the problem to a generalized system with both service reward and delay cost, studying the optimal policy to maximize total profit under a discounted reward criterion, in which a more complex index policy allowing inserted idleness is optimal for the non-preemptive single server systems. Klimov (1975) extends the $c\mu$-rule to M/G/1 systems with feedback, where a job exiting the system has some probability of being replaced by another job. The static index policy remains optimal in various extensions; see, for instance, Tcha and Pliska (1977), Buyukkoc et al. (1985) and Hirayama et al. (1989). Discussion of the cost structure, particularly the convex cost structure, also gained attention. Van Mieghem (1995) offers a generalized $c\mu$-rule for multi-class scheduling with convex delay costs under heavy traffic assumption. Additionally, Mandelbaum and Stolyar (2004) extends the asymptotic optimality of the generalized $c\mu$-rule to multi-server system in heavy-traffic settings. In summary, the classical queueing literature builds its results assuming that the system manager knows the delay cost or the service reward information perfectly. In contrast, we relax this assumption and study scenarios where the system manager does not have perfect information on customer delay costs or service rewards. Instead, the system manager learns this information within a Bayesian framework by observing the realizations of rewards and costs upon service completion.

The second stream of literature is applications of Multi-arm bandit problem (MAB) and its extensions (Gittins et al. 2011), e.g. branching bandit and restless bandit, on queueing scheduling systems. Applications of the MAB model to study the queue scheduling decision may be categorized into two streams: 1) each customer is modeled as an arm, and 2) each class of customers is modeled as an arm.

The literature modeling each customer in the queue as an arm stems from the scheduling problem of $M/G/1$ queue, where the distribution of a customer (job)'s remaining processing time is updated after being in service for some time. Whittle (1982), Weiss (1988) are among the first to model the $M/G/1$ queue as an arm-acquiring/branching multiarmed bandit (MAB) problem where each customer is modeled as an arm. Gittins et al. (1989) is often cited in the literature as proving the Gittins index policy's optimality in the $M/G/1$ when a preemptive policy is utilized. Lai and Ying (1988) reexamined work by Klimov (1975) on the non-preemptive $M/G/1$ queue with feedback, connected it to the MAB problem, and extended it to the preemptive $M/M/1$ queue with feedback. A performance analysis by Whittle (2005) provides a dynamic programming perspective on the optimality proof of the Gittins index policy in the $M/G/1$ queue. This approach, which models each customer in an $M/G/1$ queue as an arm in the MAB problem, has been extensively surveyed by Scully and Harchol-Balter (2021); for more details, we refer the readers to their review. Most recently, the Gittins index policy using this modeling approach has been shown to be near optimal in minimizing the mean response time in a more general preemptive $M/G/m$ queue (Scully et al. 2020). However, this classical approach may not be applicable to our setting. In our scenario, updating beliefs about the reward of one customer within a class can change beliefs about the rewards of all customers within that class, which violates the 'frozen' assumption of MAB problems.

The approach of modeling each class of customer as an arm often utilizes the restless arm bandit framework, where queue lengths evolve independently of which class is in service due to ongoing arrivals. This stream of work is highly relevant to our paper, as we model each customer as an arm to investigate the queueing scheduling problem with learning of reward and/or cost. Whittle (1996) first modeled the $M/G/1$ queue scheduling problem as a RMAB problem. However, he concluded that the Whittle index, a well-known solution technique for RMAB problems, generally does not reduce to the $c\mu$-rule. Glazebrook et al. (2003), Ansell et al. (2003) extended the multiclass

$M/G/1$ system with non-preemptive service, incorporating a server utilization cost and convex holding costs, respectively. They provided proof of indexability and derived an expression for the Whittle index. Argon et al. (2009) applied an RMAB framework to study the queueing routing problem in a setting with multiple parallel queues, where servers serve both dedicated and generic customers. More recently, Ayesta et al. (2017) used an RMAB framework to study the queueing scheduling problem with customer abandonment, demonstrating that an index policy is optimal when the queue capacity is restricted to one or two. Most recently, Aalto (2024) showed that the scheduling problem of an $M/G/m$ system with abandonment is indexable when modeled as an RMAB problem and derived the corresponding Whittle index. We extend this body of literature by including information states based on the belief of rewards to investigate the trade-off between parameter learning and revenue earning in queueing scheduling decisions.Added Text: Although our model adopts the RMAB framework by treating each customer class as an arm, we further bridge the gap between queueing dynamics and bandit formulations. In particular, we derive a novel structural result showing that the Whittle index for each class is bounded by a busy period driven by the queue length. This condition—where the optimal stopping time for an arm's 'active' phase cannot exceed its busy period—marks a new interface between queueing scheduling problem and RMAB framework. As a result, our Whittle index not only accounts for belief updates but also explicitly incorporates queueing-related constraints that have not been previously characterized in the RMAB literature.

The Bayesian framework (Bernardo and Smith 2009) is the most classical sequential learning framework that synergizes with queueing models, making it a natural paradigm for studying the sequence of access services. Additionally, Bayesian learning is often combined with the multi-armed bandit problem by incorporating an information state to model the explore-exploit trade-off in operations, as surveyed by Bouneffouf et al. (2020). In queueing settings, literature has focused on using the Bayesian framework to learn about offered load intensities (ρ), as reviewed by Asanjarani et al. (2021). Added Text This stream of literature focuses on estimating queue parameters for a system whose scheduling rule is fixed, treating learning and control as separate stages. Our work instead integrates learning and scheduling—each service decision both updates reward beliefs and sets the next priority. The body of literature studying decision making in queueing systems

incorporating Bayesian learning is still growing, with notable contributions. Aktekin and Soyer (2012) surveyed multiple prior distributions in queueing settings and provided an analysis on their implications for the total cost function using examples from a call center. Afèche and Ata (2013) combined the Bayesian learning framework with the admission pricing problem in queueing, studying a setting where the system manager learns the revenue characteristics of the total customer profile sequentially after the completion of the service. Lingenbrink and Iyer (2019) examined a system in which customers are strategic and learn the system's service rate through a signaling mechanism in a Bayesian manner. Krishnasamy et al. (2021) studies a discrete-time queueing and routing problem where the service success rate distribution parameters between each class of customers and each server are unknown and learned via a Bayesian framework. Their paper analyzes the asymptotic regret of total throughput for heuristics including Upper Confidence Bound and Thompson Sampling algorithms with forced exploration. In Krishnasamy et al. (2021), an MAB model is used to characterize the learning-earning trade-off. Our paper complements their study in two aspects: 1) we study the structure of the optimal Bayesian scheduling policy instead of regret bounds of heuristics, and 2) We focus on queueing scheduling decisions while incorporating the learning of unknown service reward parameters. Our paper complements prior studies by being the first to use the RMAB framework to study the queueing scheduling problem with Bayesian learning of customer characteristics. Unlike existing queueing literature that employs bandit frameworks to facilitate learning through regret analysis of heuristics, we analyze the structure of the optimal queue scheduling policy under Bayesian learning, providing insights into how learning shapes optimal scheduling decisions.

Furthermore, literature focusing on general concepts of learning (not limited to the Bayesian framework) and decision-making in queueing settings has recently gained traction. Krishnasamy et al. (2018) were among the first to study the regret bounds on queueing scheduling systems with information learning. Based on the work-conserving observation, they concluded a constant holding cost regret for learning the $c\mu$ rule via straightforward updates of the empirical mean. Zhong et al. (2024) proposed a learn-then-schedule algorithm in a similar setting but considered customer abandonment. In the exploration phase, they provided a point estimate of $C\mu/\theta$; during the exploitation phase, the estimated $C\mu/\theta$ rule was activated. Under this algorithm, the smallest achievable regret

for the estimate grows logarithmically. Walton and Xu (2021) reviewed information learning topics in stochastic networks and queues, with an additional focus on adversarial learning between the exploration and exploitation of information in queues. Chen et al. (2024) determined admission price and the service rate jointly using an online learning algorithm. Freund et al. (2023) address the problem of selecting servers for different classes of customers, with a focus on measuring the cost of learning in queues. Our paper complements the above research by studying the Bayesian optimal policy without relying on regret analysis. As such, the structure of our results provides insight into the value of learning in the queueing scheduling problem. Existing studies have primarily investigated learning in queueing systems through regret analysis of various heuristics. In contrast, our paper directly analyzes the Bayesian optimal policy and derive a closed-form characterization of the structure of the optimal scheduling policy. This more refined analysis of the optimal policy provides precise insights into how queue length directly affects optimal scheduling with learning.

3. The Model

In this section, we first present the dynamic programming formulation for the queueing scheduling problem involving a single server, K customer classes, and Bayesian updating of reward parameters (Section 3.1). Next, we approach the problem as a restless multi-armed bandit (RMAB) problem and formulate the dynamic program for each arm (customer class) in Section 3.2.

3.1 Multi-class Dynamic Scheduling Problem

We consider the problem of deciding a dynamic scheduling policy for a single-server queueing system serving K classes of customers. The service system manager ('manager' hereafter) knows that customers of class $k \in \{1, 2, \ldots, K\}$ arrive according to a Poisson process with rate λ_k (independent of other classes), and their service times are independent of each other and follow an exponential distribution with mean $1/\mu_k$. The manager is able to identify each customer's class upon their arrival. Upon completing serving a customer of class k, the manager receives a reward that is an independent realization of a random variable $R_k \geq 0$. The manager learns about the distribution of R_k using a Bayesian method (detailed shortly). For analytical tractability, we assume a preemptive service discipline, i.e., at any time, the manager can choose to serve a different customer class. The objective is to maximize the expected discounted reward over an infinite horizon, with a discount

rate $\gamma \in (0, 1]$.[1]

We next detail the process of learning about the reward. For each class $k \in \{1, 2, \ldots, K\}$, the reward R_k follows a probability distribution $p_k(r_k \mid \theta_k)$, which is a probability mass (or density) function if R_k is a discrete (or continuous) random variable. The manager knows the functional form of $p_k(r_k \mid \theta_k)$, but does not know the value of the parameter θ_k. The manager learns about θ_k over time using the Bayesian approach. Specifically, upon completing servicing a customer from class k, the manager receives a reward r_k, which is a realization of R_k, and updates the belief about the parameter θ_k as follows:

$$f_k(\theta_k \mid r_k) = \frac{p_k(r_k \mid \theta_k)\, f_k(\theta_k)}{\int p_k(r_k \mid \theta_k)\, f_k(\theta_k)\, d\nu(\theta_k)}, \tag{1}$$

where $\nu(\theta_k)$ is the counting (or Lebesgue) measure if θ_k is a discrete (or continuous) random variable, $f_k(\theta_k)$ is the belief about θ_k prior to receiving the reward r_k, $f_k(\theta_k \mid r_k)$ is the posterior belief, and these belief distributions are probability mass (or density) functions.

Following conventions in Bayesian analysis, we assume that the prior $f_k(\theta_k)$ has a conjugate distribution with respect to the likelihood $p_k(r_k \mid \theta_k)$. Under the assumption of the conjugate prior, the posterior belief $f_k(\theta_k \mid r_k)$ and the prior $f_k(\theta_k)$ belong to the same family of distributions, allowing for efficient repeated Bayesian updating using (1). This process reduces to updating the parameters of the distributions, as detailed below.

Let $\chi_k \in \Sigma_k$ be the parameters for the family of belief distributions, and write the conjugate prior as $f_k(\theta_k; \chi_k)$. According to Bernardo and Smith (2009), there exists a function $g_k : \Sigma_k \times \mathbb{R} \to \Sigma_k$ that updates the parameters χ_k, such that the posterior belief from (1) is:

$$f_k(\theta_k \mid r_k) = f_k(\theta_k; \chi_k') \quad \text{with} \quad \chi_k' = g_k(\chi_k, r_k). \tag{2}$$

Therefore, the parameter χ_k contains all the information needed to determine the predictive distribution of the reward R_k and serves as state variable for our dynamic scheduling decision formulated below.

Let $q_k \in \mathbb{Z}_+$ be a non-negative integer denoting the number of class k customers in the system and define $\boldsymbol{q} := (q_1, \ldots, q_K)$ as the vector of customer counts in the system. Let $\boldsymbol{\chi} := (\chi_1, \ldots, \chi_K)$

[1] If there is a holding cost of h_k per unit of time for a class k customer in the system, adopting the techniques by Harrison (1975a,b), we can equivalently consider a pure-reward system with zero holding costs and service rewards $R_k' = R_k + h_k/\gamma$. We update the belief about R_k' only after each service completion.

be the vector of knowledge state about the reward distributions. Given the assumption of Poisson arrivals, exponential service times, and preemptive queueing discipline, the manager's decision on which customer class to serve can be based on $(\boldsymbol{q}, \boldsymbol{\chi})$ alone, without needing information about the currently served customer class or the length of time they have been in service. Therefore, the state of the system is $s := (\boldsymbol{q}, \boldsymbol{\chi})$, which belongs to the state space $S := \mathbb{Z}_+^K \times \prod_{k=1}^K \Sigma_k$.

Let $A := \{0, 1, \ldots, K\}$ denote the action space, where action 0 corresponds to idling and action $k \geq 1$ corresponds to serving customer class k. For state $s = (\boldsymbol{q}, \boldsymbol{\chi})$, the set of customer classes the manager can choose to serve is

$$A(\boldsymbol{q}, \boldsymbol{\chi}) := \{k : q_k > 0, k = 1, 2, \ldots, K\}. \tag{3}$$

We note that the optimal decision must satisfy the following property: the server stays active if and only if there is at least one customer in the system. This follows from the fact that, under a preemptive queueing discipline and non-negative rewards, idling (action 0) is suboptimal when customers of any class are present.

The system state evolves as follows. If the current state is $(\boldsymbol{q}, \boldsymbol{\chi})$, upon the arrival of a customer of class k, the system state changes to $(\boldsymbol{q} + \boldsymbol{e}_k, \boldsymbol{\chi})$, where $\boldsymbol{e}_k$ is a K-dimensional vector with the k-th entry equal to 1 and all other entries equal to 0. Upon completing serving a customer of class k, the manager collects a reward r_k and updates the belief about θ_k; the system state transitions to $(\boldsymbol{q} - \boldsymbol{e}_k, \boldsymbol{\chi} + (g_k(\chi_k, r_k) - \chi_k)\boldsymbol{e}_k)$. For notational convenience, we denote the state of updated knowledge as $g(\boldsymbol{\chi}, r_k) := \boldsymbol{\chi} + (g_k(\chi_k, r_k) - \chi_k)\boldsymbol{e}_k$. The manager decides which customer class to serve immediately after system state changes, triggered by either an arrival or a service completion.

We first formulate the dynamic scheduling problem as a continuous-time Markov decision process with discounted reward. According to Puterman (1994), it suffices to focus our attention on the set of stationary and deterministic policies. Let $\pi = (d)^\infty$ denote a stationary policy with a deterministic decision rule, $d : S \to A$. Let $(\boldsymbol{Q}^\pi(t), \boldsymbol{X}^\pi(t))$ denote the state of the system at time t under policy π, where $\boldsymbol{Q}^\pi(t) = (Q_1^\pi(t), \ldots, Q_K^\pi(t))$ denotes the number of customers in each class at time t and $\boldsymbol{X}^\pi(t) = (X_1^\pi(t), \ldots, X_K^\pi(t))$ denotes the state of knowledge about rewards. Let $D_{k,n}^\pi$ denote the departure time of the n-th customer within class k under policy π and their reward

generated as $R_{k,n}$.[2] Then, the total expected discounted reward under policy π given an initial state $(\boldsymbol{q}, \boldsymbol{\chi})$ can be written as:

$$V^{\pi}(\boldsymbol{q}, \boldsymbol{\chi}) := \mathbb{E}\left[\sum_{k=1}^{K}\sum_{n=1}^{\infty} e^{-\gamma D^{\pi}_{k,n}} R_{k,n} \,\middle|\, (\boldsymbol{Q}^{\pi}(0), \boldsymbol{X}^{\pi}(0)) = (\boldsymbol{q}, \boldsymbol{\chi})\right].$$

The policy π determines which customer class to serve at each arrival or departure, directly influencing both the rewards accrued and the information accumulated about the reward distributions. Since the reward received at time t is discounted by $e^{-\gamma t}$, dynamic scheduling is crucial, as delays reduce the present value of rewards. A well-designed policy π should strike a balance between two objectives: (1) exploiting the current knowledge of reward distributions to maximize immediate discounted rewards, and (2) exploring certain customer classes to acquire information that improves future scheduling decisions.

Let Π denote the set of stationary and deterministic policies. Given the initial state $(\boldsymbol{q}, \boldsymbol{\chi})$, the manager's problem can be written as:

$$V(\boldsymbol{q}, \boldsymbol{\chi}) = \sup_{\pi \in \Pi} V^{\pi}(\boldsymbol{q}, \boldsymbol{\chi}). \tag{4}$$

Next, we rewrite the problem in (4) as a discrete-time Markov decision problem by employing uniformization (Puterman 1994). Let $m \geq \max_k \{\mu_k\} + \sum_{i=1}^{K} \lambda_i$ denote the uniformization constant. Then, the value function $V(\boldsymbol{q}, \boldsymbol{\chi})$ defined in (4) satisfies the following Bellman equations:

$$\begin{aligned} V(\boldsymbol{q}, \boldsymbol{\chi}) = \max_{k \in A(\boldsymbol{q}, \boldsymbol{\chi})} \Bigg\{ \beta \Bigg(& \frac{\mu_k}{m} \mathbb{E}\left[R_k + V(\boldsymbol{q} - \boldsymbol{e}_k, g(\boldsymbol{\chi}, R_k)) \mid \boldsymbol{\chi}\right] \\ & + \sum_{i=1}^{K} \frac{\lambda_i}{m} V(\boldsymbol{q} + \boldsymbol{e}_i, \boldsymbol{\chi}) + \Big(1 - \frac{\mu_k + \sum_{i=1}^{K} \lambda_i}{m}\Big) V(\boldsymbol{q}, \boldsymbol{\chi}) \Bigg) \Bigg\}, \quad \boldsymbol{q} \neq \boldsymbol{0}, \end{aligned} \tag{5}$$

$$V(\boldsymbol{0}, \boldsymbol{\chi}) = \beta \Bigg(\sum_{i=1}^{K} \frac{\lambda_i}{m} V(\boldsymbol{e}_i, \boldsymbol{\chi}) + \Big(1 - \frac{\sum_{i=1}^{K} \lambda_i}{m}\Big) V(\boldsymbol{0}, \boldsymbol{\chi}) \Bigg), \tag{6}$$

where $\beta = m/(m + \gamma)$ is the discounting factor in the equivalent uniformized discrete-time model. On the right side of (5), the first term is the uniformized probability of service completion in the next period, μ_k/m, times the expected reward generated by the service completion and the future value under the updated belief state condition on current belief state, the second term is the uniformized probability of customer arrival in the next period, λ_i/m, times the corresponding

[2] The reward $R_{k,n}$ is independent of everything else, as assumed at the beginning of Section 3.1. In particular, the reward does not depends on the policy π, but its contribution to the objective, $e^{-\gamma D^{\pi}_{k,n}} R_{k,n}$, depends on policy π.

value function, and the last term represents the case where no arrival or departure occurs in the next period and the state remains the same.

It is well known that the index policy given by the Gittins index is optimal for an $M/G/1$ queue under general conditions. However, this policy is not applicable when reward distributions are unknown. The classic proof of the optimality of the Gittins index formulates the problem as an arm acquiring bandit problem, which is a special case of multiarm bandit (MAB) problem, in which each customer in the system is modeled as an individual arm and the scheduling problem is to decide which arm to pull (Gittins et al. 2011). However, this argument fails in the system with unknown reward distributions, because when the manager updates the belief about the rewards of customer class k, the beliefs about the rewards of all customers (arms) of class k currently in the system are updated as well. In other words, the states of the other arms are no longer frozen, violating the critical assumption of the MAB. Recognizing the special characteristics of our problem, we take an alternative approach to model the problem as a restless multiarmed bandit problem (RMAB) in the next section.

3.2 Restless Multi-Arm Bandit (RMAB) Approach

In this section, we model each customer *class* as a restless arm. Thus, there are K arms, and the state of arm k is (q_k, χ_k). Every decision involves pulling exactly one arm or staying idle if no customer is in the system. Upon pulling an arm (i.e., choosing to serve a customer of a certain class), the states of all arms evolve concurrently due to the possible arrivals of customers of all classes, making this problem an RMAB problem.

A widely studied solution to the RMAB problem is the Whittle index policy, introduced by Whittle (1988). This approach treats each arm as an separate Markov decision process and computes a corresponding Whittle index, which serves as a measure of the arm's priority for activation. In general, multiple arms can be activated, but in the specific context of our single-server queue, only one arm can be activated in each period.

We formally define the Markov decision process for each arm as follows. For an arm k ($k = 1, 2, \ldots, K$), the state is (q_k, χ_k), and the manager chooses between two actions, A(Active) and P(Passive), to maximize the long-run expected discounted reward from arm k. When active, arm k serves customers and generates a random reward upon service completion. When passive, arm k

collects a deterministic reward, $w_k \geq 0$, known as the *passive reward*, representing the compensation for not serving customers. Intuitively, w_k quantifies how much the manager would be willing to "accept" as a reward to forgo serving customers of class k in the queue. A larger w_k indicates a higher opportunity cost of not serving customers of class k. This passive reward, w_k, serves as a parameter for the Markov decision problem and is critical in determining the Whittle index in the next section.

Specifically, the arrival of the class k customer occurs with probability $\tilde{\lambda}_k = \frac{\lambda_k}{m}$ in all periods, which makes all arms restless. While arm k is active and class k is not empty ($q_k \geq 1$), a service completion of a class k customer occurs with probability $\tilde{\mu}_k = \frac{\mu_k}{m}$, and the state remains unchanged with probability $1 - \frac{\mu_k + \lambda_k}{m}$. Upon service completion, r_k, a realization of R_k, is received, and the belief of the reward, χ_k, is updated to $g_k(\chi_k, r_k)$. If the queue is empty (i.e., $q_k = 0$), and arm k stays active, no reward is received. If arm k is in passive mode, the manager receives a constant reward of w_k per period, a customer arrival occurs with probability $\tilde{\lambda}_k$, and the state remains unchanged with probability $1 - \tilde{\lambda}_k$.

The Bellman equation for arm k when the passive reward is $w_k \geq 0$ can be written as follow: For $(q_k, \chi_k) \in \mathbb{Z}_+ \times \Sigma_k$:

$$J_k(q_k, \chi_k, w_k) = \max\left\{J_k^A(q_k, \chi_k, w_k), J_k^P(q_k, \chi_k, w_k)\right\}, \tag{7}$$

where $J_k(q_k, \chi_k, w_k)$ denotes the maximum long-run expected discount reward when the current state is (q_k, χ_k), $J_k^A(q_k, \chi_k, w_k)$ and $J_k^P(q_k, \chi_k, w_k)$ are the respective long-run expected discounted reward of choosing active or passive mode for one period and then acting optimally. Specifically,

$$J_k^A(q_k, \chi_k, w_k) = \begin{cases} \beta\Big[\tilde{\mu}_k \mathbb{E}\big[R_k + J_k\big(q_k - 1, g_k(\chi_k, R_k), w_k\big) \mid \chi_k\big] & \\ \quad + \tilde{\lambda}_k J_k(q_k + 1, \chi_k, w_k) + (1 - \tilde{\lambda}_k - \tilde{\mu}_k) J_k(q_k, \chi_k, w_k)\Big], & \text{if } q_k \geq 1 \\ \beta\Big[\tilde{\lambda}_k J_k(1, \chi_k, w_k) + (1 - \tilde{\lambda}_k) J_k(0, \chi_k, w_k)\Big], & \text{if } q_k = 0. \end{cases} \tag{8}$$

$$J_k^P(q_k, \chi_k, w_k) = w_k + \beta\Big[\tilde{\lambda}_k J_k(q_k + 1, \chi_k, w_k) + (1 - \tilde{\lambda}_k) J_k(q_k, \chi_k, w_k)\Big]. \tag{9}$$

Note that when there is no class k customers ($q_k = 0$), we have $J_k^P(q_k, \chi_k, w_k) \geq J_k^A(q_k, \chi_k, w_k)$ for any $w_k \geq 0$.

4. The Index Policy for Dynamic Scheduling

This section derives the Whittle index for the RMAB problem discussed in section 3.2. In Section 4.1, we establish the indexability and derive an expression for the Whittle index for the RMAB problem. Furthermore, in Section 4.2, we examine a special case where the index policy based on the Whittle index is optimal. Finally, in Section 4.3, we discuss the optimality of the Whittle index scheduling policy in more general settings.

4.1 Indexability and Whittle Index

The challenge associated with the index-based policy given by the Whittle index is that the index policy is only defined for RMAB problems that are indexable, a condition that is often difficult to establish. Moreover, it is often hard to find an expression for the Whittle index. In this subsection, we establish the indexability of the problem and derive an expression for the Whittle index by studying the problem formulated in (7)-(9).

When the passive reward is w_k, we let $S_k(w_k)$ denote the set of states where the passive action is optimal. Formally,

$$S_k(w_k) := \left\{(q_k, \chi_k) \in \mathbb{Z}_+ \times \Sigma_k : J_k^P(q_k, \chi_k, w_k) \geq J_k^A(q_k, \chi_k, w_k)\right\}. \tag{10}$$

For Whittle index to exist for each state of the restless arm k, Whittle (1988) requires that for each restless arm $k \in \{1 \dots K\}$ to have the following monotonicity property: As the passive reward increases, the collection of states which choose the passive action also increases. Whittle (1988) refers to this property as indexability.

Definition 1. *(Whittle 1988) The restless arm k is indexable if $S_k(w_k)$ is increasing in w_k, namely,*

$$S_k(w_k) \subseteq S_k(w_k'), \qquad \text{for } w_k \leq w_k'.$$

Furthermore, if a restless arm k is indexable, we can find the Whittle index for each state (q_k, χ_k) of arm k. The definition of the Whittle index for state (q_k, χ_k) is given in Definition 2.

Definition 2. *(Whittle 1988) When the restless arm k is indexable, the Whittle index for state $(q_k, \chi_k) \in \mathbb{Z}_+$ is defined as*

$$v_k(q_k, \chi_k) := \inf\left\{w_k : (q_k, \chi_k) \in S_k(w_k)\right\}. \tag{11}$$

Should we establish the indexability of the restless arm k, by Definition 2, the Whittle index for arm k is the minimum passive reward such that the manager is indifferent in choosing active or passive mode for arm k.

In period t, let $Q_k^A(t)$ and $X_k^A(t)$ denote the number of customers of class k in the system and the state of the belief about the reward R_k, respectively, when arm k operates under a special policy, $\tilde{\pi}^A$, which always chooses active action in all states. We define $B_k(q_k)$ as the length of a busy period starting with q_k customers in the system, which is given by:

$$B_k(q_k) = \inf\left\{t \geq 0 : Q_k^A(t) = 0 \;\mid\; Q_k^A(0) = q_k\right\}. \tag{12}$$

The busy period $B_k(q_k)$ represents the number of consecutive periods until the class k queue is empty if it stays in the active mode in all periods.

Let $R_{k,t}^A$ denote the conditional reward based on the manager's belief regarding the reward class k customer up to time t under policy $\tilde{\pi}^A$, formally defined as: $R_{k,t}^A := R_k \mid X_k^A(t)$. Here, $X_k^A(t)$ encapsulates the manager's belief about customers of class k up to and including time t. It is important to note that $R_{k,t}^A$ includes all reward information that has been collected up to and including period t, particularly accounting for any customer departure occurring in period t. Thus, $R_{k,t}^A$ effectively describes the conditional reward to be collected at the next departure after the period t. The remainder of this subsection focuses on proving that the restless arm k is indexable and that its Whittle index can be written as follows: For $q_k > 0$,

$$v_k(q_k, \chi_k) = \beta\tilde{\mu}_k \sup_{1\leq\tau\leq B_k(q_k)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t R_{k,t}^A \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k\right]}, \tag{13}$$

where the supremum is taken over all possible stopping times, τ, constrained to be less than or equal to the busy period $B_k(q_k)$. We define $v_k(q_k, \chi_k) = 0$ for $q_k = 0$ to ensure the completeness of the definition.

Observe that the index defined in (13) bears resembalance to the expression for a Gittins index. Lemma 1 establishes that the index, $v_k(q_k, \chi_k)$, corresponds to the Gittins index of an auxiliary Multi-Armed Bandit (MAB) problem. Corollary 1 then leverages this connection to demonstrate the properties of index $v_k(q_k, \chi_k)$, which serve as important building blocks to establish that index $v_k(q_k, \chi_k)$ is the Whittle index for the RMAB problem.

To construct such an auxiliary MAB problem, we recognize that the functional form of $v_k(q_k, \chi_k)$

is the same as that of a Gittins index (Gittins et al. 2011), except for the part where the stopping time is restricted by the busy period. Therefore, the intuition for the construction of this problem is to force the optimal stopping time within the busy period by "terminating" the arrival process once the queue becomes empty. Furthermore, when an arm is not pulled, its state is frozen, as required by the MAB problems.

Specifically, the auxiliary MAB problem consists of K arms, where only one arm can be pulled in each period. When arm k is pulled, it operates according to the active mode dynamics of the restless arm k described in Section 3.2, except that when the arm is empty (i.e., $q_k = 0$), it remains empty, meaning that no further arrival events occur in the system regardless of the action taken, and generates a reward of 0. Lemma 1 formalizes the construction of the auxiliary MAB problem.

Lemma 1. *The index specified in equation (13) is the Gittins index for an auxiliary MAB with K arms whose states are queue length and belief about the rewards for each class, denoted by $Q_k(t)$ and $X_k(t)$, for all $k \in 1 \dots K$, respectively. In each period, the manager chooses arm k with nonempty queue and serves that class. Conditioning on the states, $Q_k(t) = q_k, X_k(t) = \chi_k$, the state transition follows:*

(i) If $Q_k(t) = 0$, then $(Q_k(t+1), X_k(t+1)) = (0, \chi_k)$;

(ii) If $Q_k(t) \geq 1$, then

$$(Q_k(t+1), X_k(t+1)) = \begin{cases} (q_k + 1, \chi_k) & w.p.\ \tilde{\lambda}, \\ (q_k - 1, \tilde{X}_k) & w.p.\ \tilde{\mu}, \\ (q_k, \chi_k) & w.p.\ 1 - \tilde{\lambda} - \tilde{\mu}, \end{cases}$$

where the random variable $\tilde{X}_k$ is given by $\tilde{X}_k = g(R_k, \chi_k)$.

For the rest of the arms, the states are frozen, i.e. $(Q_j(t+1), X_j(t+1)) = (Q_j(t), X_j(t))$ for $j \neq k$.

Recognizing that the index, $v_k(q_k, \chi_k)$, is a Gittins index for an MAB problem, Corollary 1 summarizes its technical properties to support the proof showing that the index, $v_k(q_k, \chi_k)$, is the Whittle index for the RMAB problem.

Corollary 1. *The following hold:*

(i) The supremum in equation (13) is attainable by a stopping time;

(ii) Let τ be the stopping time that attains the supreme in equation (13). Then, For each (q_k, χ_k) such that $q_k \geq 1$, the stopping time τ has a stopping set, denoted by $\Omega_0(q_k, \chi_k)$, which may be chosen to be at set such that

$$\{(q_k', \chi_k') : v_k(q_k', \chi_k') < v_k(q_k, \chi_k)\} \subseteq \Omega_0(q_k, \chi_k) \subseteq \{(q_k', \chi_k') : v_k(q_k', \chi_k') \leq v_k(q_k, \chi_k)\};$$

In addition to providing the analytical properties necessary to show that $v_k(q_k, \chi_k)$ is the Whittle index for the RMAB problem, Lemma 1 and Corollary 1 also shed light on the computation methods of $v_k(q_k, \chi_k)$. The computation of the Whittle index is often a challenging task, as it is heterogeneous across problems. However, by showing that $v_k(q_k, \chi_k)$ corresponds to the Gittins index of the auxiliary MAB problem, we are equipped well-established methods to compute this index, e.g., the state elimination method, the largest remaining index algorithm, etc.

Importantly, Lemma 2 demonstrates that the index, $v_k(q_k, \chi_k)$, is increasing in the number of customers in the system of that class.

Lemma 2. *$v_k(q_k, \chi_k)$ is increasing in q_k.*

As established in Lemma 2, the index $v_k(q_k, \chi_k)$ (later shown to be the Whittle index) increases as the queue length of that class increases. This observation suggests that the manager's incentive to explore a customer class with unknown reward parameters also increases with queue length.

We are now in the position to establish Theorem 1 on the indexability of the RMAB problem and the index given in (13) being the Whittle index for the dynamic scheduling problem.

Theorem 1. *For all $k \in \{1, 2, \ldots, K\}$, we have*
(i) The arm k is indexable;
(ii) $v_k(q_k, \chi_k)$ defined in Equation (13) is the Whittle index of the restless arm k defined in Equation (11).

Theorem 1 establishes the basis for the Whittle index scheduling policy in the single-server queueing system with the learning of service rewards, formally defined below.

Definition 3 (Whittle Index Scheduling Policy)**.** *Let the Whittle index for class k be $v_k(q_k, \chi_k)$ as expressed in (13). Upon each arrival or departure of a customer (with at least one customer in the*

system after the departure), the Whittle index scheduling policy selects the customer class k^* with the highest Whittle index to be served. Formally, this is expressed as:

$$k^* \in \arg \max_{k \in \{1,\ldots,K\}} v_k(q_k, \chi_k),$$

and ties are resolved by selecting at random with equal probability among the tied classes.

While this policy is not generally optimal for the queueing scheduling problem with Bayesian learning, as formulated by (5)-(6), it is optimal in the specific case of two customer classes: one with a known reward distribution and the other with an unknown reward distribution, as discussed in the next subsection.

4.2 Optimality of the Whittle Index Scheduling Policy for Two-Class System

This subsection considers a system with two classes of customers, that is, $K = 2$. In particular, we consider that the system consists of one class of customers whose distribution of rewards is known to the manager and one class of customers with unknown reward distribution. We prove that the Whittle index scheduling policy in Definition 3 is optimal for this special case.

In this system, we assume that the manager knows the distribution of the reward R_1 of class 1 and learns about the reward distribution parameters for class 2 customers, R_2, via Bayesian updating. The following lemma provides a characterization of the Whittle index of class 1.

Lemma 3. *Suppose class 1 customers have a known reward distribution with mean r_1. Then, the Whittle index of class 1 is given as follows:*

$$v_1(q) = \begin{cases} \beta\tilde{\mu}_1 r_1, & q \geq 1, \\ 0, & q = 0. \end{cases}$$

Note that if there is at least one customer of class 1 in the system, more arrivals of class 1 customers do not change the Whittle index of class 1. This contrasts with Lemma 2 in that the Whittle index generally increases in the number of customers in the system.

We now establish that the non-idling scheduling policy that follows the Whittle index is optimal for this special case.

Theorem 2. *The Whittle index scheduling policy is optimal for the single-server queueing system with two classes: one customer class with known reward distribution and the other customer class with unknown reward distribution.*

Theorem 2 holds for any general belief model of unknown class rewards.

For the rest of this subsection, we discuss a special reward model in which the belief parameter is one-dimensional. Specifically, we assume that the reward for a class-2 customer is modeled as a random variable following a Bernoulli distribution with an unknown parameter θ, which governs the probability of generating either a high reward, r_h, or a low reward, r_l, where $0 \le r_l \le r_h$. The parameter θ can take one of two possible values: p_m or p_b, with $0 \le p_m \le p_b \le 1$. The scenario corresponding to $\theta = p_m$ (m for "malo") is referred to as the pessimistic scenario, while the scenario corresponding to $\theta = p_b$ (b for "bono") is referred to as the optimistic scenario. Using the Bayesian model described in Section 3.1, the reward probability distributions are given by:

$$p(r_h \mid \theta = p_m) = p_m, \quad p(r_l \mid \theta = p_m) = 1 - p_m;$$

$$p(r_h \mid \theta = p_b) = p_b, \quad p(r_l \mid \theta = p_b) = 1 - p_b.$$

The manager's belief state is represented by a scalar $\chi \in (0, 1)$, which denotes the probability that class-2 customers are in the optimistic scenario. Specifically, the probability that θ takes the value p_m is given by $f(\theta = p_m) = 1 - \chi$, while the probability that θ takes the value p_b is given by $f(\theta = p_b) = \chi$. This structure is commonly referred to as the Bernoulli-Bernoulli prior-posterior structure.

Let $\overline{\chi}$ and $\underline{\chi}$ denote the Bayesian posterior belief updates corresponding to a prior belief χ, depending on whether the system manager observes r_h or r_l upon service completion. These updates are given by:

$$\overline{\chi} = g(\chi, r_h) = \frac{\chi p_b}{\chi p_b + (1-\chi)p_m},$$
$$\underline{\chi} = g(\chi, r_l) = \frac{\chi(1-p_b)}{\chi(1-p_b) + (1-\chi)(1-p_m)}.$$

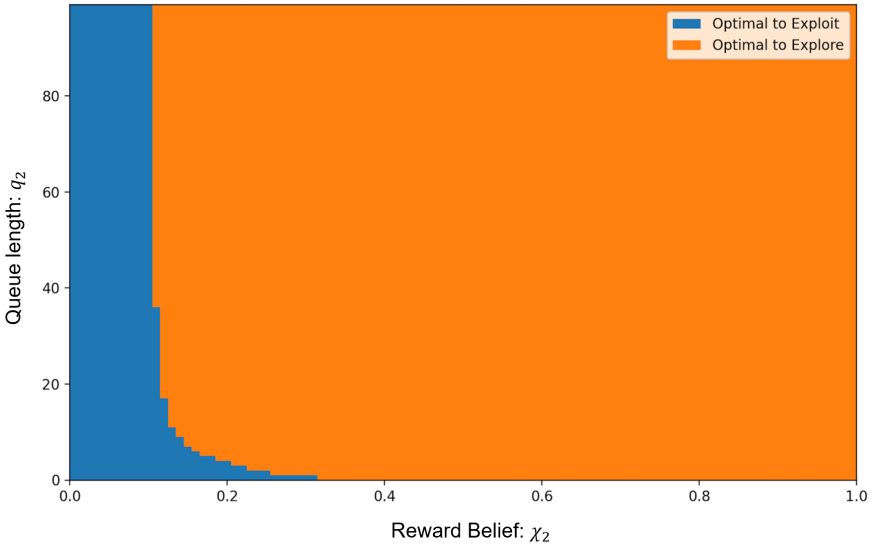
The expected reward $\overline{r}_2(\chi)$ when the system manager's belief is χ is given as follows: For $\chi \in (0, 1)$,

$$\overline{r}_2(\chi) = \mathbb{E}[R_2 \mid \chi] = \chi(p_b r_h + (1-p_b)r_l) + (1-\chi)(p_m r_h + (1-p_m)r_l).$$

To avoid trivial cases, we assume that $\mu_2 \overline{r}_2(0) \le \mu_1 r_1 \le \mu_2 \overline{r}_2(1)$. Lemma 4 states the monotonicity property of the Whittle index $v_2(q, \chi)$ on the scalar χ. In addition, this lemma also provides lower and upper bounds of the Whittle index $v_2(q, \chi)$.

Lemma 4. *The index $v_2(q, \chi)$ increases in χ. Moreover, the following holds: For $q \ge 1$ and*

Figure 1: Optimal scheduling policy for a class of customer with known reward and a class of customer with unknown reward parameters—$\lambda_1 = 3, \lambda_2 = 3, \mu_1 = 9, \mu_2 = 10, r_1 = 3, r_h = 4, r_l = 2, p_b = 0.6, p_m = 0.2$.



$\chi \in [0,1]$, $\beta \bar{r}_2(\chi)\mu_2 \leq v_2(q,\chi) \leq \bar{v}_2(\chi)$ *where*

$$\bar{v}_2(\chi) = \beta\tilde{\mu}_2 \sup_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t R_{2,t} \mid X_2(0) = \chi\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \mid X_2(0) = \chi\right]}.$$

Thus, Theorem 2 and Lemma 4 immediately lead to the following corollary stating that optimal policy can be characterized by a switch curve $q_2(\chi)$.

Corollary 2. *There exists a decreasing function $q_2 : [0,1] \to [1, +\infty]$ that characterizes the optimal policy: The server follows a nonidling policy and serves class 2 if $Q_2(t) \geq q_2(X_2(t))$. In particular, $q_2(\chi) = 1$ if $\tilde{\mu}_2\bar{r}_2(\chi) \geq \tilde{\mu}_1 r_1$ and $q_2(\chi) = \infty$ if $\bar{v}_2(\chi) \geq \beta\tilde{\mu}_1 r_1$.*

Figure 1 illustrates an example function from Corollary 2 that characterizes the optimal scheduling policy. The blue region denotes the states where the optimal action is to serve class 1 customer; the orange region denotes the states where the the optimal action is to serve class 2 customer. The boundary between the two regions showcases the decreasing function that characterizes the optimal scheduling policy, as specified in Corollary 2. The bottom left part of the curve shows that for a given belief, χ, if $Q_2(t) \geq q_2(\chi)$ the optimal policy is to serve class 1, otherwise the optimal policy is to serve class 2.

Moreover, for this particular reward model, we observe that calculating the Whittle index can be significantly simplified.

Lemma 5. *Suppose the rewards r_h and r_l for both types of customers undergo a linear transformation defined by $f(x) = ax + b$, resulting in transformed rewards $r'_h = ar_h + b$ and $r'_l = ar_l + b$. Then, the Whittle index for class $k \in \{1, \ldots, K\}$, denoted by $v_k(q_k, \chi_k \mid r'_h, r'_l)$, satisfies:*

$$v_k(q_k, \chi_k \mid r'_h, r'_l) = a \cdot v_k(q_k, \chi_k \mid r_h, r_l) + b.$$

In other words, the Whittle index under the transformed rewards is equivalent to the linear transformation of the original Whittle index.

Lemma 5 provides an efficient method for determining the Whittle index for any combination of rewards. Suppose that we have computed the Whittle index for specific rewards r_h and r_l and want to determine the index policy for a different set of rewards r'_h and r'_l. It suffices to solve the system of equations $r'_h = ar_h + b$ and $r'_l = ar_l + b$, then determine the corresponding index policy by the transformed Whittle index.

Using a special case of a two-class system with unknown reward distributions, Lemma 6 shows that the Whittle index responds differently to changes in the service rate μ and the expected reward r, even when their product $r\mu$ remains constant. This contrasts with classical queue scheduling results, such as the $r\mu$ rule.

Lemma 6. *Consider a system with $K = 2$ classes of customers, whose reward distributions are unknown and follow a Bernoulli-Bernoulli structure. In the ground truth, the expected reward for class 1 is H, and for class 2 it is L, with $H > L$. The system manager knows the value of H, L, μ_1, and μ_2, but learns which class generates the higher reward. If μ_1 increases by a factor $a > 1$ and H decreases by a factor $\frac{1}{a}$, keeping $r_1\mu_1 = H\mu_1$ constant, the Whittle index for class 1 strictly increases, while that for class 2 strictly decreases.*

Lemma 6 underscores the adaptability of the Whittle index in systems with changing parameters, particularly on changing service rates. Unlike classical scheduling rules such as the $r\mu$ rule, which treat the product of the reward and service rate as the sole determinant of priority, the Whittle index responds dynamically to changes in system parameters. By increasing the index value with service rates, the Whittle index policy implicitly focuses on facilitating the system's ability to gather information about uncertain rewards faster. In practical terms, this adaptability makes the Whittle index particularly effective in environments where service rates and rewards changes. We

numerically demonstrate the implication of this results in section 5.4.

4.3 Whittle Index Scheduling Policy in the General System

In general settings, e.g., a single-server system with two known classes and one unknown class, or both classes with unknown rewards, the Whittle index scheduling policy is not necessarily optimal; we illustrate this with an example.

For example, consider a system with three customer classes: two with known rewards and a third with unknown rewards that are learned over time. Suppose that one of the known classes, class 1, generates a reward r_1 upon completion of the service, which is significantly higher than the reward of the other known class (class 2), denoted r_2, such that $r_1 \gg r_2$. In addition, we assume that class 1 has a very high traffic intensity. Furthermore, the reward r_1 exceeds the upper bound of the index for the unknown class (class 3): $\beta\tilde{\mu}_1 r_1 > \bar{v}_3(\chi)$. In this scenario, it is optimal to prioritize class 1 whenever its queue is not empty, regardless of reward beliefs or queue lengths. In a state where the queue for class 1 is empty, but the queues for classes 2 and 3 are not, and class 2 has a higher expected reward than class 3 based on current beliefs, the Whittle index scheduling policy might recommend serving a class 3 customer. This decision would forgo the immediate higher reward from class 2 in favor of exploring the unknown reward parameters of class 3. However, such exploitation-for-exploration trade-offs can be suboptimal. After severing the class 3 customer for the period, instead of utilizing the information learned from exploration, the system will spend a long time serving class 1 customers in the next future periods, as class 1 has a very high traffic intensity. As a result, the potential value of exploring class 3 can be significantly diminished, as the discounting of future rewards reduces the value of learning as the system will be busy serving class 1. In this scenario, it is suboptimal to forgo immediate rewards from class 2 in order to explore class 3, because the value of exploration is significantly reduced by the extended periods spent serving class 1 customers, whereas collecting the higher immediate reward could increase the total discounted reward.

Although the Whittle index scheduling policy does not guarantee an optimal solution in general settings, it remains near-optimal, as we will show in Section 5.4 and is computationally efficient. The portion of states where the action of the index policy diverges from the optimal policy, as described in the above cases, is small, so the Whittle index scheduling policy performs well overall. In

terms of computational efficiency, the Whittle index for each class can be computed independently. Therefore, increasing the number of classes in the system leads to a linear increase in computation for the Whittle index, in contrast to the exponential growth in computation associated with the dimensionality of the state space.

5. Numerical Studies

In this section, we explore the numerical properties of the Whittle index we derived in Section 4.1 and evaluate the performance of the Whittle index scheduling policy using Monte Carlo simulation. We begin by reviewing the Bernoulli-Bernoulli prior-posterior structure employed throughout our numerical study in Section 5.1. In Section 5.2, we examine the sensitivity of the Whittle index to the state of the system, including the queue length and the reward belief. Section 5.3 analyzes the impact of arrival and service rates on the Whittle index. Finally, in Section 5.4, we demonstrate that the Whittle index scheduling policy performs near optimally and outperforms classical heuristics in various parameter settings.

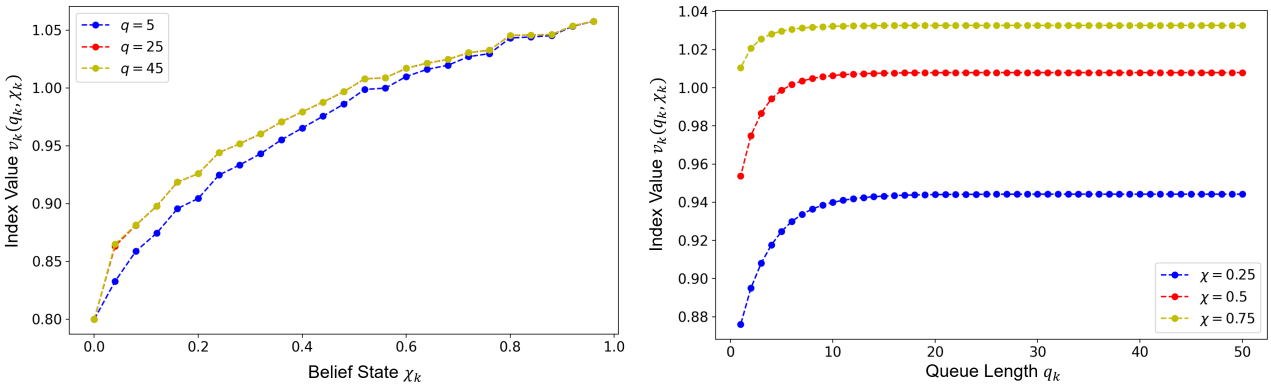
5.1 Numerical Setting

In Section 5, we examine the changes in the Whittle index values and the performance of the Whittle index scheduling policy, assuming that a class with an unknown reward distribution follows the Bernoulli-Bernoulli prior-posterior structure described in Section 4.2. Specifically, we model the reward for a class-2 customer as a Bernoulli random variable with an unknown parameter θ, which governs the probability of a high reward r_h or a low reward r_l $(0 \leq r_l \leq r_h)$. The parameter θ can take two values: p_m (pessimistic) or p_b (optimistic), where $0 \leq p_m \leq p_b \leq 1$. The manager's belief state, $\chi \in (0, 1)$, represents the probability that class-2 customers are in the optimistic scenario. Specifically, $f(\theta = p_m) = 1 - \chi$ and $f(\theta = p_b) = \chi$. We select this prior-posterior model because it is one of the few where the manager's belief state is one-dimensional, enabling an intuitive demonstration of the numerical results.

5.2 Sensitivity of the Whittle Index to States

In this section, we analyze how the Whittle index varies across system states (q_k, χ_k) for a class k customer, whose reward distribution is unknown and learned by the manager using the Bernoulli-Bernoulli prior-posterior structure described in Section 5.1. The unknown class k has the following

Figure 2: Sensitivity of Whittle Index to System States



(a) Sensitivity of Whittle Index to Belief

(b) Sensitivity of Whittle Index to Queue Length

parameters: In the optimistic scenario, 60% of customers are high-reward, i.e., $p_b = 0.6$, while in the pessimistic scenario, 40% are high-reward, i.e., $p_m = 0.4$. For class k, the arrival rate is $\lambda_k = 4$ and the service rate is $\mu_k = 10$.

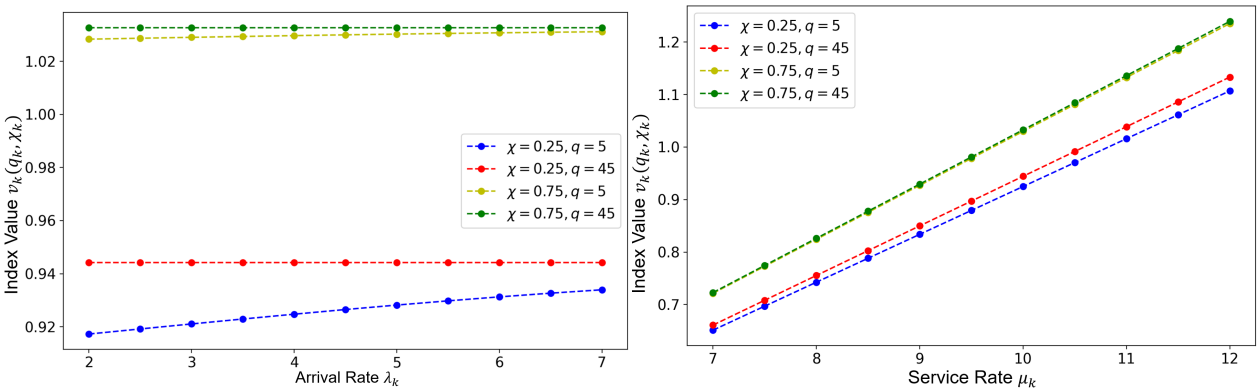
Figure 2(a) demonstrates the changes in $v_k(q_k, \chi_k)$ with respect to χ_k for queue lengths of 5, 25, and 45. The index value increases with belief across all queue lengths, with shorter queue lengths being more sensitive to these increases. Furthermore, the relationship between index value and belief exhibits a concave shape, and shorter queue lengths reach the platform at a slower rate. In particular, in belief states 0 or 1, all queue lengths yield the same index value, regardless of the queue length. This is due to the structure of the prior distribution: at 0 or 1, the belief remains the same for all future state transitions.

Figure 2(b) illustrates how $v_k(q_k, \chi_k)$ changes with respect to q_k using the same example, with samples in beliefs of low, medium and high reward. As demonstrated in Lemma 2, the index $v_k(q_k, \chi_k)$ increases with q_k. Figure 2(b) also shows that the index value is concave with respect to the queue length. In states with higher beliefs, the effect of the queue length on the index diminishes more rapidly. This is because in higher belief states, the busy period is less likely to become a binding constraint, which aligns with our discussion of Figure 2(a).

5.3 Sensitivity of the Whittle Index to Queue Parameters

In this section, we examine how the Whittle index changes with respect to the queueing parameters λ_k and μ_k for a class k customer, whose reward distribution is unknown and follows the Bernoulli-Bernoulli prior-posterior structure described in Section 5.1. The reward parameters in this study

Figure 3: Sensitivity of Whittle Index to Queue Parameters



(a) Sensitivity of Whittle Index to Arrival Rate

(b) Sensitivity of Whittle Index to Service Rate

are the same as those in Section 5.2, namely $p_b = 0.6$ and $p_m = 0.4$. To explore the impact of queueing parameters, we sample four distinct states that represent a combination of long/short queue lengths and high/low reward belief states.

Figure 3 shows the Whittle index value at different arrival and service rates for a given class. Figure 3(a) plots the index value of a class of customers with unknown rewards with the Bernoulli prior structure , denoted by k, in four different states: $(5, 0.25)$, $(45, 0.25)$, $(5, 0.75)$, and $(45, 0.75)$ representing states with long/short queue lengths and high/low reward beliefs where the service rate μ_k is 10 and arrival rates varying from 2 to 7 with an increment of 0.5. The states are represented by different colors in the figure. Figure 3(a) shows that the arrival rate increases the index value in states with a short queue length and has little effect in states with a long queue. The monotonicity follows from the property of the Whittle index $v_k(q_k, \chi_k)$ in equation (13) that a higher arrival rate stochastically increases the busy period in $v_k(q_k, \chi_k)$ and therefore increases the index value. For states with a short queue, the increase in the busy period by arrival rates is more significant; on the other hand, for states already with a long queue, the busy period in $v_k(q_k, \chi_k)$ is not likely to be a binding restriction; therefore, an increase in arrival rates has little effect on the index value. Furthermore, the increase in index value in the arrival rate is less significant in high belief states. For a high belief state, the optimal stopping time is shorter than that of a low belief state because a high belief state could possible transition to more states that lower rewards, i.e. a larger stopping set. Therefore, in a high belief state, it is less likely that the busy period in $v_k(q_k, \chi_k)$ is a binding restriction.

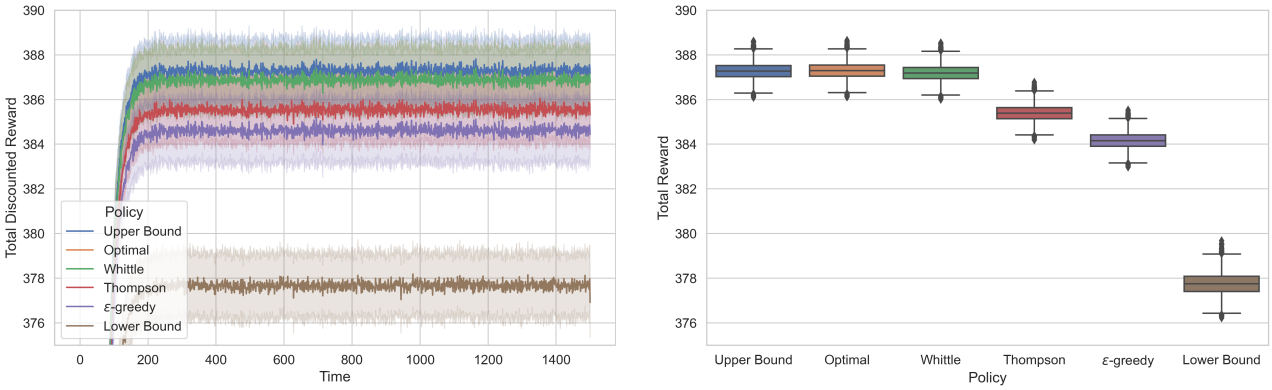
Similarly, Figure 3(b) shows the index values of a class of customers with unknown rewards, denoted by k, in the same four states where the service rate λ_k is 4 and the service rates μ_k increase from 7 to 12 with an increment of 0.5. Figure 3(b) shows that a higher service rate leads to a higher index value, since faster service would lead to a higher reward per unit collected. Interestingly, we observe that the slope of the index value with respect to the increase in service rate is state-dependent, particularly when the reward belief is low ($\chi = 0.25$). In such cases, the index value for states with longer queues increases more rapidly with the service rate. As the service rate increases from 7 to 12, we see that the state with a shorter queue length increases slower than that with a longer queue length. This is because the busy period in $v_k(q_k, \chi_k)$ is shorter with a faster service rate, which is more likely to be a binding restriction when the queue length is shorter. However, we see that the index value in high belief states increases similarly to the $r\mu$ rule, where the queue length has little impact on the index value, when the reward belief is high enough. This result is supported by the index change curve that overlaps with the service rates in states $(5, 0.75)$ and $(45, 0.75)$ in Figure 3(b).

5.4 Performance of Whittle Index Scheduling Policy

In this section, we evaluate the performance of the Whittle index policy against different benchmark scheduling policies. Using Monte Carlo simulations, we assess the performance of these policies based on their total discounted rewards. Our baseline analysis demonstrates that the Whittle index scheduling policy achieves near-optimal performance across diverse scenarios and identifies parameter regions where it significantly outperforms classical learning heuristics. Furthermore, by leveraging the properties of the Whittle index identified in Section 4.2, we demonstrate that it dynamically adjusts to changes in the expected reward r and the service rate μ even if their product $r\mu$ stays the same, highlighting the unique roles of r and μ in the queue scheduling problem with learning of customer characteristics.

In our simulation, we consider a system with two customer classes ($K = 2$), each with an unknown reward distribution to be learned by the manager using the Bernoulli-Bernoulli prior-posterior structure described in Section 5.1. For both classes, each customer generates either a high reward of $r_h = 20$ or a low reward of $r_l = 2$. In the optimistic scenario, 60% of customers generate high rewards ($p_b = 0.6$), while in the pessimistic scenario, 40% generate high rewards

Figure 4: Performance of Whittle Index Scheduling Policies Compared to Becnchmark Policies



(a) Total Discounted Reward along Time

(b) Total Discounted Reward Distribution

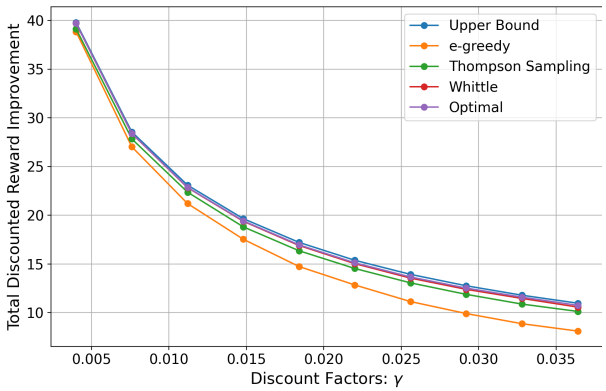
Note: In Figure 4(a), the trajectories of the optimal policy and the Whittle index scheduling policy closely align, as their actions are nearly identical.

($p_m = 0.4$). We assume that class 1 belongs to an optimistic reward scenario, and class 2 belongs to an optimistic reward scenario. Note that this information is the ground truth and is not known by the system manager. In this case, giving class 1 customer priority to the service results in an upper bound to the total discounted reward, and giving class 2 customer priority to the service results in a lower bound to the total discounted reward.

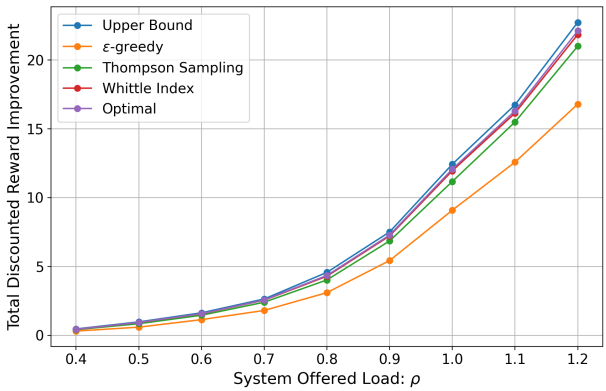
We consider five benchmark policies to evaluate the performance of Whittle index scheduling policies: the upper and lower bound policies, the optimal policy by solving the dynamic programming via value iteration methods, the Thompson sampling policy (Thompson 1933) and the ϵ-greedy policy Sutton and Barto (1998). Classical heuristics such as Thompson Sampling and the ϵ-greedy policy in our problem make scheduling decisions based on the reward belief χ. Specifically, at each arrival or departure of the customer, Thompson sampling draws a random sample ξ_k for each class $k = 1$ or 2. ξ_k takes value p_b with probability χ_k and takes value p_m with probability $1 - \chi_k$ ($p_b > p_m$). The system manager chooses the class with the $\mu_k(\xi_k r_h + (1 - \xi_k)r_l)$ value to serve. For the ϵ-greedy algorithm, let $r(\chi_k) = \chi_k(p_b r_h + (1 - p_b)r_l) + (1 - \chi_k)(p_m r_h + (1 - p_m)r_l)$ denote the expected reward given the belief χ_k, with probability $1-\epsilon$, the algorithm serves the class with the highest $r_k(\chi_k)\mu_k$ to serve, and with probability ϵ, the algorithm serves each nonempty class with equal probability.

Figure 4 compares the performance of the Whittle index scheduling policy with benchmark

Figure 5: Reward Improvement at Different Discount Factor (γ) and Offerled Load (ρ)



(a) Reward Improvement with Discount Rates



(b) Reward Improvement with Traffic Intensities

policies in terms of total discounted reward. In this analysis, arrival rates are set to $\lambda_1 = 1.5$ and $\lambda_2 = 2$, while service rates are $\mu_1 = 3$ and $\mu_2 = 4.1$, ensuring a overall traffic intensity of the system at $\rho = 0.98$. Figure 4(a) shows the average total discounted reward over time with a 95% confidence intervals, highlighting that the Whittle index closely approximates the performance of the optimal policy, while the performance of classical heuristic policies falls behind with a significant gap. Figure 4(b) presents box plots of the total reward distribution, showing that the Whittle policy achieves a reward distribution similar to the optimal policy with lower variability compared to Thompson Sampling and ε-greedy.

Figure 5(a) illustrates the average improvement in the total discounted reward of all policies compared to the lower bound at various discount factors (γ). In this analysis, the parameters are the same as in the previous study except for the discount rates, where the traffic intensity of the system is maintained at $\rho = 0.98$. We analyze discount factors ranging from 0.004 to 0.04, corresponding to valuations where one unit of reward translates to 0.996 and 0.96 units of reward after one time period, respectively.

As shown in Figure 5(a), we observe that the Whittle index scheduling policy consistently outperforms traditional learning strategies, such as the ϵ-greedy and Thompson Sampling algorithms, while achieving near-optimal performance across all discount factors. Additionally, the performance gap between the Whittle index scheduling policy and the classical learning policies widens as the discount rate increases. These findings can be attributed to the adaptive and near-optimal nature of the Whittle index scheduling policy. Whittle index scheduling policy facilitates faster learning of

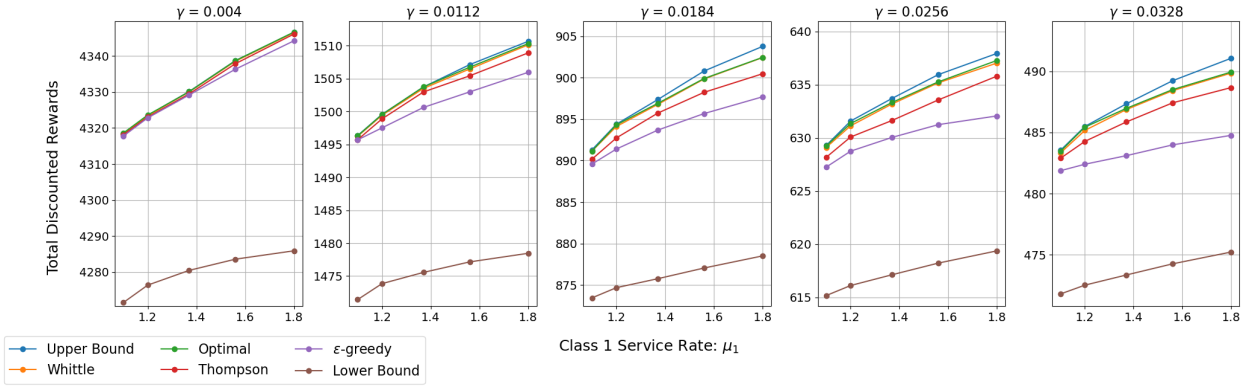
the underlying reward structure by dynamically accounting for discount rates. Unlike ϵ-greedy and Thompson Sampling algorithms, the Whittle index policy incorporates discount rates directly into its formulation, maintaining near-optimal performance regardless of the discount rate. To illustrate the faster learning provided by the Whittle index, consider the Thompson Sampling algorithm. Like the Whittle index policy, Thompson Sampling eventually learns the ground truth and adheres to the optimal action after sufficient learning. However , the better performance of the Whittle index policy indicates its ability to learn the ground truth faster.

Figure 5(b) presents the average increase in the total discounted reward for all policies compared to the lower bound at various levels of traffic intensity (ρ). For this study, the service rates increase, while the arrival rates are the same as in the baseline study. Consequently, the system's offered load $\rho = \rho_1 + \rho_2$ spans values from $\{0.4, 0.5, \ldots 1.2\}$. We specify that each class has an equal offered load of $\rho_1 = \rho_2$ and the discount rate is consistently $\gamma = 0.004$ throughout this example. As noted in earlier examples, the Whittle index scheduling policy consistently displays near-optimal performance in all traffic conditions. This robustness is due to the way that the Whittle index accounts for changes in service rates and offered loads via its busy period component. In addition, higher traffic intensities further highlight the advantages of the Whittle index scheduling policy.

Through the numerical studies, we also explore the impact of the service rate and rewards plays in the queueing scheduling problem with learning. The $r\mu$ rule in the classical queueing scheduling problem suggests that the priority rule depends only on the product of the reward r and the service rate μ. However, in the learning problem, we have shown that for a class k, for a given product $r_k\mu_k$ in the ground truth, the Whittle index increases in μ_k; see Lemma 6 for details. While the Whittle index policy adjusts for the change of the $r\mu$ compositions, the classical policies of Thompson sampling and ϵ-greedy do not, i.e., for a given state, Thompson sampling and ϵ-greedy policies have the same probability of serving a class k with the same probability regardless of the $r\mu$ composition.

To explore the implications of this result, we keep the product $r\mu$ constant for a class of customers and examine the performance of the policy in different compositions of r and μ. We increase the service rate μ_1 and decrease the portion of the high-reward customer of class 1 accordingly so that the product $r_1\mu_1$ is constant. Note that here the expected reward for class 1, in ground truth,

Figure 6: Total Discounted Rewards of Different Policies at Different $r_1\mu_1$ compositions



is defined by $r_1 = p_1 r_h + (1 - p_1) r_l$. To control for the impact of the traffic intensity so that the load offered for class 1, $\rho_1 = 0.5$ throughout all simulations, we increase the arrival rate of class 1, λ_1, accordingly. In addition, all the parameters for class 2 customers remain unchanged. In this study, regardless of the parameters of the change, given the ground truth, prioritizing class 1 generates the most total discounted rewards. Figure 6 illustrates the total discounted rewards of different policies at varying discount rates as we simultaneously increase the value of μ_1 and decrease r_1 at the same time. We observe that the Whittle index scheduling policy consistently remains near-optimal, closely approaching the upper bound across all parameter configurations. Although the total discounted rewards of the Thompson sampling and ϵ-greedy policies increase with higher service rates, the performance gap between these policies and the Whittle index policy widens as μ_1 increases. In addition, the differences in policy performance are amplified by higher discount factors, suggesting that most of the divergence occurs early in the service process when policies are still learning the underlying reward structure.

The increasing performance advantage of the Whittle index scheduling policy over Thompson sampling and ϵ-greedy policies can be explained by its dynamic adaptability and efficient learning process. For any given state, the Whittle index policy adjusts to changes in the composition of $r\mu$ by prioritizing class 1 more frequently as μ_1 increases. In contrast, the Thompson sampling and ϵ-greedy policies maintain fixed probabilities of serving each class in all states, regardless of parameter changes. Moreover, when μ_1 is higher, serving class 2 customers delays learning about class 1 rewards because the time spent serving class 2 could have been used to exploit the higher

service rates of class 1. The Whittle index policy, by favoring class 1 in such scenarios, consistently outperforms the other two policies. This can be attributed to the adaptability of the Whittle index policy to different parameter setups, allowing it to capitalize on increased service rates for class 1 more effectively than Thompson sampling and ϵ-greedy policies.

6. Conclusion

In this paper, we formulate the queueing scheduling problem with Bayesian updating of the reward parameters as an RMAB problem. Our analysis provides the corresponding Whittle index and demonstrates that the scheduling policy given by the Whittle index is optimal in a specific case and remains near optimal in general settings. We also provide numerical analysis on the robustness of the Whittle index policy under varying conditions and assumptions. Conceptually, our paper offers a framework for studying optimal Bayesian queueing scheduling problems with learning aspects. Practically, we provide a Whittle index scheduling policy whose computational complexity grows linearly with the number of classes instead of exponentially, as in classical dynamic programming algorithms. This improvement in solution efficiency offers a feasible solution for real-world applications.

Future work on learning in queueing could explore more complex queueing systems with learning of other types of information and the impact of customer behaviors. A service system with customer abandonment is particularly interesting, as customer abandonment leads to both revenue loss and loss of learning opportunities. Additionally, exploring optimal Bayesian scheduling policies where the system manager learns both service rates and service rewards could extend insights for data-driven decision making in learning-based queueing systems. Moreover, learning of customer characteristics in a queueing network and the resulting routing problem provide another interesting extension of our paper. One could investigate settings where services at different stations generate different rewards with respect to customer types, where rewards correlated are correlated, and knowledge about customer is shared among stations. With active learning of the customer reward parameters in a network, the system manager needs to decide the routing of the customer, that is, the optimal order of customer visiting stations so that the total discounted reward is maximized.

References

Aalto, S. 2024. Whittle index approach to multiserver scheduling with impatient customers and DHR service times. *Queueing Systems* **107** 1–30.

Afèche, P., B. Ata. 2013. Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. *Manufacturing & Service Operations Management* **15**(2) 292–304.

Aktekin, T., R. Soyer. 2012. Bayesian analysis of queues with impatient customers: Applications to call centers. *Naval Research Logistics* **59**(6) 441–456.

Ala, A., F. Chen. 2022. Appointment scheduling problem in complexity systems of the healthcare services: A comprehensive review. *Journal of Healthcare Engineering* **2022**(1) 5819813.

Ansell, P., K. D. Glazebrook, J. Nino-Mora, M. O'Keeffe. 2003. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research* **57**(1) 21–39.

Argon, N. T., L. Ding, K. D. Glazebrook, S. Ziya. 2009. Dynamic routing of customers with general delay costs in a multiserver queuing system. *Probability in the Engineering and Informational Sciences* **23**(2) 175–203.

Asanjarani, A., Y. Nazarathy, P. Taylor. 2021. A survey of parameter and state estimation in queues. *Queueing Systems* **97**(1) 39–80.

Ayesta, U., P. Jacko, V. Novak. 2017. Scheduling of multi-class multi-server queueing systems with abandonments. *Journal of Scheduling* **20**(2) 129–145.

Bernardo, J. M., A. F. Smith. 2009. *Bayesian Theory*. John Wiley & Sons.

Bouneffouf, D., I. Rish, C. Aggarwal. 2020. Survey on applications of multi-armed and contextual bandits. *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1–8.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* **100**(469) 36–50.

Buyukkoc, C., P. Varaiya, J. Walrand. 1985. The cμ rule revisited. *Advances in Applied Probability* **17**(1) 237–238.

Chen, X., Y. Liu, G. Hong. 2024. An online learning approach to dynamic pricing and capacity sizing in service systems. *Operations Research* **72**(6) 2677–2697. doi:10.1287/opre.2020.0612. URL https://doi.org/10.1287/opre.2020.0612.

Coro. 2023. Account prioritization: What it is and why it's important for sales teams. URL https://www.coro.io/en/blog/account-prioritization.

Cox, D., W. L. Smith. 1961. *Queues, Methuen & Co*, vol. 2. Spottiswoode, Ballantyne & Co. Ltd.

Déry, J., A. Ruiz, F. Routhier, V. Bélanger, A. Côté, D. Ait-Kadi, M.-P. Gagnon, S. Deslauriers, A. T. Lopes Pecora, E. Redondo, et al. 2020. A systematic review of patient prioritization tools in non-emergency healthcare services. *Systematic reviews* **9** 1–14.

Freund, D., T. Lykouris, W. Weng. 2023. Quantifying the cost of learning in queueing systems. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine, eds., *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 6532–6544. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1502957929fc4257dd1b6daf7d869c2f-Paper-Conference.pdf.

Gittins, J., K. Glazebrook, R. Weber. 1989. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons.

Gittins, J., K. Glazebrook, R. Weber. 2011. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons.

Glazebrook, K. D., R. Lumley, P. Ansell. 2003. Index heuristics for multiclass $M/G/1$ systems with nonpreemptive service and convex holding costs. *Queueing Systems* **45**(2) 81–111.

Harrison, J. M. 1975a. Dynamic scheduling of a multiclass queue: Discount optimality. *Operations Research* **23**(2) 270–282.

Harrison, J. M. 1975b. A priority queue with discounted linear costs. *Operations Research* **23**(2) 260–269.

Hirayama, T., M. Kijima, S. Nishimura. 1989. Further results for dynamic scheduling of multiclass $G/G/1$ queues. *Journal of Applied Probability* **26**(3) 595–603.

Klimov, G. P. 1975. Time-sharing service systems. i. *Theory of Probability & Its Applications* **19**(3) 532–551.

Krishnasamy, S., A. Arapostathis, R. Johari, S. Shakkottai. 2018. On learning the $c\mu$ rule in single and parallel server networks. *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 153–154.

Krishnasamy, S., R. Sen, R. Johari, S. Shakkottai. 2021. Learning unknown service rates in queues: A multiarmed bandit approach. *Operations research* **69**(1) 315–330.

Lai, T. L., Z. Ying. 1988. Open bandit processes and optimal scheduling of queueing networks. *Advances in Applied Probability* **20**(2) 447–472.

Lingenbrink, D., K. Iyer. 2019. Optimal signaling mechanisms in unobservable queues. *Operations research* **67**(5) 1397–1416.

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

Oliveira, M., V. Bélanger, I. Marques, A. Ruiz. 2020. Assessing the impact of patient prioritization on operating room schedules. *Operations Research for Health Care* **24** 100232.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Roy, D., E. Spiliotopoulou, J. de Vries. 2023. How data-driven decisions help restaurants stay competitive. *Harvard Business Review* .

Scully, Z., I. Grosof, M. Harchol-Balter. 2020. The Gittins policy is nearly optimal in the $M/G/k$ under extremely general conditions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* **4**(3) 1–29.

Scully, Z., M. Harchol-Balter. 2021. The Gittins policy in the $M/G/1$ queue. *2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*. IEEE, 1–8.

Strum, D. P., J. H. May, L. G. Vargas. 2000. Modeling the uncertainty of surgical procedure times. *Anesthesiology* **92**(5) 1284–1297.

Sutton, R. S., A. G. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Tcha, D.-W., S. R. Pliska. 1977. Optimal control of single-server queuing networks and multi-class $M/G/1$ queues with feedback. *Operations Research* **25**(2) 248–258.

Teguh, W., I. Setiawan. 2012. Analysis of expected and actual waiting time in fast food restaurants. *Service Science* **4**(3) 216–227.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**(3/4) 285–294.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability* **5**(3) 809–833.

Walton, N., K. Xu. 2021. Learning and information in stochastic networks and queues. *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*. INFORMS, 161–198.

Weiss, G. 1988. Branching bandit processes. *Probability in the Engineering and Informational Sciences* **2**(3) 269–278.

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

Whittle, P. 1982. *Optimization Over Time*. John Wiley & Sons, Inc.

Whittle, P. 1996. *Optimal Control: Basics and Beyond*. John Wiley & Sons, Inc.

Whittle, P. 2005. Tax problems in the undiscounted case. *Journal of Applied Probability* **42**(3) 754–765.

Zhong, Y., J. R. Birge, A. R. Ward. 2024. Learning to schedule in multiclass many-server queues with abandonment. *Operations Research* **Forthcoming**.

Online Appendix: Proofs

Proof of Lemma 1:

Proof. We construct an auxiliary MAB with K arms. Therefore, the state of the arm does not change under continuation control. When the arm is frozen, the state of the arm k does not change and no reward is collected. It is well known that the optimal policy is an index policy given by the Gittins index (Gittins et al. 2011).

In general formulation, Gittins index, $v(s)$, for a state s of an arm is defined as:

$$v(s) = \sup_{\tau} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha^t r_t(s) \mid s_0 = s\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \alpha^t \mid s_0 = s\right]}, \tag{14}$$

where τ is the stopping time (the decision epoch to stop exploring the current arm), s is the state of the arm at the current time, α is the discount factor ($0 < \alpha < 1$), $r_t(s)$ is the expected reward to receive if the arm is active in state s at time t, $\mathbb{E}[\cdot]$ denotes the expectation over the stochastic process governing rewards and transitions, $s_0 = s$ representing the initial state of the arm, and the supremum is taken over all possible stopping times τ. To verify that the index specified in the equation (13) is the Gittins index for this auxiliary MAB, we show that the expectation over the stochastic process conditioning on the initial states $\mathbb{E}\left[\cdot \mid s_0 = s\right]$ is equivalent to $\mathbb{E}\left[\cdot \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)\right]$, and $r_t(s)$ corresponds to $\beta\tilde{\mu}_k E[R_{k,t}^A]$.

In the Gittins index formulation specified in Equation (14), before an arm achieves the stopping time, it collects the reward assuming that the arm stays in the active mode. Thus, the stochastic process, $(Q_k^A(t), X_k^A(t) \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)$, under policy $\tilde{\pi}^A$, which always chooses active action in all states, is equivalent to the stochastic process, $(Q_k(t), X_k(t) \mid Q_k(0) = q_k, X_k(0) = \chi_k)$, for $t \leq \tau$, before the stopping time τ. Therefore, the expectation over the stochastic process conditioning on the initial states for the Gittins index: $\mathbb{E}\left[\cdot \mid s_0 = s\right]$ is $\mathbb{E}\left[\cdot \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)\right]$. By staying active in period t, the probability of seeing a service complete in the next period is $\tilde{\mu}_k$. Should completion of the service occur, the reward to collect conditioning on the belief in the period t is $R_k^A \mid X_k^A(t) = R_{k,t}^A$. Taking into account the discount factor, $r_t(s)$ for the Gittins index is $\beta\tilde{\mu}_k\mathbb{E}[R_{k,t}^A] := \beta\tilde{\mu}_k\mathbb{E}[R_k \mid X_k(t)]$. Thus, the Gittins index of arm k of the auxiliary MAB in state

(q_k, χ_k), is given by: For $q_k \geq 1$ and $\chi_k \in \Sigma_k$,

$$v_k(q_k, \chi_k) = \beta\tilde{\mu}_k \sup_{1\leq\tau} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \mathbb{E}[R_{k,t}^A] \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k\right]}. \tag{15}$$

Note that the reward sequences, $R_{k,t}^A$, for $t = 1, 2 \ldots$, are conditionally independent on the state $X_k(t)$. Using the law of iterated expectations, we can combine the expectation operations. Thus, the Gittins index for the MAB problem is:

$$v_k(q_k, \chi_k) = \beta\tilde{\mu}_k \sup_{1\leq\tau} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t R_{k,t}^A \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k\right]}, \tag{16}$$

and $v_k(0, \chi_k) = 0$ for $\chi_k \in \Sigma_k$.

Note that the arm k for the auxiliary MAB freezes when the queue is empty, that is, when $q_k = 0$, then $(Q_k(t+1), X_k(t+1)) = (0, \chi_k)$. Therefore, the optimal stopping time is constrained by the time when the queue is empty by construction, as continuation would only lower the index. Thus, the optimal stopping time τ satisfies $\tau \leq B_k(q_k)$, where

$$B_k(q_k) = \inf\left\{t \geq 0 : Q_k^A(t) = 0 \mid Q_k^A(0) = q_k\right\}. \tag{17}$$

Given that the optimal stopping time is less than $B(q_k)$, we can also write the Gittins index as:

$$\beta\tilde{\mu}_k \sup_{1\leq\tau\leq B_k(q_k)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t R_{k,t}^A \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k\right]}, \tag{18}$$

which is the Whittle index we showed in Theorem 1 □

Proof of Corollary 1:

Proof. The proof follows directly from Lemma 2.2 in Gittins et al. (2011). □

Proof of Lemma 2:

Proof. The proof follows that for each fixed sample path, $B_k(q_k)$ increases in q_k. Thus, the function $v_k(q_k, \chi_k)$ increases in q_k. □

Proof of Theorem 1:

Proof. We show that the arm k is indexable and $v_k(q_k, \chi_k)$ defined in Equation (13) is the Whittle index of the restless arm k defined in Definition 2 by showing that the following policy is optimal for the restless arm k defined in Equation (7):

$$\tilde{\pi}(q_k, \chi_k, w_k) = \begin{cases} A \text{ (Active)}, & \text{if } v_k(q_k, \chi_k) > w_k \\ P \text{ (Passive)}, & \text{if } v_k(q_k, \chi_k) < w_k \\ \text{Indifferent between } A \text{ and } P, & \text{if } v_k(q_k, \chi_k) = w_k. \end{cases}$$

We show that the optimality of the policy $\tilde{\pi}$ for the restless arm k implies indexability and identifies the Whittle index to be $v_k(q_k, \chi_k)$, respectively:

i.) *If policy $\tilde{\pi}$ is optimal for the restless arm k defined in Equation (7), then the arm k is indexable:* By Equation (10) and Definition 2, the stopping sets are the set of states where the passive action is optimal:

$$S_k(w_k) = \left\{(q_k, \chi_k) \in \mathbb{Z}_+ \times \Sigma_k : J_k^P(q_k, \chi_k, w_k) \geq J_k^A(q_k, \chi_k, w_k)\right\} \tag{19}$$

If the policy $\tilde{\pi}$, is optimal, then the stopping set of the restless arm k is defined by the by the states where the index $v_k(q_k, a_k)$ less that passive reward w_k, that is

$$S_k(w_k) \equiv \{(q_k, \chi_k) \in \mathbb{Z}_+ \times \Sigma_k : (q_k, \chi_k) \in \mathbb{Z}_+ \times \Sigma_k : v_k(q_k, \chi_k) \leq w_k\} \tag{20}$$

To prove the indexability of the restless arm k, we show that the stopping set is increasing in the passive reward w_k. Let $w_k' \geq w_k$ be another passive reward, we have

$$\{(q_k, \chi_k) \in \mathbb{Z}_+ \times \Sigma_k : v_k(q_k, \chi_k) \leq w_k\} \subseteq \left\{(q_k, \chi_k) \in \mathbb{Z}_+ \times \Sigma_k : v_k(q_k, \chi_k) \leq w_k'\right\},$$

which satisfies the monotonicity needed for the indexability stated in definition (1).

ii.) *If policy $\tilde{\pi}$ is optimal for the restless arm k defined in Equation (7), then $v_k(q_k, \chi_k)$ is the Whittle index for arm k:*

For the policy $\tilde{\pi}$ to be optimal for a single arm problem defined by Equation (7), the following statements hold for the index $v_k(q_k, \chi_k)$:

(a) If $v_k(q_k, \chi_k) > w_k$, $J_k^A(q_k, \chi_k, w_k) > J_k^P(q_k, \chi_k, w_k)$.

(b) If $v_k(q_k, \chi_k) < w_k$, $J_k^A(q_k, \chi_k, w_k) < J_k^P(q_k, \chi_k, w_k)$.

(c) If $v_k(q_k, \chi_k) = w_k$, $J_k^A(q_k, \chi_k, w_k) = J_k^P(q_k, \chi_k, w_k)$.

By definition (2), the Whittle index for a state is the minimum passive reward for the restless arm k to take the passive reward at that state. If the above three statements hold, we can establish that $v_k(q_k, \chi_k)$ is such a minimum reward satisfying Definition 2. To see this, note that with a passive reward of $w_k = v_k(q_k, \chi_k)$, the arm k is indifferent between active and passive action, with any passive reward lower than $v_k(q_k, \chi_k)$, the arm k should be active, and with any reward higher than $v_k(q_k, \chi_k)$, the arm k should remain passive.

Now, it suffices to show that the policy $\tilde{\pi}$ is optimal for the restless arm k defined in Equation (7).

Because for any $w_k < 0$, $J_k^A(q_k, \chi_k, w_k) > J_k^P(q_k, \chi_k, w_k)$, it suffices to consider the cases where $w_k \geq 0$. Moreover, we can see that for any $w_k \geq 0$, $J_k^A(0, \chi_k, w_k) \leq J_k^P(0, \chi_k, w_k)$ and $v_k(0, \chi_k) = 0 \leq w_k$. Thus, for any $w_k \geq 0$, it suffices to show that the policy $\tilde{\pi}$ is optimal in cases of $q_k > 0$.

For the case of $q_k > 0$, we show that the policy $\tilde{\pi}$ is optimal for the restless arm k using an interchange argument with logic similar to the proof of Theorem 2.1 in Gittins et al. (2011). To facilitate the argument, we assume that the realization of the events, i.e., the realized rewards and the arrival and service completion events, are associated with the number of periods that the system is in either mode in any fixed sample path. This assumption leads to two specific implications:

1. For a fixed sample path, any two policies see the same event on the same order of active/passive periods. In explain, if one policy sees a service completion upon the n^{th} active period, then the event on n^{th} active period of the other policy will also be a service completion.

2. For a fixed sample path, the server stays in the active mode for t_1 periods and in the passive mode for t_2 periods before hitting $q_k = 0$, then the system state is the same at the beginning of period $t_1 + t_2 + 1$, independent of the sequence of customer service. For any two policies, only the first $t_1 + t_2$ periods differ. Suppose that the queue lengths are positive in the first $t_1 + t_2$ periods under both policies. If the system stays in active mode for t_1 periods and in passive mode for t_2 periods under both policies, then the two policies produce the same sample paths after the first $t_1 + t_2$ periods.

The structure of our proof combines an interchange argument with induction, outlined as follows. Let Π_i denote the set of policies that deviates from the proposed policy $\tilde{\pi}$ for **at most** i periods where $i \geq 1$. We first show that if policies are restricted to Π_1, then there is no state of the restless bandit in which the single option of deviating from policy $\tilde{\pi}$ should be taken. Note that $\tilde{\pi} \subseteq \Pi_1 \subseteq \Pi_2 \ldots$, by an inductive argument, we may conclude that the policy $\tilde{\pi}$ is optimal within the class Π_i for all $i \geq 1$ (i.e., from the above it follows that the last option to deviate need not be used, and so we may restrict our attention to Π_{i-1}, and inductively to $\Pi_{i-2}, \ldots, \Pi_0$, where $\Pi_0 = \{\tilde{\pi}\}$.)

We use an interchange argument to show that if the policies are restricted to the set Π_1, then it is suboptimal to deviate from policy π^*. Consider a policy $\pi \in \Pi_1$ that deviates exactly once from the policy $\tilde{\pi}$. Without loss of generality, we can assume that π deviates in the first period. We show that by constructing a policy π' that interchanges the order of active and passive periods in policy π, the total reward collected under π' is strictly greater than that under π. Ultimately, under such construction, π' corresponds to the policy $\tilde{\pi}$ for $\pi \in \Pi_1$, therefore, it is strictly suboptimal to deviate from $\tilde{\pi}$.

For policy π, let (q_k, χ_k) denote the initial state of the system and $(Q_k^{\pi}(t), X_k^{\pi}(t))$ denote the state of the system at the end of the period t under policy π. By the index value $v_k(q_k, \chi_k)$, of different states, we divide our proof into three cases:

1. **Case:** $v_k(q_k, \chi_k) < w_k$. In this case, policy $\tilde{\pi}$ chooses the passive mode in period 0, and stays in the active mode until the index of the state is lower than w_k. Under policy π, the arm k stays in the active mode in period 0 and follows the policy $\tilde{\pi}$. Let the stopping time $\sigma \geq 2$ denote the period when the system switches to the passive mode, i.e., $\sigma = \inf\{\sigma : v_k(Q^{\pi}(\sigma), X^{\pi}(\sigma)) \leq w_k\}$. Note that $\sigma \leq B_k(q_k)$ because $v_k(0, \chi_k) = 0 \leq w_k$ for all $\chi_k \in \Sigma_k$. By interchanging the passive and the active time period, we construct π' which stays in passive mode in period 0 and then switches to the active mode in periods $2, 3, \ldots, \sigma + 1$. Then, policy π' follows policy $\tilde{\pi}$ afterwards.

   Note that $Q_k^{\pi'}(1) \geq q_k$ because under period π', period 0 can only see an arrival event or remain unchanged as the restless arm k stayed in the passive mode. Thus, the queue is nonempty in periods $2, \ldots, \sigma + 1$ under policy π' under the interchange assumption. Furthermore, the

system dynamics are identical after period $\sigma+1$ under the two policies by the interchange assumption stated at the beginning of the proof. Therefore, the rewards under the two policies, π and π', only differ in the first $\sigma+1$ periods.

Let $R^{\pi}_{k,t}$ and $R^{\pi'}_{k,t}$ denote the conditional reward based on the manager's belief regarding the reward of class k customers up to time t under policies π and π', respectively. We formally define these conditional rewards as $R^{\pi}_{k,t} := R_k \mid X^{\pi}_k(t)$, and $R^{\pi'}_{k,t} := R_k \mid X^{\pi'}_k(t)$, then the following holds:

$$\begin{aligned} v_k(q_k,\chi_k) &\geq \beta\tilde{\mu}_k \frac{\mathbb{E}\left[\sum_{t=0}^{\sigma-1}\beta^t R^{\pi}_{k,t} \mid Q^{\pi}_k(0)=q_k, X^{\pi}_k(0)=\chi_k\right]}{\mathbb{E}\left[\sum_{t=1}^{\sigma-1}\beta^t \mid Q^{\pi}_k(0)=q_k, X^{\pi}_k(0)=\chi_k\right]} \\ &= \beta\tilde{\mu}_k \frac{\mathbb{E}\left[\sum_{t=1}^{\sigma}\beta^{t-1} R^{\pi'}_{k,t} \mid Q^{\pi'}_k(0)=q_k, X^{\pi'}_k(0)=\chi_k\right]}{\mathbb{E}\left[\sum_{t=1}^{\sigma}\beta^{t-1} \mid Q^{\pi'}_k(0)=q_k, X^{\pi'}_k(0)=\chi_k\right]}. \end{aligned}$$

The inequality follows from Equation (13) and the equality follows from the interchange assumption. Note that the sum of the discounted time $\mathbb{E}\left[\sum_{t=1}^{\sigma}\beta^{t-1}\right]$ can be expressed as $\frac{\mathbb{E}[1-\beta^{\sigma}]}{1-\beta}$ by the geometric sum formula. Therefore, we have:

$$\begin{aligned} v_k(q_k,\chi_k) &\geq \beta\tilde{\mu}_k \frac{(1-\beta)\mathbb{E}\left[\sum_{t=0}^{\sigma-1}\beta^t R^{\pi}_{k,t} \mid Q^{\pi}_k(0)=q_k, X^{\pi}_k(0)=\chi_k\right]}{\mathbb{E}\left[1-\beta^{\sigma} \mid Q^{\pi}_k(0)=q_k, X^{\pi}_k(0)=\chi_k\right]} \\ &= \beta\tilde{\mu}_k \frac{\mathbb{E}\left[(1-\beta)\sum_{t=1}^{\sigma}\beta^{t-1} R^{\pi'}_{k,t} \mid Q^{\pi'}_k(0)=q_k, X^{\pi'}_k(0)=\chi_k\right]}{\mathbb{E}\left[1-\beta^{\sigma} \mid Q^{\pi'}_k(0)=q_k, X^{\pi'}_k(0)=\chi_k\right]}. \end{aligned}$$

Therefore, it follows from the assumption $w_k > v_k(q_k,\chi_k)$ that the following holds:

$$w_k > \beta\tilde{\mu}_k \frac{(1-\beta)\mathbb{E}\left[\sum_{t=0}^{\sigma-1}\beta^t R^{\pi}_{k,t} \mid Q^{\pi}_k(0)=q_k, X^{\pi}_k(0)=\chi_k\right]}{\mathbb{E}\left[1-\beta^{\sigma} \mid Q^{\pi}_k(0)=q_k, X^{\pi}_k(0)=\chi_k\right]}.$$

Simplify the inequality, we have:

$$\begin{aligned} &w_k + \beta\tilde{\mu}_k\mathbb{E}\left[\sum_{t=1}^{\sigma}\beta^t R^{\pi'}_{k,t} \mid Q^{\pi'}_k(0)=q_k, X^{\pi'}_k(0)=\chi_k\right] \\ &> \beta\tilde{\mu}_k\mathbb{E}\left[\sum_{t=0}^{\sigma-1}\beta^t R^{\pi}_{k,t} \mid Q^{\pi}_k(0)=q_k, X^{\pi}_k(0)=\chi_k\right] + \mathbb{E}[\beta^{\sigma} \mid Q^{\pi}_k(0)=q_k, X^{\pi}_k(0)=\chi_k]w_k. \end{aligned}$$

Note that the left-hand side of the inequality is the expected discounted reward in the first $\sigma+1$ periods under policy π', while the right-hand side is the expected discounted reward in the first $\sigma+1$ periods under policy π. Note that π' is also in Π_1 because $v_k(Q^{\pi'}_k(t), X^{\pi'}_k(t)) \geq v_k(Q^{\pi}_k(t-1), X^{\pi}_k(t-1)) \geq w_k$ for $t=1,\ldots,\sigma$ by the definition of stopping time σ and the

interchange assumption. Thus, in this case, if π deviates from policy $\tilde{\pi}$ in period 1, there exists a policy $\pi' \in \Pi_1$ that strictly improves policy π. Furthermore, note that policy π' in effect is the policy $\tilde{\pi}$ as it stays passive in the first period where $v_k(q_k, \chi_k) < w_k$ and stay active until σ, where $v_k(Q^\pi(\sigma), X^\pi(\sigma)) \leq w_k$.

2. **Case** $v_k(q_k, \chi_k) > w_k$. In this case, under policy $\tilde{\pi}$, the restless arm k stays in active mode until τ, where $v_k(Q^{\tilde{\pi}}(\tau), X^{\tilde{\pi}}(\tau)) \leq w_k$. Deviating from policy $\tilde{\pi}$ in the first period, under policy π, the restless arm k stays in the passive mode in period 1 and stops until σ, where $v_k(Q^\pi(\sigma), X^\pi(\sigma)) \leq w_k$.

   Under policy π, we have that $Q_k^\pi(1) \geq q_k$ and $\tilde{X}_k^\pi(1) = \chi_k$ since policy π stayed passive in the first period. Thus, it follows from (ii) of Corollary 1 that $v_k(\tilde{Q}_k^\pi(1), \tilde{X}_k^\pi(1)) > w_k$. Thus, the system switches to the active mode in period 2. The system stays in the active mode for σ periods until the index of the state falls below w_k. Let $\delta = \tilde{Q}_k^\pi(1) - q_k$ denote the number of new arrivals in the first period. In addition, let σ denote the stopping time that stops when the index of $v_k(Q_k^\pi(t) - \delta, X_k^\pi(t))$ falls below $v_k(q_k, \chi_k)$. Thus, it follows from (ii) of Corollary 1 and Lemma 2 that $\tau \leq \sigma$ almost surely. Therefore, policy π can stay active for at least τ periods.

   We construct a policy π' that stays active in periods $0, \ldots, \tau - 1$ and stays in the passive mode in period τ. Note that the system state under the two policies is the same after period τ. Thus, policy π' acts the same with policy π after period τ. Note that by (ii) of Corollary 1, the stopping time $\tau - 1$ attains $v_k(q_k, \chi_k)$.

   Similar to the previous case, let $R_{k,t}^\pi$ and $R_{k,t}^{\pi'}$ denote the conditional reward based on the manager's belief regarding the reward of class k customers up to time t under policies π and π', respectively. The following holds by the case assumption, $v_k(q_k, \chi_k) > w_k$, attainability of $\tau - 1$ on $v_k(q_k, \chi_k)$, and the transformation of the geometric sum similar in the previous case:

$$v_k(q_k, \chi_k) = \beta\tilde{\mu}_k \frac{(1-\beta)\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t R_{k,t}^\pi \mid Q_k^\pi(0) = q_k, X_k^\pi(0) = \chi_k\right]}{\mathbb{E}\left[1 - \beta^\tau \mid Q_k^\pi(0) = q_k, X_k^\pi(0) = \chi_k\right]} > w_k.$$

   Simplify the expression, we have:

$$\beta\tilde{\mu}_k\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t R_{k,t}^{\pi'} \mid Q_k^{\pi'}(0)=q_k, X_k^{\pi'}(0)=\chi_k\right]+\mathbb{E}\left[\beta^\tau \mid Q_k^{\pi'}(0)=q_k, X_k^{\pi'}(0)=\chi_k\right]w_k$$
$$> w_k+\beta\tilde{\mu}_k\mathbb{E}\left[\sum_{t=1}^{\tau}\beta^t R_{k,t}^{\pi} \mid Q_k^{\pi}(0)=q_k, X_k^{\pi}(0)=\chi_k\right].$$

Note that the left-hand side is the expected discounted payoff of the first τ periods under policy π', whereas the right-hand side is that under policy π. Note that π' is also in set Π_1. Note that π' is also in Π_1 because π' doesn't not deviate from policy $\tilde{\pi}$. Thus, in this case, if π deviates from policy $\tilde{\pi}$ in period 1, there exists a policy $\pi' \in \Pi_1$ that strictly improves policy π. Furthermore, note that policy π' in effect is the policy $\tilde{\pi}$ as it stays active until τ, where $v_k(Q^\pi(\tau), X^\pi(\tau)) \leq w_k$.

3. **Case** $v_k(q_k, \chi_k) = w_k$. In this case, it is indifferent to stay in the active or passive modes. First consider the case when the system is in the active mode. The server switches to the passive mode only when its index is less than or equal to w_k. Let τ be the time when the server switches to class 1. It follows from (ii) of Corollary 1 that τ attains $v_k(q_k, \chi_k)$. Thus, the total expected discounted reward in first τ periods is $w_k \sum_{t=0}^{\tau}\beta_t$ by the definition of $v_k(q_k, \chi_k)$. Now consider a policy that the system stays in the passive mode in period 1 and switches to the active mode in periods $2, \ldots, \tau$. It is easy to see that the alternative policy also yields a total expected discounted reward of $w_k \sum_{t=0}^{\tau}\beta_t$ in first τ period. The sample paths of the two policies are the same after period τ. Thus, there is no difference with the policy of staying in the active mode in period 1 and the policy that staying in the passive mode in period 1 and then switches to the active mode. An induction argument can take care of the case when the system stays in the passive mode for multiple periods before switching to the active mode.

As a conclusion, we have shown that if the policies are restricted to the set Π_1, then it is suboptimal to deviate from the policy $\tilde{\pi}$. Using an inductive argument, we can show that if policies are restricted to the set Π_k, it is optimal to consider only policies in the set Π_{k-1}. Thus, we conclude that policy $\tilde{\pi}$ is optimal among the set of policies Π_i for all integers $i \geq 1$.

Establishing that $\tilde{\pi}$, the policy determined by $v_k(q_k, \chi_k)$, is optimal for all $q_k, \chi_k \in \mathbb{Z}_+ \times \Sigma_k$ completes our proof. This conclusion is based on our explanation earlier in the proof, where we

show that the optimality of $\tilde{\pi}$ directly implies that the restless arm k is indexable and the function $v_k(q_k, \chi_k)$ serves as the Whittle index for arm k. □

Proof of Lemma 3:

Proof. It can be viewed as a special case of Theorem 1 when the rewards are i.i.d. and has a mean of r_1. If the rewards are i.i.d., for any stopping time that satisfies $1 \leq \tau \leq B_1(q)$ almost surely, the following holds:

$$\frac{\beta\tilde{\mu}_1\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t R_1\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} = \frac{\beta\tilde{\mu}_1 r_1\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} = \beta\tilde{\mu}_1 r_1.$$

Class 1 can also be viewed as a special case of Ansell et al. (2003). The result is consistent with the analysis there. □

Proof of Theorem 2:

Proof. The proof is very similar to that of Theorem 1.

If $q_1 = 0$, the optimal action is to serve the class 2 customer because we assume a preemptive discipline. Similarly, if $q_2 = 0$, the optimal action is to serve a class 1 customer. Therefore, it suffices to consider the cases $q_1 > 0$ and $q_2 > 0$, for which the index $v_1(q_1) = \beta\tilde{\mu}_1 r_1$.

In the cases $q_1 > 0$ and $q_2 > 0$, intuitively, the scheduling problem with one known-reward class and one unknown-reward class is equivalent to the single restless arm problem defined in Equation (7) with passive reward $w_k = \beta\tilde{\mu}_1 r_1 = v_1(q_1)$ for $q_1 > 0$. To see the similarities, in each period, the manager chooses an active action by serving a class 2 customer and earns a random reward and update belief χ, or chooses a passive action by serving class 1 customer and collect a known reward of $\beta\tilde{\mu}_1 r_1$. As we have shown in Theorem 1, the optimal policy is given by the relationship between $v_1(q_1)$ and $v_2(q_2, \chi)$.

We call the proposed policy stated in the theorem the non-idling index policy. To facilitate the proof, let Π_i denote the set of policies that deviates from the proposed index policy in i periods at most for $k \geq 1$. Thus, the policy of non-idling index is set in Π_i for all $i \geq 1$. We follow a similar induction argument to show the optimality of the proposed index policy within Π_i. Now, let us consider a policy $\pi \in \Pi_1$ that deviates from the index policy once. Without loss of generality, we can assume that π deviates in the first period. Let (q_1, q_2, χ) denote the initial state of the system.

We consider two cases: $v_2(q_2, \chi) < v_1(q_1)$ and $v_2(q_2, \chi) > v_1(q_1)$. Using similar logic in the proof of Theorem 1, we show that the Whittle scheduling policy is optimal in 3 cases:

1. Case 1: $v_2(q_2, \chi) < v_1(q_1)$. In this case, under policy π, the server serves a class 2 customer or stays idle in period 1. We first discuss the case where $q_2 \geq 1$ and the server starts serving a class 2 customer. This case is the same as the case where $v_k(q_k, \chi_k) < w_k$ in the proof of Theorem 1, so we omit the analysis here.

   Next, we consider the scenario in which the policy π sets the server idle in period 1. Then, according to policy π, the server serves class 2 in periods $2, \ldots, \sigma$ to $v_2(Q_2^\pi(\sigma+1), X^\pi(\sigma+1)) < \mu_1 r_1$ for the first time and then starts serving class 1. The stopping time $\sigma = 1$ if $v_2(Q_2^\pi(2), X^\pi(2)) < \mu_1 r_1$. In this case, the server starts serving class 1 immediately after idling in period 1. We consider a policy π' that serves class 1 in period 1, follows policy π in periods $2, \ldots, \sigma$, and then idle the server in period $\sigma + 1$. The states of the system are the same under both policies in period $\sigma + 2$ and afterward. Thus, we can let policy π' follow policy π after period $\sigma + 1$. The policy π' yields a strictly higher pay-off than the policy π.

2. Case 2: $v_2(q_2, \chi) > v_1(q_1)$. The server either servers a customer of class 1 or stays idle in period 1 under policy π. The argument of staying idle in period 1 is similar to the previous case discussed. Thus, we omit it and only consider the case when π serves class 1 in period 1. In this case, $q_1 \geq 1$ and $v_1(q_1) = \mu_1 r_1$. We can follow a similar argument to that of Case 2 of the proof of Theorem 1.

3. Case 3: $v_2(q_2, \chi) = v_1(q_1)$. We can follow a similar argument to that of Case 3 of the proof of Theorem 1 to see that either policy yields a reward of $\beta \tilde{\mu}_1 r_1 \sum_{t=0}^{\tau} \beta_t$ within τ.

□

Proof of Lemma 4:

Proof. Denote by $\overline{\chi}$ and $\underline{\chi}$, respectively, the Bayesian posteriors corresponding to a prior χ depending on whether the system manager receives a high or low reward upon service completion. They

satisfy

$$\overline{\chi} = g(\chi, r_h) = \frac{\chi p_b}{\chi p_b + (1-\chi)p_m} \text{ and}$$
$$\underline{\chi} = g(\chi, r_l) = \frac{\chi(1-p_b)}{\chi(1-p_b) + (1-\chi)(1-p_m)}$$

First, notice that both posterior believes $\overline{\chi}$ and $\underline{\chi}$ increase in initial belief, χ. In addition, the expected reward $\bar{r}_2(\chi)$ also increases in χ. Since the posterior believes and the expected reward both increase strictly in χ, their compositions, the posterior expected rewards, $\overline{r}(\overline{\chi})$ and $\overline{r}(\underline{\chi})$ also increase in χ. Thus, for any stopping time $1 \leq \sigma \leq B_2(q_2)$ a.s., the following holds: For $0 \leq \chi_2' \leq \chi_2 \leq 1$,

$$\frac{\mathbb{E}\left[\beta\tilde{\mu}_2 \sum_{t=0}^{\sigma-1} \beta^t R_{2,t} \mid \tilde{X}(0) = \chi_2'\right]}{\mathbb{E}\left[\sum_{t=0}^{\sigma-1} \beta^t\right]} \leq \frac{\beta\tilde{\mu}_2 \mathbb{E}\left[\sum_{t=1}^{\sigma-1} \beta^{t-1} R_{2,t} \mid \tilde{X}(0) = \chi_2\right]}{\mathbb{E}\left[\sum_{t=0}^{\sigma-1} \beta^t\right]} \leq v_2(q_2, \chi_2).$$

Taking the supremum on all possible stopping times, we obtain $v_2(q_2, \chi_2') \leq v_2(q_2, \chi_2)$.

Note that $\sigma = 2$ is a stopping time if $q_2 \geq 1$ because $B_2(q_2) \geq 2$ for all $q_2 \geq 1$. Thus, the following holds: For $q_2 \geq 1$ and $\chi_2 \in [0,1]$,

$$v_2(q_2, \chi_2) \geq \beta\tilde{\mu}_2 \mathbb{E}\left[\bar{r}_2(\tilde{X}(0))\right] = \beta\tilde{\mu}_2 \bar{r}_2(\chi_2).$$

The inequality $v_2(q, \chi) \leq \bar{v}_2(\chi_2)$ holds as any stopping satisfying $1 \leq \sigma \leq B_2(q)_2$ satisfies $1 \leq \sigma$. □

Proof of Lemma 5:

Proof. Whittle index is specified by the following Equation:

$$v_k(q_k, \chi_k) = \beta\tilde{\mu}_k \sup_{1 \leq \tau \leq_{a.s.} B_k(q_k)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t R_{k,t}^A \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t\right]}, \tag{21}$$

where $R_{k,t}^A$ is the expected reward of the class k conditioning on the belief state $X_k^A(t)$. In the setting of our specific reward case, where the parameter takes value p_b, p_m, r_h and r_l, we have

$$R_{k,t}^A = r_h\left(X_k^A(t)p_b + \left(1 - X_k^A(t)\right)p_m\right) + r_l\left(X_k^A(t)\left(1-p_b\right) + \left(1 - X_k^A(t)\right)\left(1-p_m\right)\right). \tag{22}$$

Let $P(X_k^A(t))$ denote the value $X_k^A(t)p_b + \left(1 - X_k^A(t)\right)p_m$, we have

$$R_{k,t}^A = r_h P(X_k^A(t)) + r_l(1 - P(X_k^A(t))). \tag{23}$$

Plugging it into the equation (21), we have

$$v_k(q_k, \chi_k) = \beta\tilde{\mu}_k \sup_{1\le\tau\le_{a.s.}B_k(q_k)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\left(P(X_k^A(t)) + (1-P(X_k^A(t)))\right) \mid Q_k^A(0)=q_k, X_k^A(0)=\chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]}. \tag{24}$$

Suppose that we transform the rewards for both classes with the linear operation $f(x) = ax + b$, such that $r_h' = ar_h + b$ and $r_l' = ar_l + b$. Let $v_k(q_k, \chi_k \mid r_h', r_l')$ denote the Whittle index after the linear transformation, we have

$$\begin{aligned}
&v_k(q_k, \chi_k \mid r_h', r_l') \\
&= \beta\tilde{\mu}_k \sup_{1\le\tau\le_{a.s.}B_k(q_k)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\left((ar_h+b)P(X_k^A(t)) + (ar_l+b)(1-P(X_k^A(t)))\right) \mid Q_k^A(0)=q_k, X_k^A(0)=\chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} \\
&= \beta\tilde{\mu}_k \sup_{1\le\tau\le_{a.s.}B_k(q_k)} \frac{a\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\left(r_hP(X_k^A(t)) + r_l(1-P(X_k^A(t))) + \frac{b}{a}\right) \mid Q_k^A(0)=q_k, X_k^A(0)=\chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]}. \\
&= \beta\tilde{\mu}_k \sup_{1\le\tau\le_{a.s.}B_k(q_k)} \frac{a\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\left(r_hP(X_k^A(t)) + r_l(1-P(X_k^A(t)))\right) \mid Q_k^A(0)=q_k, X_k^A(0)=\chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} + a\frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\frac{b}{a}\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]}. \\
&= a\beta\tilde{\mu}_k \sup_{1\le\tau\le_{a.s.}B_k(q_k)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\left(r_hP(X_k^A(t)) + r_l(1-P(X_k^A(t)))\right) \mid Q_k^A(0)=q_k, X_k^A(0)=\chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} + b,
\end{aligned} \tag{25}$$

which is equivalent to applying the linear transformation, $f(x) = ax + b$ on the original Whittle index with reward r_h and r_l. □

Proof of Lemma 6:

Proof. Suppose that we have a system with two classes of customer, class 1 and class 2, whose rewards are unknown. We assume that the ground truth is that class 1 generates a higher expected reward, that is, the portion of the high-reward customers of class 1 is p_h and class 2 generates a lower expected reward, that is, the portion of the high-reward customer of class 2 is p_l. Given the ground truth, the expected reward for each class 1 customer is a high reward, denoted by $H = p_h(r_h - r_l) + r_l$ and for each class 2 customer is a low reward denote by $L = p_l(r_h - r_l) + r_l$ for the class 2 customer. Note that these are ground truth information that the system manager does not know.

In the the ground truth, we fix the product of $r_1\mu_1$ and $r_2\mu_2$. Additionally, we fix $r_2 = L$ and μ_2, exactly the same, which means that the customer characteristics of the class 2 customer remain unchanged. It is important to note that the system manager knows the values of H, L, μ_1, and

μ_2, as well as any changes to these values. The only information for the manager to learn is which class generates the higher expected reward, H, and which generates the lower expected reward, L. Now, if we change the composition $r_1\mu_1$ in the truth of the ground by increasing μ_1 and decreasing r_1 at the same time so that the $r_1\mu_1$ product is the same, we want to show that the Whittle index for class 1 will increase and thus more likely to recommend serving class 1 customer.

To show this, we need to go back to the definition of the Whittle index. In general, the Whittle index takes the form of following:

$$v_k(q_k, \chi_k) = \beta\tilde{\mu}_k \sup_{1\leq\tau\leq B_k(q_k)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t R_k^A \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t\right]}, \tag{26}$$

In the single server queue scheduling problem with Bernoulli-Bernoulli prior-posterior structure, the state of the system for each class is characterized by (q_k, χ_k)—the length of the queue for that class and the probability that the manager believes that the reward for that class belongs to a optimistic scenario. The Whittle index takes the following form:

$$v_k(q_k, \chi_k) = \beta\tilde{\mu}_k \sup_{1\leq\tau\leq B_k(q_k)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \Big(\chi_k H + (1-\chi_k)L\Big) \mid Q_k^A(0) = q_k, X_k^A(0) = \chi_k)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t\right]}. \tag{27}$$

Suppose that we increase μ_1 by a, for some $a > 1$, to keep the $r_1\mu_1 = H\mu_1$ product the same, we need to decrease H by $\frac{1}{a}$. Plugging this into the Whittle index, we have

$$v_1'(q_1, \chi_1) = \beta a\tilde{\mu}_1 \sup_{1\leq\tau\leq B_1(q_1)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \Big(\chi_1 \frac{1}{a} H + (1-\chi_1)L\Big) \mid Q_1^A(0) = q_1, X_k^A(0) = \chi_1)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t\right]}. \tag{28}$$

Since we fix ρ_1, the busy period (for the uniformized queueing system) has the same distribution before and after changing the composition of $r_1\mu_1$. This allows us to intuitively show the difference before after the composition change:

$$
\begin{aligned}
&v_1'(q_1,\chi_1) - v_1(q_1,\chi_1) = \\
&\beta a\tilde{\mu}_1 \sup_{1\le\tau\le B_1(q_1)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\Big(\chi_1\frac{1}{a}H + (1-\chi_1)L)\Big) \mid Q_1^A(0)=q_1, X_1^A(0)=\chi_1)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} \\
&-\beta\tilde{\mu}_1 \sup_{1\le\tau\le B_1(q_1)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\Big(\chi_1 H + (1-\chi_1)L)\Big) \mid Q_1^A(0)=q_1, X_1^A(0)=\chi_1)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} \\
&=\beta\tilde{\mu}_1 \sup_{1\le\tau\le B_1(q_1)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\Big(\chi_1 H + a(1-\chi_1)L)\Big) \mid Q_1^A(0)=q_1, X_1^A(0)=\chi_1)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} \\
&-\beta\tilde{\mu}_1 \sup_{1\le\tau\le B_1(q_1)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\Big(\chi_1 H + (1-\chi_1)L)\Big) \mid Q_1^A(0)=q_1, X_1^A(0)=\chi_1)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} \\
&=\beta\tilde{\mu}_1 \sup_{1\le\tau\le B_1(q_1)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\Big((a-1)(1-\chi_1)L)\Big) \mid Q_1^A(0)=q_1, X_1^A(0)=\chi_1)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]} \\
&>0.
\end{aligned}
\tag{29}
$$

The inutition is that the μ_1 term on the outside increased by a factor of a, the term inside decreased less because it is weighted by the term χ_1, therefore, the overall index value increases strictly for any $\chi_1 > 0$. The increase of μ_1 affects the Whittle index linearly; however, the reward affects the Whittle index sub-linearly because it is weighted by the belief. Thinking along this line, we can show that this change only reduces the Whittle index for class 2 customers, making serving class 1 customers more favorable:

$$
\begin{aligned}
v_2(q_2,\chi_2) &= \beta\tilde{\mu}_2 \sup_{1\le\tau\le B_2(q_2)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\Big(\chi_2 H + (1-\chi_2)L)\Big) \mid Q_k^A(0)=q_2, X_k^A(0)=\chi_2)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]}, \\
v_2'(q_2,\chi_2) &= \beta\tilde{\mu}_2 \sup_{1\le\tau\le B_2(q_2)} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\Big(\chi_2\frac{1}{a}H + (1-\chi_2)L)\Big) \mid Q_2^A(0)=q_2, X_2^A(0)=\chi_2)\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t\right]},
\end{aligned}
\tag{30}
$$

where $v_2(q_2,\chi_2) > v_2'(q_2,\chi_2)$.

Therefore, since $v_1'(q_1,q_2) > v_1'(q_1,q_2)$ and $v_2'(q_2,\chi_2) < v_2(q_2,\chi_2)$, we can conclude that the decrease in H and the increase in μ_1 strictly increase the likelihood that the Whittle index policy recommends serving the class 1 customer. The intuition is that the service rate is a common

knowledge whose increase will always be given full weight. However, the changes in p_h will be weighted by belief. □