Whittle Index-Based Q-Learning for Wireless Edge Caching With Linear Function Approximation

Guojun Xiong, *Student Member, IEEE*, Shufan Wang, Jian Li, *Member, IEEE*, and Rahul Singh, *Member, IEEE*

***Abstract*— We consider the problem of content caching at the wireless edge to serve a set of end users via unreliable wireless channels so as to minimize the average latency experienced by end users due to the constrained wireless edge cache capacity. We formulate this problem as a Markov decision process, or more specifically a restless multi-armed bandit problem, which is provably hard to solve. We begin by investigating a discounted counterpart, and prove that it admits an optimal policy of the threshold-type. We then show that this result also holds for average latency problem. Using this structural result, we establish the indexability of our problem, and employ the Whittle index policy to minimize average latency. Since system parameters such as content request rates and wireless channel conditions are often unknown and time-varying, we further develop a model-free reinforcement learning algorithm dubbed as `Q+-Whittle` that relies on Whittle index policy. However, `Q+-Whittle` requires to store the Q-function values for all state-action pairs, the number of which can be extremely large for wireless edge caching. To this end, we approximate the Q-function by a parameterized function class with a much smaller dimension, and further design a `Q+-Whittle` algorithm with linear function approximation, which is called `Q+-Whittle-LFA`. We provide a finite-time bound on the mean-square error of `Q+-Whittle-LFA`. Simulation results using real traces demonstrate that `Q+-Whittle-LFA` yields excellent empirical performance.**

***Index Terms*— Wireless edge caching, restless bandits, whittle index policy, reinforcement learning, finite-time analysis.**

I. Introduction

THE dramatic growth of wireless traffic due to an enormous increase in the number of mobile devices is posing many challenges to the current mobile network infrastructures. In addition to this increase in the volume of traffic, many emerging applications such as Augmented/Virtual Reality, autonomous vehicles and video streaming, are *latency-sensitive*. In view of this, the traditional approach of offloading the tasks to remote data centers is becoming less attractive. Furthermore, since these emerging applications typically require unprecedented computational power, it is not possible to run them on mobile devices, which are typically resource-constrained.

To provide such stringent timeliness guarantees, mobile edge computing architectures have been proposed as a means to improve the quality of experience (QoE) of end users, which move servers from the cloud to edges, often wirelessly that are closer to end users. Such edge servers are often empowered with a small wireless base station, e.g., the storage-assisted future mobile Internet architecture and cache-assisted 5G systems [1]. By using such edge servers, content providers are able to ensure that contents such as movies, videos, software, or services are provided with a high QoE (with minimal latency). The success of edge servers relies upon "content caching", for which popular contents are placed at the cache associated with the wireless edge. If the content requested by end users is available at the wireless edge, then it is promptly delivered to them. Unfortunately, the amount of contents that can be cached at the wireless edge is often limited by the wireless edge cache capacity. These issues are further exacerbated when the requested content is delivered over *unreliable* channels.

In this work, we are interested in minimizing *the average latency* incurred while delivering contents to end users, which are connected to a wireless edge via unreliable channels. We design dynamic policies that decide *which contents should be cached at the wireless edge so as to minimize the average latency of end users.*

A. Whittle Index Policy for Wireless Edge Caching

We pose this problem as a Markov decision process (MDP) [2] in Section III. Here, the system state is the number of outstanding requests from end users for each content that needs to be satisfied, and the cost is measured as the latency experienced by end users to obtain the requested contents. The available actions are the choices of caching each content or not given that the wireless edge cache capacity is much smaller than the total number of distinct requested contents. This MDP turns out to be an infinite-horizon average-cost restless multi-armed bandit (RMAB) problem [3]. Even though in theory this RMAB can be solved by using relative value iteration [2], this approach suffers from the curse of dimensionality, and fails to provide any insight into the solution. Thus, it is desirable to derive low-complexity solutions and provide guarantees

Manuscript received 22 February 2023; revised 14 September 2023, 16 January 2024, and 11 April 2024; accepted 9 June 2024; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor L. Huang. Date of publication 24 June 2024; date of current version 17 October 2024. The work of Guojun Xiong, Shufan Wang, and Jian Li was supported in part by the National Science Foundation (NSF) under Grant 2148309 and Grant 2315614, in part by U.S. Army Research Office (ARO) under Grant W911NF-23-1-0072, and in part by U.S. Department of Energy (DOE) under Grant DE-EE0009341. The work of Rahul Singh was supported in part by the Science and Engineering Research Board under Grant SRG/2021/002308. *(Corresponding author: Jian Li.)*

Guojun Xiong, Shufan Wang, and Jian Li are with the Data Science Program, Departments of Applied Mathematics and Statistics, and Computer Science, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: guojun.xiong@stonybrook.edu; shufan.wang@stonybrook.edu; jian.li.3@stonybrook.edu).

Rahul Singh is with Indian Institute of Science, Bengaluru, Karnataka 560012, India (e-mail: rahulsingh@iisc.ac.in).

Digital Object Identifier 10.1109/TNET.2024.3417351

on their performance. A celebrated policy for RMAB is *the Whittle index policy* [3]. We propose to use the Whittle index policy to solving the above problem for wireless edge caching.

Following the approach taken by Whittle [3], we begin by relaxing the hard constraint of the original MDP, which requires that the number of cached contents at each time is *exactly* equal to the cache size. These are relaxed to a constraint which requires that the number of cached contents is equal to the cache size *on average*. We then consider the Lagrangian of this relaxed problem, which yields us a set of decoupled average-cost MDPs, which we call the per-content MDP. Instead of optimizing the average cost (latency) of this per-content MDP, we firstly consider a discounted per-content MDP, and prove that the optimal policy for each discounted per-content MDP has an appealing *threshold structure*. This structural result is then shown to also hold for *the average latency problem*. We use this structural result to show that our problem is indexable [3], and then derive Whittle indices for each content. Whittle index policy then prioritizes contents in a decreasing order of their Whittle indices, and caches the maximum number of content constrained by the cache size. Whittle index policy is computationally tractable since its complexity increases linearly with the number of contents. Moreover it is known to be asymptotically optimal [4], [5] as the number of contents and the cache size are scaled up, while keeping their ratio as a constant. Our contribution in Section IV is non-trivial since establishing indexability of RMAB problems is typically intractable in many scenarios, especially when the probability transition kernel of the MDP is convoluted [6], and Whittle indices of many practical problems remain unknown except for a few special cases.

B. Whittle Index-Based Q-Learning With Linear Function Approximation for Wireless Edge Caching

The Whittle index policy needs to know the controlled transition probabilities of the underlying MDPs, which in our case amounts to knowing the statistics of content request process associated with end users, as well as the reliability of wireless channels. However, these parameters are often unknown and time-varying. Hence in Section V, we design an efficient reinforcement learning (RL) algorithm to make optimal content caching decisions dynamically without knowing these parameters. We do not directly apply off-the-shelf RL methods such as UCRL2 [7] and Thompson Sampling [8] since the size of state-space grows exponentially with the number of contents, and hence the computational complexity and the learning regret would also grow exponentially. Thus, the resulting algorithms would be too slow to be of any practical use.

To overcome these limitations, we first derive a model-free RL algorithm dubbed as `Q+-Whittle`, which is largely inspired by the recent work [9] that proposed a tabular Whittle index-based Q-learning algorithm, which we call `Q-Whittle` for ease of exposition. The key aspect of `Q-Whittle` [9] is that the updates of Q-function values and Whittle indices form a two-timescale stochastic approximation (2TSA) [10] with the former operating on a faster timescale and the latter on a slower timescale. Though [9] provided a rigorous asymptotic convergence analysis, such a 2TSA usually suffers from slow convergence in practice (as we numerically verify in Section VII). To address this limitation, our key insight is that we can further leverage the threshold-structure of the optimal policy to each per-content MDP to learn Q-function values of *only* those state-action pairs which are visited under the current threshold policy, rather than all state-action pairs as in `Q-Whittle`. This novel update rule enables `Q+-Whittle` to significantly improve the sample efficiency of `Q-Whittle` using the conventional ϵ-greedy policy.

We note that `Q+-Whittle` needs to store Q-function values for all state-action pairs, the number of which can be very large for wireless edge caching. To address this difficulty, we further study `Q+-Whittle` with linear function approximation (LFA) by using low-dimensional linear approximation of Q-function. We call this algorithm the `Q+-Whittle-LFA`, which can be viewed through the lens of a 2TSA. We provide a finite-time bound on the mean-square error of `Q+-Whittle-LFA` in Section VI. To the best of our knowledge, our work is the first to consider a model-free RL approach with LFA towards a Whittle index policy in the context of wireless edge caching over unreliable channels, and the first to provide a finite-time analysis of a Whittle index based Q-learning with LFA. We note that our model-free framework with LFA and its finite-time analysis under Markovian noise is of independent interest, and could be useful for other large-scale network problems.

Finally, we provide extensive numerical results using both synthetic and real traces to support our theoretical findings in Section VII, which demonstrate that `Q+-Whittle-LFA` produces significant performance gain over state of the arts.

II. Related Work

Although edge caching has received a significant amount of attentions, we are not aware of any prior work proposing an analytical model for latency-optimal wireless edge caching over unreliable channels, designing a computationally efficient index based policy and a novel RL augmented algorithm in face to unpredictable content requests and unreliable channels. We provide an account of existing works in two areas closely related to our work: content caching and restless bandits.

Content Caching [11] has been studied in numerous domains with different objectives such as minimizing expected delay [12], operational costs [13] or maximizing utility [14], [15]. The joint caching and request routing has also been investigated, e.g., [16] and [17]. Most prior works formulated the problem as a constrained/stochastic optimization problem, etc. None of those works provided a formulation using the RMAB framework and developed an index based caching policy. Furthermore, all above works assumed full knowledge of the content request processes and hence did not incorporate a learning component. A recent line of works considered caching from an online learning perspective, e.g., [18] and [19], and used the performance metric of learning regret or competitive ratio. Works such as [20], [21], [22], [23], and [24] used deep RL methods. However, deep RL methods lack of theoretical performance guarantees. Our model, objective and formulation significantly depart from those considered in aforementioned works, where we pose the wireless edge caching problem as a MDP and develop the Whittle index policy that can be easily learned through a model-free RL framework.

Restless Multi-Armed Bandit (RMAB) is a general framework for sequential decision making problems, e.g., [25] and [26]. However, RMAB is notoriously intractable [27]. One celebrated policy is the Whittle index policy [3]. However,

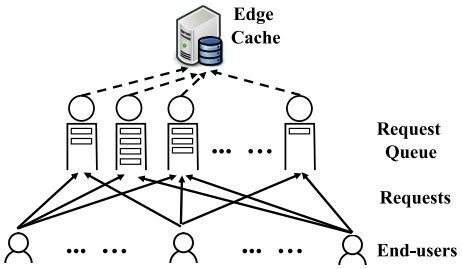

Fig. 1. Wireless edge caching over unreliable channels.

Whittle index is well-defined only when the indexability condition is satisfied, which is in general hard to verify. Additionally, the application of Whittle index requires full system knowledge. Thus it is important to examine RMAB from a learning perspective, e.g., [28], [29], [30], and [31]. However, these methods did not exploit the special structure available in RMAB and contended directly with an extremely high dimensional state-action space yielding the algorithms to be too slow to be useful. Recently, RL based algorithms have been developed [9], [32], [33], [34], [35], [36], [37], [38], [39], [40] to explore the problem structure through index policies. However, [9], [32], [34], and [35] lacked finite-time performance analysis and multi-timescale SA algorithms often suffer from slow convergence. References [33] and [36] depended on a simulator for explorations which cannot be directly applied here since it is difficult to build a perfect simulator in complex wireless edge environments. Reference [38] leveraged the threshold policy via a deep neural network without finite-time performance guarantees. References [37], [39], and [40] either studied a finite-horizon setting or developed model-based RL solutions, while we consider an infinite-horizon average-cost setting and develop model-free RL algorithms. Specifically, we propose `Q`$^+$`-Whittle-LFA`, a low-complexity Whittle index based Q-learning algorithm with linear function approximation. Our finite-time analysis of `Q`$^+$`-Whittle-LFA` further distinguishes our work.

III. System Model and Problem Formulation

In this section, we present the system model and formulate the average latency minimization problem for wireless edge caching over unreliable channels.

A. System Model

Consider a wireless edge system as shown in Figure 1, where the wireless edge is equipped with a cache of size B units to store contents that are provided to end users. We denote the set of distinct contents as $\mathcal{M} = \{1, \cdots, M\}$ with $|\mathcal{M}| = M$. We assume that all contents are of unit size. End users make requests for different contents to the wireless edge. If the requested content is available (i.e., cached) at the wireless edge, then it is delivered to end users directly through a wireless channel that is unreliable [41]. The goal of the wireless edge is to decide at each time which contents to cache so that the cumulative value of the average content request latency experienced by end users is minimal.

Content Request and Delivery Model. We assume that requests for content $m \in \mathcal{M}$ arrive at the wireless edge from end users according to a Poisson process[1] with arrival rate λ_m. To each content m, we associate a "request queue" at the wireless edge, which stores the number of outstanding requests for content m at time t. The queue length associated with the number of such requests at time t is denoted by $S_{m,t}$. The rationality of this model is that the number of content requests may be larger than the service capacity of the wireless edge server. Hence, the content requested from an end user might not be served immediately so that there will be a latency associated with the end user getting content. Another point is due to the fact that the wireless channels between the edge cache and end users are unreliable. This motivates us to consider a queuing model which captures the latency experienced by end users.

The time taken by the wireless edge to deliver the content, i.e. serve end users' requests, is modeled by appropriate random variables, which heavily relies on the wireless channel quality between the wireless edge and end users.[2] More specifically, we assume that the time taken to deliver content m to end users is exponentially distributed with mean $1/\nu_m$ [41]. The service times are independent across different contents and requests. Thus, when $S_{m,t} \geq 1$, the request of content m departs from the corresponding request queue with rate ν_m.

Decision Epochs. The decision epochs/times are the moments when the states of request queues change. At each decision epoch/time t, the wireless edge determines for each content whether or not it should be cached, and then delivers the cached contents to the desired end users.

B. System Dynamics and Problem Formulation

We now formulate the problem of average latency minimization for the above model as a MDP.

States. We denote the state of the wireless edge at time t as $\mathbf{S}_t := (S_{1,t}, \cdots, S_{M,t}) \in \mathbb{N}^M$, where $S_{m,t}$ is the number of outstanding requests for content $m \in \mathcal{M}$. To guarantee the stability of the Markov chain, we assume that $S_{m,t} \in [0, S_{max}], \forall m, t$, where S_{max} can be arbitrarily large but bounded. For ease of readability, we denote the state-space associated with $\mathbf{S}_t$ as $\mathcal{S}$.

Actions. At each time t, for each content m, the wireless edge has to make a decision regarding whether or not to cache it. We use $A_{m,t}$ to denote the action for content m at time t. Thus, we let $A_{m,t} = 1$ if it is cached, and $A_{m,t} = 0$ otherwise. We let $\mathcal{A} := \{0, 1\}$ be the set of decisions available for each content, and let $\mathbf{A}_t := (A_{1,t}, \cdots, A_{M,t})$ be the vector consisting of decisions for M contents. The cache capacity constraint implies that $\mathbf{A}_t$ must satisfy the following constraints,

$$\sum_{m=1}^{M} A_{m,t} \leq B, \quad \forall t. \tag{1}$$

We aim to design a policy $\pi : \mathcal{S} \mapsto \mathcal{A}^M$ maps the state $\mathbf{S}_t$ of the wireless edge to caching decisions $\mathbf{A}_t$, i.e., $\mathbf{A}_t = \pi(\mathbf{S}_t)$.

Transition Kernel. The state of the m-th request queue can change from S_m to either $S_m + 1$, or $S_m - 1$ at the beginning of each decision epoch. Let $\mathbf{e}_m$ be the M-dimensional vector

[1] Poisson arrivals have been widely used in the literature, e.g., [16] and [17] and references therein. However, our model holds for general stationary process [42] and our model-free RL based algorithm and analysis in Sections V and VI holds for any request process.

[2] Though the wireless channel can be explicitly modeled as in physical-layer communication models [43], it requires additional beamforming and channel estimations, which is out of the scope of this work. With our queue model, the effect of wireless channel is incorporated in content departure rate as in [41].

whose m-th component is 1, and all others are 0. Then,

$$\mathbf{S} = \begin{cases} \mathbf{S} + \mathbf{e}_m, & \text{with transition rate } b_m(S_m, A_m), \\ \mathbf{S} - \mathbf{e}_m, & \text{with transition rate } d_m(S_m, A_m), \end{cases} \tag{2}$$

where $b_m(S_m, A_m) := \lambda_m$. We allow for state-dependent content delivery rates, which enables us to model realistic settings [25], [26]. In particular, our setup can cover the classic $M/M/k$ queue if $d_m(S_m, A_m) = \nu_m S_m A_m$. This models the general multicast scenario in which the wireless edge can simultaneously serve end users whose requested contents are cached at the edge.

Average Latency Minimization Problem. It follows from Little's Law [44] that the objective of minimizing the average latency faced by end users is equivalent to that of minimizing the average number of cumulative outstanding requests in the system. Let $C_{m,t}(S_{m,t}, A_{m,t}) := S_{m,t}$ be the instantaneous cost incurred by user m at time t, so that the cumulative cost incurred in the system at time t is given by

$$C_t(\mathbf{S}_t, \mathbf{A}_t) = \sum_{m=1}^{M} C_{m,t}(S_{m,t}, A_{m,t}) = \sum_{m=1}^{M} S_{m,t}. \tag{3}$$

With this choice of instantaneous cost, the average cost incurred in the system is proportional to the average latency faced by end users. Our objective is to derive a policy π that makes content caching decisions at the capacity-constrained wireless edge for solving the following MDP:

$$\min_{\pi \in \Pi} \; C_\pi := \limsup_{T\to\infty} \sum_{m=1}^{M} \frac{1}{T} \mathbb{E}_\pi \left[\int_0^T S_{m,t} dt \right], \quad \text{s.t.} \; \sum_{m=1}^{M} A_{m,t} \le B, \; \forall t, \tag{4}$$

where the subscript denotes the fact that the expectation is taken with respect to the measure induced by the policy π, and Π is the set of all feasible wireless edge caching policies. Henceforth, we refer to (4) as the "original MDP." Since it is an infinite-horizon average-cost problem, in principle it can be solved via the relative value iteration [2]. More specifically, there exists a value function, and an average cost value for the above MDP [2, Theorem 8.4.3]:

Lemma 1: Consider the MDP (4) whose transition kernel is described in (2). There exists a β^* and a function $V : \mathcal{S} \mapsto \mathbb{R}$ that satisfy the following dynamic programming (DP) equation*:*

$$\beta^* = \min_{\sum_m A_m \le B} \left(\sum_{m=1}^{M} \Big[S_m + \lambda_m V(\mathbf{S} + \mathbf{e}_m) + \nu_m S_m A_m V(\mathbf{S} - \mathbf{e}_m) - (\lambda_m + \nu_m S_m A_m) V(\mathbf{S}) \Big] \right). \tag{5}$$

Though one can obtain an optimal policy π^* using relative value iteration, this approach suffers from the curse of dimensionality, i.e., the computational complexity grows linearly with the size of state space $\mathcal{S}$, the latter quantity in turn grows exponentially with the number of contents M. This renders such a solution impractical. Furthermore, this approach fails to provide insight into the solution structure. Thus, our focus will be on developing computationally appealing solutions.

C. Lagrangian Relaxation

We now discuss Lagrangian relaxation of the original MDP (4), and introduce the corresponding "per-content MDP." The Lagrangian multipliers together with these per-content problems form the building block of our Whittle index policy, that will be formally introduced in Section IV.

Following Whittle's approach [3], we first consider the following *"relaxed problem,"* which relaxes the "hard" constraint in (4) to an average constraint:

$$\min_{\pi \in \Pi} \; \limsup_{T\to\infty} \sum_{m=1}^{M} \frac{1}{T} \mathbb{E}_\pi \left[\int_0^T S_{m,t} dt \right], \quad \text{s.t.} \; \limsup_{T\to\infty} \sum_{m=1}^{M} \frac{1}{T} \mathbb{E}_\pi \left[\int_0^T A_{m,t} dt \right] \le B. \tag{6}$$

Next, we consider the Lagrangian associated with (6),

$$L(\pi, W) := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \int_0^T \left\{ \sum_{m=1}^{M} S_{m,t} - W \left(B - \sum_{m=1}^{M} A_{m,t} \right) \right\}, \tag{7}$$

where W is the Lagrangian multiplier, and π is a wireless edge caching policy. The corresponding dual function is defined as

$$D(W) := \min_{\pi} L(\pi, W). \tag{8}$$

The dual problem corresponding to W is to optimize the Lagrangian $L(\pi, W)$ over the choice of π. For a fixed value of W, the dual problem (8) corresponding to the relaxed problem (6) decouples the original problem (4) into M "*per-content MDPs*," each of them involving only a single content. Specifically, the per-content MDP corresponding to the m-th content is given as follows,

$$\min_{\pi_m} \bar{C}_m := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{\pi_m} \left[\int_0^T \bar{C}(S_{m,t}, A_{m,t}) dt \right], \tag{9}$$

where $\bar{C}(S_{m,t}, A_{m,t}) := S_{m,t} - W(1 - A_{m,t})$ is the instantaneous cost incurred by m-th content, and π_m is a policy that makes decisions (only) for the m-th content. It then follows that in order to evaluate the dual function (8) at W, it suffices to solve all M independent per-content MDPs (9) [2]. The relaxed problem (6) can be solved by solving each of these M per-content MDPs, and then combining their solutions.

Unfortunately, this solution does not always provide a feasible wireless edge caching policy for the original problem (4), which requires that the cache capacity constraint (1) must be met at all times, rather than just in the average sense as in the constraint (6). Whittle index policy, which we discuss next, combines these solutions corresponding to per-content MDPs in such a way that the resulting policy is also feasible for the original problem (4), i.e, it satisfies the hard constraint.

IV. Whittle Index Policy

We now describe the Whittle index policy that will be utilized for making decisions for wireless edge caching. To the best of our knowledge, Whittle index policy has not been used previously to solve this problem. More specifically, the wireless edge caching problem (4) can be posed as a RMAB problem in which each content $m \in \mathcal{M}$ can be viewed as an arm m, and playing arm m would correspond to cache content m. At each time t, the queue length $S_{m,t}$ of the corresponding

request queue is the state of arm m, and $A_{m,t}$ is the action taken for content m. $A_{m,t} = 1$ denotes that content m is cached at time t, and $A_{m,t} = 0$ otherwise. It is well-known that Whittle index policy is a computationally tractable solution to the RMAB problem since its computational complexity scales linearly with the number of arms M. For ease of readability, we relegate all proofs in this section to Appendix A.

A. Indexability and Whittle Index

Whittle index policy is defined for a RMAB only when the underlying problem is "indexable" [3]. Thus, we begin by showing that our MDP is *indexable*. Loosely speaking, to show that the problem is indexable, we need to consider the single-arm (per-content) MDP (9) and then need to show that the set of states in which the optimal action is passive (i.e., not to cache) increases as the Lagrangian multiplier W increases. We definite it formally here for completeness.

Definition 1 (Indexability): Consider the per-content MDP (9) for the m-th content. Let $D_m(W)$ be the set of states in which the optimal action for the per-content MDP (9) is to choose the passive action, i.e., $A_m = 0$. Then the m-th MDP is said to be indexable if $D_m(W)$ increases with W, i.e., if $W > W'$, then $D_m(W) \supseteq D_m(W')$. The original MDP (4) is indexable if all of the M per-content MDPs (9) are indexable.

In case that a MDP is indexable, the Whittle index for each content/arm is defined as follows.

Definition 2 (Whittle Index): If the per-content MDP (9) for the m-th content is indexable, the Whittle index in state S is the smallest value of the Lagrangian multiplier W such that the optimal policy is indifferent towards actions $A_m = 0$ and $A_m = 1$ when the Lagrange multiplier is set equal to this value. We denote this Whittle index by $W_m(S)$, satisfying $W_m(S) := \inf_{W \geq 0}\{S \in D_m(W)\}$.

B. The Per-Content MDP (9) Is Indexable

Our proof of indexability relies on the "threshold" property of the optimal policy for each per-content MDP (9), i.e., for each content $\forall m \in \mathcal{M}$, it is optimal to cache this content only when the number of outstanding requests for it is above a certain threshold; this threshold might depend upon m. We begin by analyzing this MDP for a fixed m, and thus drop the subscript m in the rest of this subsection for ease of exposition. One key observation is that although the inherent MDP in (9) operates in continuous time, decisions are made only at those time instants when either a new content request arrives, or a content delivery occurs. Otherwise, there is no state transition (see Section III-A for details). With this observation, those time instants together can be treated as in the discrete-time domain (samples from the original time domain). Hence, for the per-content MDP in (9), the state transitions under an action are sampled on discrete-time instants. For example, the MDP transitions from state s to state $s+1$ with probability $\frac{\lambda}{\lambda+\nu s a}$, and from state s to state $s-1$ with probability $1 - \frac{\lambda}{\lambda+\nu s a}$, according to the controlled birth-and-death process. This reduces our considered continuous-time "per-content" MDP (9) to a discrete-time MDP problem and each time instance when a state transition occurs corresponds to one step in the discrete-time MDP problems. To this end, we show the threshold property by analyzing the per-content MDP (9) in its discrete-time equivalently.

1) Threshold Property of an Optimal Policy: We start by analyzing an associated discounted cost MDP, rather than the average latency problem. After analyzing the discounted MDP, we extend our results to the case of average latency problem (9). The discounted latency problem corresponding to (9) is given as follows,

$$\min_{\pi} \mathbb{E}_{\pi}\left[\lim_{T\to\infty}\sum_{t=1}^{T}\alpha^{t-1}\bar{C}(S_t, A_t)|S_0 = s\right], \tag{10}$$

where $\alpha \in (0,1)$ is a discount factor. It is well-known that there exists an optimal stationary deterministic policy for this discounted latency problem [2], and hence we will restrict ourselves to the class of stationary deterministic policies while solving this problem. We apply the value iteration method to find the optimal policy.

Let U denote the Banach space of bounded real-value functions on $\mathbb{N}$ with supremum norm. Define the operator $\mathcal{T} : U \to U$ as follows,

$$(\mathcal{T}u)(s) := \min_{a\in\{0,1\}} \bar{C}(s,a) + \alpha\mathbb{E}[u(s')], \tag{11}$$

where $u(\cdot) \in U$ and the expectation is taken with respect to the distribution of state s' which results when action a is applied in state s. Let $J^{\alpha}(s)$ denote the optimal expected total discounted cost incurred by the system when it starts in state s. Then we have that $J^{\alpha}(s) = \mathcal{T}J^{\alpha}(s)$, i.e., $J^{\alpha}(s)$ is a solution of the Bellman equation satisfying

$$J^{\alpha}(s) = \min_{a\in\{0,1\}} \bar{C}(s,a) + \alpha\mathbb{E}[J^{\alpha}(s')]. \tag{12}$$

As is described in (2), s' can only assume values from the set $\{s-1, s+1\}$. Let $P_{s,a} := \frac{\lambda}{\lambda+\nu s a}$, so that (12) can be written compactly as $J^{\alpha}(s) =$

$$\min_{a\in\{0,1\}} \bar{C}(s,a) + \alpha\Big(P_{s,a}J^{\alpha}(s+1) + (1-P_{s,a})J^{\alpha}(s-1)\Big). \tag{13}$$

Define the state-action value function $\forall s \in \mathcal{S}, a \in \{0,1\}$ as:

$$Q^{\alpha}(s,a) := \bar{C}(s,a) + \alpha\Big(P_{s,a}J^{\alpha}(s+1) + (1-P_{s,a})J^{\alpha}(s-1)\Big). \tag{14}$$

Therefore, we have $J^{\alpha}(s) = \min_{a\in\{0,1\}} Q^{\alpha}(s,a)$.

We need the following assumption on the underlying MDPs in order to ensure that the Whittle indices $W(s)$ are finite.

Assumption 1: For each state s, there exists a finite $W(s) > 0$ such that the optimal action is of equal preference in activating and not activating the arm (caching or not caching the content), i.e. $Q^{\alpha}(s,1) = Q^{\alpha}(s,0)$.

We now show that for each value of W, the optimal policy for the per-content MDP (10) is of threshold-type.

Proposition 1: Consider the discounted latency MDP (10) with a fixed $W \geq 0$. There exists an optimal policy for (10) that is of threshold-type with the threshold depending on W.

Remark 1: Existing works [45], [46] among others have also used the threshold structure of an optimal policy in order to show that the underlying MDP is indexable. The key is to show that if the optimal action for state s is to keep the arm active ($a = 1$), i.e., $Q^{\alpha}(s,1) \leq Q^{\alpha}(s,0)$, then the optimal action for state $s+1$ is also to keep it active, i.e., $Q^{\alpha}(s+1,1) \leq Q^{\alpha}(s+1,0)$. The threshold structure in turn is shown by considering the corresponding discounted MDP, and proving for this discounted problem that its value function

$Q^{\alpha}(\cdot)$ of the underlying MDP is convex [45], or monotone [46]. Works such as [45] and [46] showed that these properties hold, but then the associated MDPs in these works have transition rates that are not a function of state. In contrast, the transition rates in our MDPs are a function of state, and hence we cannot use existing results directly.

The following proposition extends our results in Proposition 1 for the discounted latency problem in (10) to the original average latency problem in (9).

Proposition 2: There exists an optimal stationary policy of the threshold-type for the average latency problem in (9).

2) Indexability of the Per-Content MDP (9): We now show that the per-content MDP (9) is indexable.

Proposition 3: The per-content MDP (9) is indexable.

Proposition 4: Let $\{\phi_R(s)\}_{s=0}^{S_{max}}$ be the stationary distribution of the Markov process which results when the threshold policy with threshold value R is applied. If the function $\frac{\sum_s s\phi_R(s)-\sum_s s\phi_{R-1}(s)}{\sum_{s=0}^{R}\phi_R(s)-\sum_{s=0}^{R-1}\phi_{R-1}(s)}$ is non-decreasing in R, then the Whittle indices of the per-content MDP (9) are given as follows,

$$W(R) := \frac{\sum_s s\phi_R(s) - \sum_s s\phi_{R-1}(s)}{\sum_{s=0}^{R}\phi_R(s) - \sum_{s=0}^{R-1}\phi_{R-1}(s)}. \tag{15}$$

From (15), it is clear that the stationary distribution of the threshold policy is required to compute the Whittle indices.

Proposition 5: The stationary distribution $\{\phi_R(s)\}_{s=0}^{S_{max}}$ of the threshold policy with threshold value R satisfies

$$\begin{aligned}
\phi_R(s) &= 0,\ s = 0, 1, \ldots, R-1,\\
\phi_R(R) &= \frac{\nu(R+1)}{\lambda+\nu(R+1)}\cdot\phi_R(R+1),\\
\phi_R(R+1) &= 1/\Bigg(1+\frac{\nu(R+1)}{\lambda+\nu(R+1)}\\
&\quad+\sum_{l=2}^{S_{max}-R}\prod_{j=2}^{l}\frac{\lambda}{\lambda+\nu(R+j-1)}\frac{\lambda+\nu(R+j)}{\nu(R+j)}\Bigg),\\
\phi_R(R+l) &= \phi_R(R+1)\prod_{j=2}^{l}\frac{\lambda}{\lambda+\nu(R+j-1)}\\
&\quad\cdot\frac{\lambda+\nu(R+j)}{\nu(R+j)},\ l = 2, \cdots, S_{max}-R.
\end{aligned} \tag{16}$$

C. Whittle Index Policy

We now describe how the solutions to the relaxed problem (6) are used to obtain a policy for the original problem (4). Whittle index policy assigns an index $W_m(S_{m,t})$ to the queues of each content $m \in \mathcal{M}$. This index $W_m(S_{m,t})$ depends upon current state $S_{m,t}$ and current time. The Whittle index policy then activates (caches) B arms (contents) with the highest value of the indices $W_m(S_{m,t})$. Although this policy need not to be optimal for the original problem (4), it has been shown to be asymptotically optimal [4], [5] as the number of contents M and the cache size B are scaled up, while keeping their ratio as a constant.

V. Whittle Index Based Q-Learning With LFA

In order to implement the Whittle index policy that was discussed in Section IV, one needs to know the controlled transition probabilities of each of the M per-content MDPs. Since this information is often not available, and moreover these parameters are time-varying, we now develop reinforcement learning algorithms that learn the Whittle index policy for wireless edge caching. Specifically, we design a model-free reinforcement learning augmented algorithm with linear function approximation (LFA), which we call `Q+-Whittle-LFA`, which leverages the threshold structure of the optimal policy developed in Section IV while learning Q-functions for different state-action pairs. Similar to Section IV, we focus on learning the Whittle index for each per-content MDP, and hence drop the subscript m for ease of presentation. Again, as discussed in Section IV, though the whole system (the content caching process) operates in continuous time, decisions are made only at those time instants when either a new content request arrives, or a content delivery occurs (see Section III-A). Those time instants can be treated as in the discrete-time domain. Hence, in this section, samples are made on discrete time instants.

A. Preliminaries

We first review some preliminaries for Q-learning for Whittle index policy, which was first proposed in [32] for the discounted cost setup and further generalized in [9] for average cost setup.

Consider the dynamic programming equations associated with the per-content MDP in (9),

$$\begin{aligned}
V(s)+\tilde{\beta}^* = \min_{a\in\{0,1\}}\Bigg\{&a\Big(s+\sum_{s'}p(s'|s,1)V(s')\Big)\\
&+(1-a)\Big(s-W+\sum_{s'}p(s'|s,0)V(s')\Big)\Bigg\},
\end{aligned} \tag{17}$$

where $\tilde{\beta}^* \in \mathbb{R}$ is the optimal value of the long-term average cost of the MDP with the Lagrange multiplier set equal to W, and $V(\cdot)$ is the relative value function. The corresponding Q-function is given as follows [47],

$$Q(s,a)+\tilde{\beta}^* = s-(1-a)W(s)+\sum\nolimits_{s'}p(s'|s,a)V(s'), \tag{18}$$

where value function $V(\cdot)$ satisfies $V(s) = \min_{a\in\{0,1\}}Q(s,a)$. We now discuss a relation satisfied by the Whittle indices $\{W(s)\}_{s\in\mathcal{S}}$, that was derived in [32]. When the Lagrange multiplier W is set equal to the Whittle index $W(s)$, actions 0 and 1 are equally favorable in state s, i.e., $Q(s,0) = Q(s,1)$. Substituting for $Q(s,a)$ from (18) into the relation $Q(s,0) = Q(s,1)$, we obtain the following relation for $W(s)$,

$$W(s) = \sum\nolimits_{s'}p(s'|s,0)V(s') - \sum\nolimits_{s'}p(s'|s,1)V(s'). \tag{19}$$

The work [9] proposed a tabular Whittle index-based Q-learning algorithm, which we call `Q-Whittle` for ease of exposition. The key aspect of `Q-Whittle` is that the updates of Q-function values and Whittle indices form a two-timescale stochastic approximation (2TSA), where the Q-function values are updated at a faster timescale for a given $W(s)$, and the Whittle indices are updated at a slower timescale. More precisely, the Q-function values are updated as follows,

$$Q_{n+1}(s,a) = Q_n(s,a)+\gamma_n\mathbb{1}_{\{S_n=s,A_n=a\}}\Big(s-(1-a)W(s)$$

$$+\max_a Q_n(S_{n+1},a)-I(Q_n)-Q_n(s,a)\Big),\quad n=0,1,\ldots, \tag{20}$$

where the subscript n denotes the decision epoch for the per-content MDP in (9), and $I(\cdot)$ is a reference function [2], [48]. Recall that decision epoch represents the moment when state of the per-content MDP changes. Note that reference functions are used only when performing relative Q-learning iterations for the average cost setup, and not used while optimizing cumulative discounted rewards. $\{\gamma_n\}$ is a step-size sequence satisfying $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$. Accordingly, the Whittle indices are updated as follows,

$$W_{n+1}(s) = W_n(s) + \eta_n(Q_n(s,0) - Q_n(s,1)), \tag{21}$$

with the step-size sequence $\{\eta_n\}$ satisfying $\sum_n \eta_n = \infty$, $\sum_n \eta_n^2 < \infty$ and $\eta_n = o(\gamma_n)$. The coupled iterates (20) and (21) form a 2TSA, and a rigorous asymptotic convergence guarantee is provided in [9].

B. Q$^+$-Whittle

While [9] proposed a Q-learning based algorithm for learning Whittle indices, Q-Whittle requires a reference function $I(Q_n)$ in order to approximate the unknown parameter $\tilde{\beta}^*$. It is not clear how one should choose the reference function, since there is no unique choice and this function might be problem dependent. To circumvent this problem, a widely-adopted approach is to instead learn an optimal policy for the corresponding discounted-cost MDP, that differs from the average cost MDP only in that the future rewards are discounted. It follows from classical results on MDPs [49] that there exists a stationary deterministic policy that is optimal for all values of discount factor α that are sufficiently close to 1. Moreover this policy is also optimal for the average-cost MDP. This policy is known as the Blackwell optimal policy. Such a technique has been applied to the study of average-cost MDP in [32] and [50]. We will adopt a similar approach, and hence now focus on the discounted Q-learning.

We now use the structural result regarding an optimal policy for the per-content MDP (9) in order to reduce the exploration overhead associated with the updates of Q-functions. Specifically, by specializing the Q-learning iterations for a threshold policy with threshold R, one needs to update the Q-function values $Q(s,0)$ only for states $s = 1,2,\ldots,R-1$ (and not for states $s \geq R$), while all other state-action values are left unchanged since the optimal action for all $s < R$ is deterministic, i.e., $a = 0$. Similarly, for action $a = 1$, we only need to update the Q-function values $Q(s,1)$ for $s > R$. When the arm is in state R, it randomizes between actions 0 and 1. To keep the discussion simple, we assume that these two actions are chosen with equal probability when state is R. This key observation drastically reduces the complexity of Q-learning when it is applied to learn Whittle indices, as compared with the existing Q-Whittle [9]. Towards this end, we call this improved version of Q-Whittle algorithm, one which leverages the threshold structure of the optimal policy, as Q$^+$-Whittle.

Specifically, we consider the problem of learning the Whittle index for state $s = R$, and develop a recursive update scheme for learning it. Let $Q_n^R(S_n, A_n)$ be the Q-function value during iteration n with dependence on R. The Q-function updates of Q$^+$-Whittle are given as follows:

Case 1: When $S_n > R$, we have

$$Q_{n+1}^R(S_n,1) \leftarrow (1-\gamma_n)Q_n^R(S_n,1) + \gamma_n S_n + \gamma_n\Big(\underbrace{\alpha \mathbb{1}_{(S_{n+1}>R)} Q_n^R(S_{n+1},1)}_{\text{Term}_{11}} + \underbrace{\alpha \mathbb{1}_{(S_{n+1}<R)} Q_n^R(S_{n+1},0)}_{\text{Term}_{12}} + \underbrace{\alpha \mathbb{1}_{(S_{n+1}=R)} \min_a Q_n^R(S_{n+1},a)}_{\text{Term}_{13}}\Big), \tag{22}$$

where the step-size sequence $\{\gamma_n\}$ satisfies $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$. Term$_{11}$ follows from the above insight that only Q-function values for states greater than R, i.e. $Q_n^R(S_n,1), S_n > R$ need to be updated. This differs significantly from Q-Whittle [9], where both $Q_n^R(\cdot,1)$ and $Q_n^R(\cdot,0)$ need to be updated when $S_n > R$. This is due to the fact that our Q$^+$-Whittle leverages the threshold-type optimal policy while performing Q-function updates, which either does not exist or is not leveraged in [9] and [32]. Similar insights lead to the updates of Term$_{12}$ and Term$_{13}$.

Case 2: When $S_n < R$, we have

$$Q_{n+1}^R(S_n,0) \leftarrow (1-\gamma_n)Q_n^R(S_n,0) + \gamma_n(S_n - W_n(R)) + \gamma_n\Big(\underbrace{\alpha \mathbb{1}_{(S_{n+1}>R)} Q_n^R(S_{n+1},1)}_{\text{Term}_{21}} + \underbrace{\alpha \mathbb{1}_{(S_{n+1}<R)} Q_n^R(S_{n+1},0)}_{\text{Term}_{22}} + \underbrace{\alpha \mathbb{1}_{(S_{n+1}=R)} \min_a Q_n^R(S_{n+1},a)}_{\text{Term}_{23}}\Big), \tag{23}$$

where the updates of Term$_{21}$, Term$_{22}$ and Term$_{23}$ leverage similar insights as those in Case 1.

Case 3: When $S_n = R$, $Q_n^R(S_n, A_n)$ gets updated according to either (22) or (23) with equal probability.

In summary, the Q-function updates of Q$^+$-Whittle are given as

$$Q_{n+1}^R(s,a) = \begin{cases} (22) \text{ or } (23), & \text{if } (s,a) = (S_n, A_n), \\ Q_n^R(s,a), & \text{otherwise.} \end{cases} \tag{24}$$

With the above Q-function updates, the parameter W under the threshold policy with threshold R is updated as follows,

$$W_{n+1}(R) = W_n(R) + \eta_n\Big(Q_n^R(R,0) - Q_n^R(R,1)\Big), \tag{25}$$

where the step-size sequence $\{\eta_n\}$ satisfies $\sum_n \eta_n = \infty$, $\sum_n \eta_n^2 < \infty$ and $\eta_n = o(\gamma_n)$.

Q$^+$-Whittle is summarized in Algorithm 1. Since we are learning the Whittle index for every state, the algorithm will loop for all states. Q-function and W updates discussed above remain the same for all M contents (lines 4-8). Since the wireless edge can cache at most B contents, an easy implementation is to find the possible activation set $\mathcal{C} := \{m \in \mathcal{M} | S_m(t) \geq R\}$ for threshold R at time/epoch t and activate $\min(B, |\mathcal{C}|)$ arms with highest Whittle indices $W_{m,t}(S_{m,t})$. Note that t is the moment when the state of any of the M per-content MDPs changes.

Remark 2: Some definitions (e.g., $W(s)$) in this paper are similar to those in [9] and [32], which studied Q-Whittle through a two-timescale update. However, our Q$^+$-Whittle differs from those in [9] and [32] from two perspectives. First, [9] and [32] adopted the conventional ϵ-greedy rule for Q-function value updates. In contrast, we leverage the property

Algorithm 1 $\texttt{Q}^+\texttt{-Whittle}$ for Per-Content MDP

1: Initialize: $Q_0^{s'}(s,a)=0,\ W_0(s)=0,\ \forall s,s'\in\mathcal{S}$.
2: **for** $R\in\mathcal{S}$ **do**
3: Set the threshold policy as $\pi=R$.
4: **for** $n=1,2,\ldots,T$ **do**
5: Update $Q_n^R(s_n,a_n)$ according to (24).
6: Update $W_n(R)$ according to (25).
7: **end for**
8: $W_0(R+1)=W_T(R)$, $Q_0^{R+1}(s,a)=Q_T^R(s,a)$.
9: **end for**
10: Return: $W(s),\forall s\in\mathcal{S}$.

Algorithm 2 $\texttt{Q}^+\texttt{-Whittle-LFA}$ for Per-Content MDP

1: Initialize: $\phi(s,a),\ \theta_0, W_0(s)=0$ for $\forall s\in\mathcal{S},a\in\mathcal{A}$.
2: **for** $R\in\mathcal{S}$ **do**
3: Set the threshold policy as $\pi=R$.
4: **for** $n=1,2,\ldots,T$ **do**
5: Update θ_n^R according to (28).
6: Update $W_n(R)$ according to (29).
7: **end for**
8: $W_0(R+1)=W_T(R)$, $\theta_0^{R+1}(s,a)=\theta_T^R(s,a)$.
9: **end for**
10: Return: $W(s),\forall s\in\mathcal{S}$.

of optimal threshold-type policy into Q-function value updates as in (22) and (23). Such a threshold-type Q-function value update dramatically reduces the computational complexity since each state only has a fixed action to explore. Second, the threshold policy further enables us to update Whittle indices in an incremental manner, i.e., the converged Whittle index in state s can be taken as the initial value for the subsequent state $s+1$ (line 8 in Algorithm 1), instead of being randomly initiated as in [9] and [32]. This further speeds up the learning process. In addition, [32] lacked convergence guarantee. Recently, another line of work [51] leveraged Q-learning to approximate Whittle indices through a single-timescale SA, where Q-function and Whittle indices were learned independently. Reference [51] considered the finite-horizon MDP and cannot be directly applied to infinite-horizon discounted or average cost MDPs considered in this paper.

C. $\texttt{Q}^+\texttt{-Whittle}$ With Linear Function Approximation

When the number of state-action pairs is very large, which is often the case for wireless edge caching, $\texttt{Q}^+\texttt{-Whittle}$ can be intractable due to the curse of dimensionality. A closer look at (25) further reveals that the Whittle index is updated only when state s is visited. This can significantly slow down the convergence process of the corresponding 2TSA when the state space is large. To overcome this difficulty, we further study $\texttt{Q}^+\texttt{-Whittle}$ with linear function approximation (LFA) by using low-dimensional linear approximation of Q on a linear subspace with dimension $d\ll|\mathcal{S}||\mathcal{A}|$. We call this algorithm as $\texttt{Q}^+\texttt{-Whittle-LFA}$.

Specifically, given a set of basis functions $\phi_\ell:\mathcal{S}\times\mathcal{A}\mapsto\mathbb{R},\ \ell=1,\cdots,d$, the approximation of the Q-function $\tilde{Q}_\theta(s,a)$ parameterized by a unknown weight vector $\theta\in\mathbb{R}^d$, is given by $\tilde{Q}_\theta(s,a)=\phi(s,a)^\intercal\theta,\ \forall s\in\mathcal{S},a\in\mathcal{A}$, where $\phi(s,a)=(\phi_1(s,a),\cdots,\phi_d(s,a))^\intercal$. The feature vectors are assumed to be linearly independent and are normalized so that $\|\phi(s,a)\|\leq 1,\forall s\in\mathcal{S},a\in\mathcal{A}$.

Similar to $\texttt{Q}^+\texttt{-Whittle}$, we consider the problem of learning the Whittle index for state $s=R$, which can be equivalently formulated as the problem of learning the coefficient θ. Let θ_n^R be its value during iteration n, which depends on the value of R. Leveraging the same ideas for Q-function updates in (24), $\texttt{Q}^+\texttt{-Whittle-LFA}$ iteratively updates θ_n^R as follows:

Case 1: When $S_n>R$, we have

$$\theta_{n+1}^R\leftarrow\theta_n^R+\gamma_n\phi(S_n,1)\Big[S_n+\alpha\mathbb{1}_{(S_{n+1}>R)}\phi(S_{n+1},1)^\intercal\theta_n^R+\alpha\mathbb{1}_{(S_{n+1}<R)}\phi_n(S_{n+1},0)^\intercal\theta_n^R+\alpha\mathbb{1}_{(S_{n+1}=R)}\min_a\phi(S_{n+1},a)^\intercal\theta_n^R-\phi(S_n,1)^\intercal\theta_n^R\Big]. \tag{26}$$

Case 2: When $S_n<R$, we have,

$$\theta_{n+1}^R\leftarrow\theta_n^R+\gamma_n\phi(S_n,0)\Big[(S_n-W)+\alpha\mathbb{1}_{(S_{n+1}>R)}\phi_n(S_{n+1},1)^\intercal\theta_n^R+\alpha\mathbb{1}_{(S_{n+1}<R)}\phi_n(S_{n+1},0)^\intercal\theta_n^R+\alpha\mathbb{1}_{(S_{n+1}=R)}\min_a\phi_n(S_{n+1},a)^\intercal\theta_n^R-\phi(S_n,0)^\intercal\theta_n^R\Big]. \tag{27}$$

Case 3: When $S_n=R$, the update occurs either according to (26) or (27) with an equal probability.

The iterations can be summarized as follows,

$$\theta_{n+1}^R=\begin{cases}(26)\text{ if }S_n>R,\\(27)\text{ if }S_n<R,\\(26)\text{ or }(27),\text{ if }S_n=R.\end{cases} \tag{28}$$

We now derive a similar iterative scheme for learning the Whittle indices. To do this, we consider the Whittle index update in (25), and replace the Q-function values $Q_n^R(R,0),Q_n^R(R,1)$ by their linear function approximations $\phi(R,0)^\intercal\theta_n^R$ and $\phi(R,1)^\intercal\theta_n^R$, respectively. This gives us the following iterations,

$$W_{n+1}(R)=W_n(R)+\eta_n(\phi(R,0)^\intercal\theta_n^R-\phi(R,1)^\intercal\theta_n^R). \tag{29}$$

We summarize $\texttt{Q}^+\texttt{-Whittle-LFA}$ in Algorithm 2, which is one of our key contributions in this paper.

VI. Finite-Time Performance Analysis

In this section, we provide a finite-time analysis of our $\texttt{Q}^+\texttt{-Whittle-LFA}$ algorithm, which can be viewed through the lens of 2TSA. Our key technique is motivated by [52], which deals with a general nonlinear 2TSA. To achieve this goal, we first need to rewrite our $\texttt{Q}^+\texttt{-Whittle-LFA}$ updates in (28) and (29) in the form of a 2TSA. Throughout this section, we will perform the analysis for any threshold policy $\pi=R$, and hence we drop the superscript R for ease of presentation.

A. Two-Timescale Stochastic Approximation

Given the threshold policy $\pi = R$, the corresponding true Whittle index associated with the threshold state R is $W(R)$. Denote θ^R as the optimal θ obtained by $\mathtt{Q^+\text{-}Whittle\text{-}LFA}$ in Algorithm 2. Following the conventional ODE method [10], we begin by converting $\mathtt{Q^+\text{-}Whittle\text{-}LFA}$ in (28) and (29) into a standard 2TSA. In particular, we rewrite the updates (28) and (29) as follows,

$$\theta_{n+1} = \theta_n + \gamma_n[h(\theta_n, W_n) + \xi_{n+1}], \tag{30}$$
$$W_{n+1} = W_n + \eta_n[g(\theta_n, W_n) + \psi_{n+1}], \tag{31}$$

where $\{\xi_n\}$ is an appropriate martingale difference sequence with respect to the filtration σ-field $\mathcal{F}_n = \{\theta_0, W_0, \xi_0, \ldots, \theta_n, W_n, \xi_n\}, n = 1, 2, \ldots$; $\{\psi_n\}$ is a suitable error sequence; h and g are appropriate Lipschitz functions defined below that satisfy the conditions needed for our ODE analysis, and the step sizes γ_n, η_n satisfy Assumption 5 below. Note that the θ_n and W_n iterations are coupled. By using the operator defined in (11), we rewrite the θ update in (28) as follows,

$$\theta_{n+1} = \theta_n + \gamma_n \phi(S_n, A_n)\Big[[\mathcal{T}\theta_n](S_n, A_n) - \phi(S_n, A_n)^\intercal \theta_n + \xi_{n+1}(S_n, A_n)\Big], \quad \forall (S, A) \in \mathcal{S} \times \mathcal{A}, \tag{32}$$

where,

$$[\mathcal{T}\theta_n](S_n, A_n) = S_n - (1 - A_n)W_n + \alpha \sum_{s'} p(s'|S_n, A_n) \min_{a'} \phi(s', a')^\intercal \theta_n, \tag{33}$$

$$\xi_{n+1}(S_n, A_n) = S_n - (1 - A_n)W_n + \alpha \min_a \phi(S_{n+1}, a)^\intercal \theta_n - [\mathcal{T}\theta_n](S_n, A_n). \tag{34}$$

Hence, we have $h(\theta_n, W_n)$ in (30) as

$$h(\theta_n, W_n) := [\mathcal{T}\theta_n](S_n, A_n) - \phi(S_n, A_n)^\intercal \theta_n, \tag{35}$$

which is Lipschitz in both θ and W. Similarly, we have

$$g(\theta_n, W_n) := \phi(R, 0)^\intercal \theta_n - \phi(R, 1)^\intercal \theta_n, \tag{36}$$

which is Lipschitz in θ. W.l.o.g., we assume $\psi_n = 0$ for all n since the update of W in (29) is deterministic. After having identified these two functions, i.e., (35) and (36), the asymptotic convergence of our 2TSA can be established by using the ODE method [9], [10], [52], [53]. For ease of exposition, we temporally assume fixed step size here, then the 2TSA is reduced to the following differential equations:

$$\dot{\theta}(t) = h(\theta(t), W(t)), \quad \dot{W}(t) = \frac{\eta}{\gamma} g(\theta(t), W(t)), \tag{37}$$

where the ratio η/γ represents the difference in timescale between these two updates. Our focus here is on characterizing the finite-time convergence rate of (θ_n, W_n) to the globally asymptotically optimal equilibrium point $(\theta^R, W(R))$ of (37) for each R. Using an idea of [52], the key part of our analysis is based upon an appropriate choice of two step sizes η_n, γ_n, and a Lyapunov function. We first define the following two "error terms,"

$$\tilde{\theta}_n := \theta_n - f(W_n), \quad \tilde{W}_n := W_n - W(R), \tag{38}$$

which characterizes the coupling between θ_n and W_n. If we are able to show that $\tilde{\theta}_n$ and $\tilde{W}_n$ simultaneously converge to zero, then we would have shown $(\theta_n, W_n) \to (\theta^R, W(R))$. Thus, to prove the convergence of (θ_n, W_n) of our 2TSA to its true value $(\theta^R, W(R))$, we instead study the convergence of $(\tilde{\theta}_n, \tilde{W}_n)$ by providing the finite-time analysis for the mean squared error generated by (30)-(31). In order to simultaneously study the properties of $\tilde{\theta}_n$ and $\tilde{W}_n$, we consider the following Lyapunov function

$$M(\theta_n, W_n) := \frac{\eta_n}{\gamma_n}\|\tilde{\theta}_n\|^2 + \|\tilde{W}_n\|^2 = \frac{\eta_n}{\gamma_n}\|\theta_n - f(W_n)\|^2 + \|W_n - W(R)\|^2. \tag{39}$$

We make the following assumptions while analyzing (30)-(31).

Assumption 2: Provided any $W \in \mathbb{R}$, there exists an operator f such that $\theta = f(W)$ is the unique solution to $h(\theta, W) = 0$, where h and f are Lipschitz continuous with positive constants L_h and L_f such that

$$\|f(W) - f(W')\| \le L_f\|W - W'\|,$$
$$\|h(\theta, W) - h(\theta', W')\| \le L_h(\|\theta - \theta'\| + \|W - W'\|). \tag{40}$$

The operator g in (31) is Lipschitz continous with constant L_g, i.e.,

$$\|g(\theta, W) - g(\theta', W')\| \le L_g(\|\theta - \theta'\| + \|W - W'\|). \tag{41}$$

Remark 3: The Lipschitz continuity of the functions f, g, h guarantees the existence of solutions to the ODEs (37). Note that when h and g are linear functions of θ and W, Assumption 2 is automatically satisfied. This assumption is widely used for both linear and nonlinear 2TSA [52], [54], [55].

Assumption 3: There exist $\mu_1 > 0$ and $\mu_2 > 0$ such that

$$\tilde{\theta}^\intercal h(\theta, W) \le -\mu_1\|\tilde{\theta}\|^2, \ \forall \theta, \tilde{\theta} \in \mathbb{R}^d, \ W \in \mathbb{R},$$
$$\tilde{W} g(\theta, W) \le -\mu_2\|\tilde{W}\|^2, \ \forall \tilde{W}, W \in \mathbb{R}, \theta \in \mathbb{R}^d. \tag{42}$$

Remark 4: Assumption 3 guarantees the uniqueness of the solution to the ODEs (37). This assumption can be viewed as a relaxation of the monotone property of the nonlinear mappings [52], [54], since it is automatically satisfied if h and g are strongly monotone as is assumed in [52].

Assumption 4: Random variables ξ_n are independent of each other and across time, with zero mean and bounded variances

$$\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0, \qquad \mathbb{E}[\|\xi_n\|^2|\mathcal{F}_{n-1}] \le \Lambda,$$

where $\Lambda > 0$.

Assumption 5: The step sizes γ_n and η_n satisfy $\sum_{n=0}^{\infty}\gamma_n = \sum_{n=0}^{\infty}\eta_n = \infty, \sum_{n=0}^{\infty}\gamma_n^2 < \infty, \sum_{n=0}^{\infty}\eta_n^2 < \infty$, η_n/γ_n is non-increasing in n and $\lim_{n\to\infty}\eta_n/\gamma_n = 0$.

Remark 5: These assumptions are standard in SA literature [9], [10], [52], [53]. Assumption 4 holds since $\xi_n(s, a) = s - (1 - a)W_n + \alpha \min_a \phi(S_{n+1}, a)^\intercal\theta_n - [\mathcal{T}\theta_n](s, a)$, thus $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$.

B. Finite-Time Analysis of $\mathtt{Q^+\text{-}Whittle\text{-}LFA}$

Theorem 1: Consider the iterates $\{\theta_n\}$ and $\{W_n\}$ generated by (28) and (29) for learning the Whittle indices, and suppose that Assumptions 2-5 hold true. Let the step-sizes be chosen as $\gamma_n = \frac{\gamma_0}{(n+1)^{5/9}}, \eta_n = \frac{\eta_0}{(n+1)^{10/9}}$. Then we have

$$\mathbb{E}[M(\theta_{n+1}, W_{n+1})|\mathcal{F}_n] \le \frac{\mathbb{E}[M(\theta_0, W_0)]}{(n+1)^2}$$

$$+\frac{C_1(\|\tilde{\theta}_0\|^2+\|\tilde{W}_0\|^2)}{(n+1)^{2/3}}+\frac{\gamma_0\eta_0\Lambda}{(n+1)^{2/3}},\ n=1,2,\ldots, \quad (43)$$

where $C_1=(L_h^2+L_f^2+2L_g^2(L_f+1)^2)\alpha_0\eta_0+2L_g^2(L_f+1)^2\left(L_f^2+(1+L_h\alpha_0)^2\right)\frac{\eta_0^3}{\gamma_0^3}$.
The first term of the right hand side of (43) corresponds to the bias due to the initialization, which goes to zero at a rate $\mathcal{O}(1/n^2)$. The second term corresponds to the accumulated estimation error of the nonlinear 2TSA. The third term stands for the error introduced due to the fluctuations of the martingale difference noise sequence $\{\xi_n\}$ in (30). The second and third terms in the right hand side of (43) decay at a rate $\mathcal{O}(1/n^{2/3})$, and hence dominate the overall convergence rate in (43). The proof is presented in Appendix B.

Remark 6: Our finite-time analysis of Q^+*-Whittle-LFA consists of two steps. First, we rewrite* Q^+*-Whittle-LFA updates into a 2TSA in (30)-(31). The key is to identify two critical terms* h *and* g*. Second, we prove a bound on finite-time convergence rate of* Q^+*-Whittle-LFA by leveraging and generalizing the machinery of nonlinear 2TSA [52]. The key is the choice of two step sizes (as characterized in Theorem 1) and a Lyapunov function given in (39). Though the main steps of our proof are motivated by [52], we need to characterize the specific requirements for our settings as aforementioned. Need to mention that we do not need the assumption that* h *and* g *are strongly monotone as in [52], and hence requires a re-derivation of the main results.*

VII. Numerical Results

In this section, we numerically evaluate the performance of our Q^+-Whittle-LFA algorithm using both synthetic and real traces.

A. Baselines and Experiment Setup

We compare Q^+-Whittle-LFA to existing learning based algorithms for wireless edge caching. In particular, we focus on both Q-learning based Whittle index policy for wireless edge caching (see Remark 2) such as *Q-learning Whittle Index Controller* (QWIC) [32], Q-Whittle[3] [9], *Whittle Index Q-learning* (WIQL) [51] and *Deep Threshold Optimal Policy Training* (DeepTOP) [38]; and existing learning based algorithms for wireless edge caching such as *Follow-the-Perturbed-Leader* (FTPL) [56], *Deep Q-Learning* (DQL) [57] and *Deep Actor-Critic* (DAC) [58]. We also compare these learning based algorithms to our Whittle index policy (see Section IV), which is provably asymptotically optimal when system parameters are known. For the above algorithms using neural networks, we consider two hidden layers with size (64, 32), with external memory size being 10,000 and batch size being 10. The discount factor is $\alpha=0.98$. The learning rates are initialized to be $\gamma_0=0.1$ and $\eta_0=0.01$, and are decayed by 1.1 every $1,000$ time steps. In LFA, we set $d=20$. In addition, we consider the normalized/standardized features, where features $\phi(s,a),\forall s,a$ are normalized or standardized to have certain properties like zero mean and unit variance.

[3]For Q-Whittle, we use the same reference function as indicated in Eq. (13) in [9], i.e., $I(Q):=\frac{1}{2|\mathcal{S}|}\sum_{s\in\mathcal{S}}Q(s,0)+Q(s,1)$, with $|\mathcal{S}|$ being the cardinality of $\mathcal{S}$.

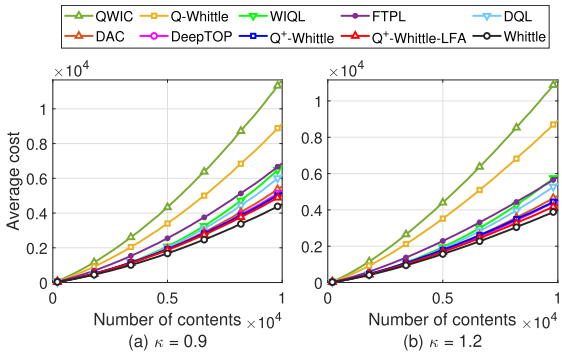

Fig. 2. Accumulated cost (latency) using synthetic traces.

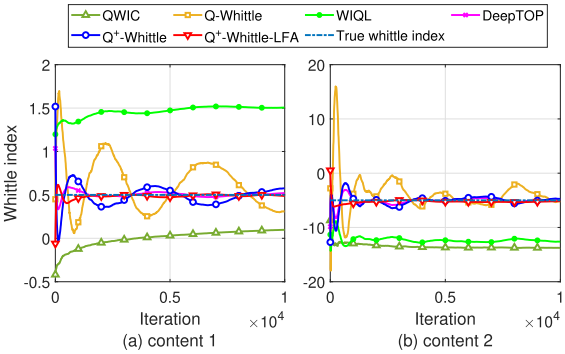

Fig. 3. Convergence in terms of iterations of Whittle index based Q-learning algorithms for two randomly selected contents.

B. Evaluation Using Synthetic Traces

We simulate a system with the number of distinct contents M ranging from 200 to $10,000$ with a step size of 200. In each case, content requests are drawn from a Zipf distribution with Zipf parameters κ of 0.9 and 1.2. As we consider a state-dependent delivery rate in our model (2), we set the "unit rate" $\nu=18$ with the true delivery rate of νSA, and the total number of requests varies across each M. The cache size is $B=M/10$.

Accumulated Cost (Latency). The accumulated costs of above learning based algorithms are presented in Figure 2, where we use the Monte Carlo simulation with $2,000$ independent trails. From Figure 2, it is clear that our Q^+-Whittle-LFA consistently outperforms its counterparts. In addition, WIQL outperforms QWIC and Q-Whittle, which is consistent with the observations made in [51]. Moreover, our Q^+-Whittle and Q^+-Whittle-LFA perform close to the Whittle index policy. This is due to the fact that both leverage the asymptotically optimal Whittle index policy to make decisions for wireless edge caching. Finally, we remark that Q^+-Whittle-LFA is much more computationally efficient compared to Q^+-Whittle, and Q-Whittle in [9], especially when the state space is large. This observation is further pronounced when we compare their convergence as illustrated below.

Convergence and Running Time. We demonstrate the convergence of Whittle index based Q-learning algorithms in terms of the number of iterations in Figure 3, and in terms of running time in Figure 4. The running time are obtained via averaging over 2,000 Monte Carlo runs of a single-threaded program on Ryzen 7 5800 × 3D desktop with 64 GB RAM. In both figures, we randomly draw two contents from the trace with Zipf parameter 0.9 due to the decoupled nature of our framework (see Section V). For ease of exposition, we only

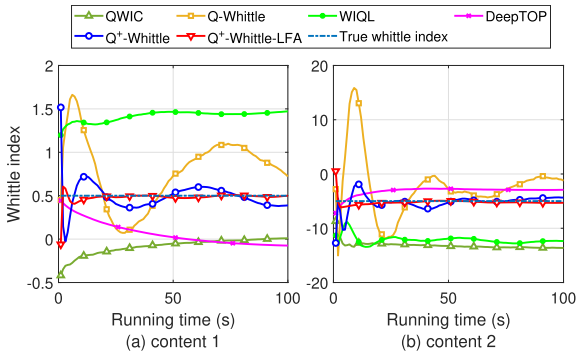

Fig. 4. Convergence in terms of running time of Whittle index based Q-learning algorithms for two randomly selected contents.

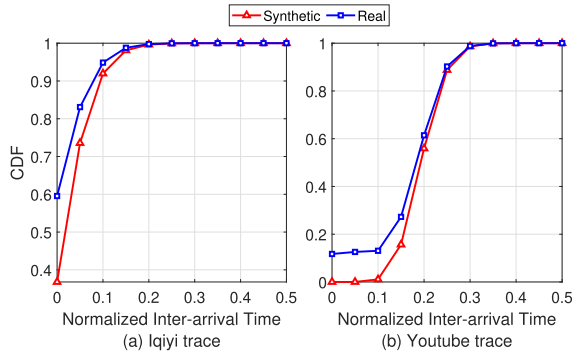

Fig. 5. Cumulative distribution function (CDF) of inter-arrival time for a randomly selected content from real traces.

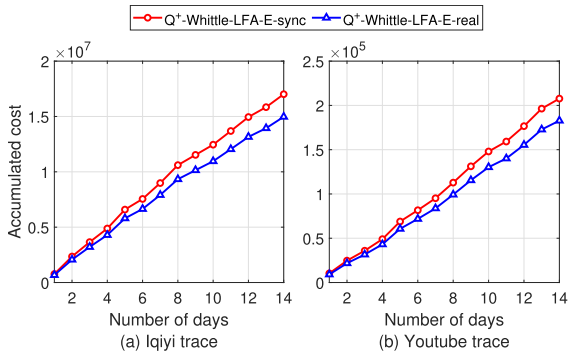

Fig. 6. Accumulated cost (latency) for theoretical and empirical results in real traces.

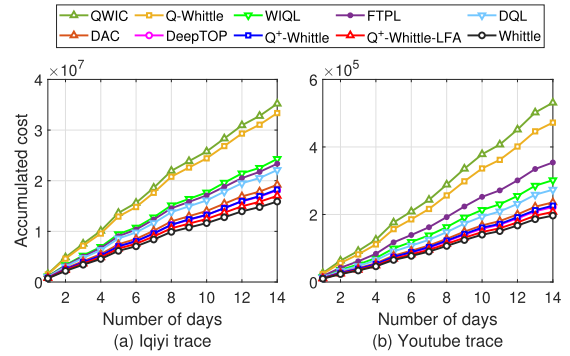

Fig. 7. Accumulated cost (latency) using real traces.

show results of Whittle indices of two states for these two particular contents.

We observe that the Whittle indices obtained by our $\mathtt{Q}^{+}\mathtt{-Whittle}$ and $\mathtt{Q}^{+}\mathtt{-Whittle-LFA}$ converge to the true Whittle indices, which are obtained under the assumption that system parameters are known. More importantly, $\mathtt{Q}^{+}\mathtt{-Whittle-LFA}$ converges much faster than $\mathtt{Q}^{+}\mathtt{-Whittle}$ both in terms of iterations and running time, as motivated earlier. In addition, $\mathtt{Q-Whittle}$ in [9] is provably convergent to the true Whittle index, however, the multi-timescale nature of $\mathtt{Q-Whittle}$ makes it converge slowly in practice (see discussions in Section II). As shown in Figure 3, $\mathtt{Q-Whittle}$ still cannot converge to the true Whittle indices after 10,000 iterations while our $\mathtt{Q}^{+}\mathtt{-Whittle-LFA}$ converges only after 1,000 iterations. We note that DeepTOP [38] also leverages a threshold policy to learn Whittle indices, which converges to the true Whittle indices in a smaller number of iterations as shown in Figure 3 but at the cost of a much larger running time as shown in Figure 4. This is due to its intrinsic nature of training a deep neural network in each iteration for making decisions. Finally, neither QWIC nor WIQL are guaranteed to converge to the true Whittle indices as observed in Figure 3. Similar observations hold for other contents in other traces, and hence are omitted here.

C. Evaluation Using Real Traces

We further evaluate $\mathtt{Q}^{+}\mathtt{-Whittle-LFA}$ using two real traces: (i) *Iqiyi* [59], which contains mobile video behaviors; and (ii) *YouTube* [60], which contains trace data about user requests for specific content collected from a campus network. For the Iqiyi (resp. YouTube) trace, there are more than 67 (resp. 0.6) million requests for more than 1.4 million (resp. 0.3) unique contents over a period of 335 (resp. 336) hours. We evaluate the accumulated cost over rough 14 days for each trace with a cache size of $B = 4,000$ (resp. $2,000$) for Iqiyi (resp. YouTube). We choose these values based on the observation of average number of active contents in the traces.[4]

Non-Poisson arrivals in real traces. We first show that the real traces do not have *strict* Poisson arrivals for any content. To that end, we generate a synthetic trace based on the real trace, where each content follows Poisson arrivals with the same average arrival rate as in the corresponding trace. We analyze the distribution of the inter-arrival times for the corresponding contents from the real and synthetic traces. In particular, Figure 5 presents the comparison for a randomly selected content from the trace. It is clearly from Figure 5 that they are visibly different. Similar trends hold for other contents in real traces considered in this paper, i.e., real traces do not have strict Poisson arrivals for any content.

Comparisons between theoretical and empirical results. We now show that despite the fact that real traces may not be strictly Poisson, our proposed solutions with Poisson assumption works well in practice. Specifically, we compare (i) *E-sync*: the empirical accumulated cost (latency) for the synthetic Poisson trace with same content arrival rates as in the real trace as described earlier; and (ii) *E-real*: empirical accumulated cost (latency) for the real trace. Figure 6 compares the curves of these two cases for our $\mathtt{Q}^{+}\mathtt{-Whittle-LFA}$. We observe that the theoretical and empirical accumulated costs only differ slightly. Similar observations hold for all baseline methods considered in this paper and hence are omitted here. We then present the theoretical accumulated cost comparisons of our $\mathtt{Q}^{+}\mathtt{-Whittle-LFA}$ and all baseline methods in Figure 7, from which we observe that

[4]A content is said to be active at time t if t lies between the first and the last requests for the content.

Q$^+$-Whittle-LFA outperforms all baselines. Importantly, despite the slight differences between theoretical and empirical results for each baseline method, the trend shown in Figure 7 also holds for comparisons when using the corresponding empirical accumulated cost and hence are omitted. Finally, we note that Q$^+$-Whittle-LFA can quickly learn the system dynamics and perform close to the Whittle index policy, which matches well with our theoretical results.

VIII. Conclusion

In this paper, we studied the content caching problem at the wireless edge with unreliable channels. Our goal is to derive an optimal policy for making content caching decisions so as to minimize the average content request latency from end users. We posed the problem in the form of a Markov decision process, and showed that the optimal policy has a simple threshold-structure and presented a closed form of Whittle indices for each content. We then developed a novel model-free reinforcement learning algorithm with linear function approximation, which is called Q$^+$-Whittle-LFA that can fully exploit the structure of the optimal policy when the system parameters are unknown. We mathematically characterized the performance of Q$^+$-Whittle-LFA and also numerically demonstrated its empirical performance.

Appendix A
Proof of Propositions in Section IV-B

A. Proof of Proposition 1

Proof: According to Assumption 1, we denote the smallest state with no preference[5] over active and passive actions as R, i.e, $Q^\alpha(R,1) = Q^\alpha(R,0)$. This implies the following two facts. First, for state $s < R$, the optimal action is 0, i.e.,

$$J^\alpha(R-1) = R - 1 - W + \alpha J^\alpha(R). \tag{44}$$

Second, equal preference over two actions at state R implies

$$R - W + \alpha J^\alpha(R+1) = R + \alpha P_{R,1} J^\alpha(R+1) + \alpha(1-P_{R,1}) J^\alpha(R-1),$$

from which we have

$$W = \alpha(1-P_{R,1})(J^\alpha(R+1) - J^\alpha(R-1)). \tag{45}$$

From (44), we establish the connection between value functions of states $R-1$ and $R+1$, i.e.,

$$J^\alpha(R-1) = R-1-W+\alpha(R - W + \alpha J^\alpha(R+1)). \tag{46}$$

Substituting (46) into (45), we have

$$J^\alpha(R+1) = \frac{\frac{W}{\alpha(1-P_{R,1})} + (R - 1 - W + \alpha(R-W))}{1-\alpha^2}.$$

As a result, $J^\alpha(R+1)$ can be updated as

$$\begin{cases} R+1-W+\alpha J^\alpha(R+2), & \text{if } a=0, \\ R+1+\alpha P_{R+1,1} J^\alpha(R+2) + \alpha(1-P_{R+1,1}) J^\alpha(R), & \text{o.w.} \end{cases}$$

[5]For an arbitrary subsidy W, it is possible that in states $1, 2, \ldots, R-1$, the passive action is the only optimal action, and in states $R, R+1, \ldots$, the active action is the only optimal action. Unfortunately, such a W value is not the Whittle index for any state. Since we are considering the Whittle index for each state s, which is defined as the value of W such that $Q(s,0) = Q(s,1)$. Hence, with loss of generality, we can always modify the current W value to make $Q(R,1) = Q(R,0)$.

In the following, we show that it is optimal to choose action 1 at state $R+1$. We first show that $a=0$ is not optimal by contradiction, and then verify that $a=1$ is optimal. Assume that the optimal action at state $R+1$ is $a=0$. Then, we have

$$\begin{aligned} W &\geq \alpha(1-P_{R+1,1})(J^\alpha(R+2) - J^\alpha(R)) \\ &= \alpha(1-P_{R+1,1})\Big(\frac{J^\alpha(R+1) - (R+1-W)}{\alpha} \\ &\quad - (R - W + \alpha J^\alpha(R+1))\Big) \\ &= \frac{1-P_{R+1,1}}{\alpha(1-P_{R,1})} W, \end{aligned} \tag{47}$$

where the inequality is due to the fact that optimal action is 0 at state $R+1$ and the last equality directly comes by plugging the closed-form expression of $J^\alpha(R+1)$. Since $\frac{1-P_{R+1,1}}{\alpha(1-P_{R,1})} > 1$, the inequality does not hold and it occurs an contradiction. This means that action 0 is not optimal for state $R+1$. We further verify that $a=1$ is optimal. When the optimal action at state $R+1$ is $a=1$, we have

$$\begin{aligned} W &\overset{(a)}{\leq} \alpha(1-P_{R+1,1})\Bigg(\frac{J^\alpha(R+1) - (R+1-W)}{\alpha} \\ &\quad - (R - W + \alpha J^\alpha(R+1))\Bigg) \\ &\overset{(b)}{\leq} \alpha(1-P_{R+1,1})\Bigg(\frac{J^\alpha(R+1) - (R+1) - \alpha(1-P_{R+1,1}) J^\alpha(R)}{\alpha P_{R+1,1}} \\ &\quad - (R - W + \alpha J^\alpha(R+1))\Bigg) \\ &= \alpha(1-P_{R+1,1})\Big(J^\alpha(R+2) - J^\alpha(R)\Big), \end{aligned} \tag{48}$$

where (a) directly follows from the contradiction implied by (47), and (b) holds as $\frac{J^\alpha(R+1) - (R+1) - \alpha(1-P_{R+1,1}) J^\alpha(R)}{\alpha P_{R+1,1}} \geq \frac{J^\alpha(R+1) - (R+1-W)}{\alpha}$. Thus the optimal action for $R+1$ is 1.

Following the same idea, the above results can be easily generalized to any state $s \geq R+1$, and hence we omit the detail here. To this end, the optimal policy of the discounted MDP (10) is of the threshold-type. ∎

B. Proof of Proposition 2

Proof: According to [61], the optimal expected total discounted latency $J_{\pi^*_\alpha}$ under the optimal policy π^*_α with discount factor α, and the optimal average latency J_{π^*} under the optimal policy π^* satisfy $\lim_{\alpha\to 1}(1-\alpha) J^\alpha_{\pi^*_\alpha}(s) = J_{\pi^*}(s), \forall s$. Since our action set is finite, there exists an optimal stationary policy for the average latency problem such that $\pi^*_\alpha \to \pi^*$ [61]. This shows that the optimal policy for (9) is of the threshold-type. ∎

C. Proof of Proposition 3

Proof: Since the optimal policy for (9) is of the threshold-type, for a given W, the optimal average cost under a threshold R satisfies

$$h(W) := \min_R \left\{ h^R(W) := \sum_{s=0}^{\infty} s\phi_R(s) - W \sum_{s=0}^{R} \phi_R(s) \right\}, \tag{49}$$

where $\phi_R(s)$ is the stationary probability of state s under the threshold policy $\pi = R$. It is easy to show that $h^R(W)$ is concave non-increasing in W since it is a lower envelope of linear non-increasing functions in W, i.e., $h^R(W) > h^R(W')$ if $W < W'$. Thus we can choose a larger threshold R when W increases to further decrease the total cost according to (49), i.e, $D(W) \subseteq D(W')$ when $W < W'$. ∎

D. Proof of Proposition 4

Proof: Following from the definition of Whittle index, the performance of a policy with threshold R equals to the performance of a policy with threshold $R+1$ [25], [26], i.e.,

$$\begin{aligned}&\mathbb{E}_R[s] - W(R)\mathbb{E}_R[\mathbb{1}_{\{A(s)=0\}}]\\ &\quad= \mathbb{E}_{R+1}[s] - W(R)\mathbb{E}_{R+1}[\mathbb{1}_{\{A(s)=0\}}],\end{aligned} \tag{50}$$

where the subscript denotes the fact that the associated quantities involve a threshold policy with the value of threshold equal to this value. Since the evolution of per-content is described by the transition kernel (a birth-and-death process) in (2), we have $\mathbb{E}_R[\mathbb{1}_{\{A(s)=0\}}] = \sum_{s=0}^{R}\phi_R(s)$. ∎

E. Proof of Proposition 5

Proof: Given the transition kernel in (2), the transition rate satisfies $q(S+1|S,0) = q(S+1|S,1) = \lambda$ and $q(S_1|S,0) = 0$ for $S \le R$, and $q(S-1|S,1) = \nu S$ for $S > R$. It is clear that $\forall S < R$ is transient because the state keeps increasing. Therefore, $\phi_R(S) = 0$, $\forall S < R$. Note that for threshold state R, the stationary probability satisfies

$$\phi_R(R) = \frac{\nu(R+1)}{\lambda+\nu(R+1)}\phi_R(R+1).$$

Based on the birth-and-death process, the stationary probabilities for states $R+l, \forall l = 2, \cdots, S_{max} - R$ satisfy

$$\phi_R(R+l)\frac{\lambda}{\lambda+\nu(R+l)} = \frac{\nu(R+l+1)}{\lambda+\nu(R+l+1)}\phi_R(R+l+1).$$

Therefore, we have the following relation

$$\phi_R(R+l) = \phi_R(R+1)\prod_{j=2}^{l}\frac{\lambda}{\lambda+\nu(R+j-1)}\frac{\lambda+\nu(R+j)}{\nu q(R+j)}.$$

Since $\phi_R(R) + \phi_R(R+1) + \cdots + \phi_R(S_{max}) = 1$, we have

$$\begin{aligned}&\phi_R(R+1)\\ &= 1/\Bigg(1 + \frac{\nu(R+1)}{\lambda+\nu(R+1)}\\ &\quad+ \sum_{l=2}^{S_{max}-R}\prod_{j=2}^{l}\frac{\lambda}{\lambda+\nu(R+j-1)}\frac{\lambda+\nu(R+j)}{\nu(R+j)}\Bigg).\end{aligned}$$

∎

Appendix B
Proof of Theorem 1

To prove Theorem 1, we need the following three key lemmas regarding the error terms defined in (38). First, we study the property of $\tilde{\theta}_n$.

Lemma 2: Consider the iterates $\{\theta_n\}$ and $\{W_n\}$ generated by (30)-(31). Under Assumptions 2-5, we have for all $n \ge 0$,

$$\begin{aligned}&\mathbb{E}\left[\|\tilde{\theta}_{n+1}\|^2|\mathcal{F}_n\right]\\ &\le \gamma_n^2\Lambda + (1 - 2\gamma_n\mu_1 + L_h^2\gamma_n^2)\|\tilde{\theta}_n\|^2\\ &\quad+ 2L_f^2L_g^2\eta_n^2\|\tilde{\theta}_n\|^2 + 2L_f^2L_g^2(L_f+1)^2\eta_n^2\|\tilde{W}_n\|^2\\ &\quad+ \left(L_f^2\gamma_n^2 + \frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2}{\gamma_n^2}\right)\|\tilde{\theta}_n\|^2\\ &\quad+ \frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2(L_f+1)^2}{\gamma_n^2}\|\tilde{W}_n\|^2.\end{aligned} \tag{51}$$

Proof: According to the definition in (38), we have

$$\begin{aligned}\tilde{\theta}_{n+1} &= \theta_{n+1} - f(W_{n+1})\\ &= \tilde{\theta}_n + \gamma_n h(\theta_n, W_n) + \gamma_n\xi_n + f(W_n) - f(W_{n+1}),\end{aligned}$$

which leads to

$$\begin{aligned}&\|\tilde{\theta}_{n+1}\|^2\\ &= \|\tilde{\theta}_n + \gamma_n h(\theta_n, W_n) + \gamma_n\xi_n + f(W_n) - f(W_{n+1})\|^2\\ &= \underbrace{\|\tilde{\theta}_n + \gamma_n h(\theta_n, W_n)\|^2}_{\text{Term}_1} + \underbrace{\|\gamma_n\xi_n + f(W_n) - f(W_{n+1})\|^2}_{\text{Term}_2}\\ &\quad+ \underbrace{2\left(\tilde{\theta}_n + \gamma_n h(\theta_n, W_n)\right)^T(f(W_n) - f(W_{n+1}))}_{\text{Term}_3}\\ &\quad+ \underbrace{2\gamma_n\left(\tilde{\theta}_n + \gamma_n h(\theta_n, W_n)\right)^T\xi_k}_{\text{Term}_4},\end{aligned} \tag{52}$$

where the second equality is due to the fact that $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\mathbf{x}^T\mathbf{y}$.

We next analyze the conditional expectation of each term in $\|\tilde{\theta}_{n+1}\|^2$ on $\mathcal{F}_n$. We first focus on Term_1.

$$\begin{aligned}&\mathbb{E}\Big[\text{Term}_1|\mathcal{F}_n\Big]\\ &= \|\tilde{\theta}_k\|^2 + 2\gamma_n\tilde{\theta}_n^\intercal h(\theta_n, W_n) + \|\gamma_n h(\theta_n, W_n)\|^2\\ &\overset{(a1)}{=} \|\tilde{\theta}_n\|^2 + 2\gamma_n\tilde{\theta}_n^\intercal h(\theta_n, W_n)\\ &\quad+ \gamma_n^2\|h(\theta_n, W_n) - h(f(W_n), W_n)\|^2\\ &\overset{(a2)}{\le} \|\tilde{\theta}_n\|^2 - 2\gamma_n\mu_1\|\tilde{\theta}_n\|^2 + L_h^2\gamma_n^2\|\tilde{\theta}_n\|^2,\end{aligned}$$

where (a1) follows from $h(f(W_n), W_n) = 0$, and (a2) holds due to the Lipschitz continuity of h in Assumption 2 and $\gamma_n\tilde{\theta}_n^T h(\theta_n, W_n) \le -\mu_1\|\tilde{\theta}_n\|^2$. For Term_2, we have

$$\begin{aligned}&\mathbb{E}\Big[\text{Term}_2|\mathcal{F}_n\Big]\\ &= \mathbb{E}[\|f(W_n) - f(W_{n+1}) + \gamma_n\xi_n\|^2|\mathcal{F}_n]\\ &\overset{(b1)}{=} \mathbb{E}[\|f(W_n) - f(W_{n+1})\|^2|\mathcal{F}_n] + \gamma_n^2\mathbb{E}[\|\xi_n\|^2|\mathcal{F}_n]\\ &\overset{(b2)}{\le} L_f^2\mathbb{E}[\|W_n - W_{n+1}\|^2|\mathcal{F}_n] + \gamma_n^2\Lambda\\ &= L_f^2\mathbb{E}[\|\eta_n g(\theta_n, W_n)\|^2|\mathcal{F}_n] + \gamma_n^2\Lambda\\ &= L_f^2\eta_n^2\|g(\theta_n, W_n)\|^2 + \gamma_n^2\Lambda\\ &\overset{(b3)}{\le} 2L_f^2\eta_n^2\|g(\theta_n, W_n) - g(f(W_n), W_n)\|^2 + \gamma_n^2\Lambda\\ &\quad+ 2L_f^2\eta_n^2\|g(f(W_n), W_n) - g(f(W(R)), W(R))\|^2\\ &\overset{(b4)}{\le} 2L_g^2L_f^2\eta_n^2\|\tilde{\theta}_n\|^2 + 2L_g^2L_f^2\eta_n^2\Big(\|f(W_n) - f(W(R))\|\\ &\quad+ \|W_n - W(R)\|\Big)^2 + \gamma_n^2\Lambda\\ &\overset{(b5)}{\le} 2L_f^2L_g^2\eta_n^2\|\tilde{\theta}_n\|^2 + 2L_f^2L_g^2(L_f+1)^2\eta_n^2\|\tilde{W}_n\|^2 + \gamma_n^2\Lambda,\end{aligned} \tag{53}$$

where (b1) is due to $\mathbb{E}[\xi_n|\mathcal{F}_n]=0$, (b2) is due to the Lipschitz continuity of f, and (b3) holds since $\|g(\theta_n,W_n)\|^2 \le 2\|g(\theta_n,W_n)-g(f(W_n),W_n)\|^2+2\|g(f(W_n),W_n)-g(f(W(R)),W(R))\|^2$ when $g(f(W_n),W_n)=0$, (b4) and (b5) hold because of the Lipschitz continuity of g and f. Next, we have the conditional expectation of Term_3 as

$$\begin{aligned}
&\mathbb{E}\Big[\text{Term}_3|\mathcal{F}_n\Big]\\
&\le 2\mathbb{E}\|\tilde{\theta}_n+\gamma_n h(\theta_n,W_n)\|\cdot\|f(W_n)-f(W_{n+1})\|\\
&\overset{(c1)}{\le} 2L_f\eta_n\|\tilde{\theta}_n+\gamma_n h(\theta_n,W_n)\|\cdot\|g(\theta_n,W_n))\|\\
&\le 2L_f\eta_n(1+L_h\gamma_n)\|\tilde{\theta}_n\|\left(L_g\|\tilde{\theta}_n\|+L_g(L_f+1)\|\tilde{W}_n\|\right)\\
&\overset{(c2)}{\le} L_f^2\gamma_n^2\|\tilde{\theta}_n\|^2+\frac{(1+L_h\gamma_n)^2\eta_n^2}{\gamma_n^2}\\
&\qquad\cdot\left(L_g\|\tilde{\theta}_n\|^2+L_g(L_f+1)\|\tilde{W}_n\|\right)^2\\
&\le\left(L_f^2\gamma_n^2+\frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2}{\gamma_n^2}\right)\|\tilde{\theta}_n\|^2\\
&\quad+\frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2(L_f+1)^2}{\gamma_n^2}\|\tilde{W}_n\|^2, \qquad (54)
\end{aligned}$$

where (c1) is due to the Lipschitz continuity of f and (c2) holds because $2\mathbf{x}^T\mathbf{y}\le\beta\|\mathbf{x}\|^2+1/\beta\|\mathbf{y}\|^2,\forall\beta>0$. Since $\mathbb{E}\Big[\text{Term}_4|\mathcal{F}_n\Big]=0$, combining all terms leads to the final expression in (51). ∎

Lemma 3: Consider the iterates $\{\theta_n\}$ and $\{W_n\}$ generated by (30)-(31). Under Assumptions 2-5, for any $n\ge 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{W}_{n+1}\|^2|\mathcal{F}_n\right]&\le\|\tilde{W}_n\|^2+2\eta_n^2L_g^2\|\tilde{Q}_n\|^2\\
&\quad+2\eta_n^2L_g^2(L_h+1)^2\|\tilde{W}_n\|^2. \qquad (55)
\end{aligned}$$

Proof: According to (38), we have $\tilde{W}_{n+1}=W_{n+1}-W(R)=\tilde{W}_n+\eta_n g(\theta_n,W_n)$, which leads to

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{W}_{n+1}\|^2|\mathcal{F}_n\right]\\
&=\|\tilde{W}_n\|^2+2\eta_n\tilde{W}_n^Tg(\theta_n,W_n)+\eta_n^2\|g(\theta_n,W_n)\|^2\\
&\overset{(d1)}{\le}\|\tilde{W}_n\|^2-2\eta_n\mu_2\|\tilde{W}_n\|^2+\eta_n^2\|g(\theta_n,W_n)|^2\\
&\overset{(d2)}{\le}\|\tilde{W}_n\|^2-2\eta_n\mu_2\|\tilde{W}_n\|^2\\
&\quad+2\eta_n^2L_g^2\|\tilde{\theta}_n\|^2+2\eta_n^2L_g^2(L_f+1)^2\|\tilde{W}_n\|^2, \qquad (56)
\end{aligned}$$

where (d1) is due to $2\eta_n\tilde{W}_n^T(g(\theta_n,W_n))\le-2\mu_2\|\tilde{W}_n\|^2$ and (d2) is due to (b3)-(b5). ∎

Lemma 4: Consider the iterates $\{\theta_n\}$ and $\{W_n\}$ generated by (30)-(31). Assume that $\gamma_n\le\min\left(\frac{1}{2\mu_1},\frac{2\mu_1}{L_h^2+L_f^2}\right)$, $\eta_n\le\min\left(\frac{1}{2\mu_2},\frac{\mu_2}{L_g^2(L_f+1)^2(L_f^2+1)}\right)$ and $\eta_n\ll\gamma_n$. Then under Assumptions 2-5, we have

$$\lim_{n\to\infty}\mathbb{E}\left[\|\tilde{\theta}_n\|^2+\|\tilde{W}_n\|^2|\mathcal{F}_n\right]\to 0 \ \textit{almost surely}.$$

Proof: Providing Lemma 2 and Lemma 3, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{\theta}_n\|^2+\|\tilde{W}_n\|^2|\mathcal{F}_n\right]\\
&\le\gamma_n^2\Lambda+(1-2\gamma_n\mu_1+L_h^2\gamma_n^2)\|\tilde{\theta}_n\|^2\\
&\quad+\left(2L_f^2L_g^2\eta_n^2\|\tilde{\theta}_n\|^2+2L_f^2L_g^2\eta_n^2(L_f+1)^2\|\tilde{W}_n\|^2\right)\\
&\quad+\left(L_f^2\gamma_n^2+\frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2}{\gamma_n^2}\right)\|\tilde{\theta}_n\|^2\\
&\quad+\frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2(L_f+1)^2}{\gamma_n^2}\|\tilde{W}_n\|^2\\
&\quad+(1-2\eta_n\mu_2)\|\tilde{W}_n\|^2+2\eta_n^2L_g^2\|\tilde{\theta}_n\|^2+2\eta_n^2L_g^2(L_f+1)^2\|\tilde{W}_n\|^2\\
&\le\gamma_n^2\Lambda+(1-2\gamma_n\mu_1)\|\tilde{\theta}_n\|^2+(1-2\eta_n\mu_2)\|\tilde{W}_n\|^2\\
&\quad+\Bigg(L_h^2\gamma_n^2+2L_f^2L_g^2\eta_n^2+L_f^2\gamma_n^2+\frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2}{\gamma_n^2}\\
&\qquad+2\eta_n^2L_g^2\Bigg)\|\tilde{\theta}_n\|^2\\
&\quad+2L_g^2(L_f+1)^2\left(L_f^2\eta_n^2+\eta_n^2+\frac{(1+L_h\gamma_n)^2\eta_n^2}{\gamma_n^2}\right)\|\tilde{W}_n\|^2.
\end{aligned}$$

Since $\eta_n\le\min\left(\frac{1}{2\mu_2},\frac{\mu_2}{L_g^2(L_f+1)^2(L_f^2+1)}\right)$, $\gamma_n\le\min\left(\frac{1}{2\mu_1},\frac{2\mu_1}{L_h^2+L_f^2}\right)$ and $\eta_n\ll\gamma_n$, we have

$$\begin{aligned}
-1\le D_1:&=-2\gamma_n\mu_1+\Bigg(L_h^2\gamma_n^2+2L_f^2L_g^2\eta_n^2+L_f^2\gamma_n^2\\
&\quad+\frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2}{\gamma_n^2}+2\eta_n^2L_g^2\Bigg)\le 0,\\
-1\le D_2:&=-2\eta_n\mu_2+2L_g^2(L_f+1)^2\\
&\quad\cdot\left(L_f^2\eta_n^2+\eta_n^2+\frac{(1+L_h\gamma_n)^2\eta_n^2}{\gamma_n^2}\right)\le 0.
\end{aligned}$$

Define $x_n=\min(D_1,D_2)$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{\theta}_n\|^2+\|\tilde{W}_n|^2|\mathcal{F}_n\right]\\
&\le\gamma_n^2\Lambda+(1+x_n)\left(\|\tilde{\theta}_n\|^2+\|\tilde{W}_n\|^2\right)\\
&=\prod_{t=0}^{n}(1+x_t)\left(\|\tilde{\theta}_0\|^2+\|\tilde{\lambda}_0\|^2\right)\\
&\quad+\left(1+\sum_{t=0}^{n}\prod_{\tau=0}^{t}(1+x_{n-\tau})\right)\gamma_n^2\Lambda.
\end{aligned}$$

Since $0\le 1+x_n\le 1,\forall n$ and $\lim_{n\to\infty}\gamma_n\to 0$, $\lim_{n\to\infty}\mathbb{E}\left[\|\tilde{\theta}_n\|^2+\|\tilde{W}_n\|^2|\mathcal{F}_n\right]\to 0$ *almost surely*. ∎

Now we are ready to prove Theorem 1. Providing Lemmas 2-4, if $\frac{\eta_n}{\alpha_n}$ is non-increasing, we have

$$\begin{aligned}
&\mathbb{E}\Big[M(\theta_{n+1},W_{n+1})\Big|\mathcal{F}_n\Big]\\
&\le\frac{\eta_n}{\gamma_n}\gamma_n^2\Lambda+\frac{\eta_n}{\gamma_n}(1-2\gamma_n\mu_1+L_h^2\gamma_n^2)\|\tilde{\theta}_n\|^2\\
&\quad+\frac{\eta_n}{\gamma_n}\left(2L_f^2L_g^2\eta_n^2\|\tilde{\theta}_n\|^2+2L_f^2L_g^2\eta_n^2(L_f+1)^2\|\tilde{W}_n\|^2\right)\\
&\quad+\frac{\eta_n}{\gamma_n}\left(L_f^2\gamma_n^2+\frac{2(1+L_h\gamma_n)^2\eta_n^2L_g^2}{\gamma_n^2}\right)\|\tilde{\theta}_n\|^2\\
&\quad+\frac{2(1+L_h\gamma_n)^2\eta_n^3L_g^2(L_f+1)^2}{\gamma_n^3}\|\tilde{W}_n\|^2\\
&\quad+(1-2\eta_n\mu_2)\|\tilde{W}_n\|^2+2\eta_n^2L_g^2\|\tilde{\theta}_n\|^2\\
&\quad+2\eta_n^2L_g^2(L_f+1)^2\|\tilde{W}_n\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(e1)}{\leq} \frac{\eta_n}{\gamma_n}\gamma_n^2\Lambda + \frac{\eta_n}{\gamma_n}(1-2\gamma_n\mu_1)\|\tilde{\theta}_n\|^2 + (1-2\eta_n\mu_2)\|\tilde{W}_n\|^2 \\
&\quad + \left((L_h^2+L_f^2)\gamma_n\eta_n + 2(L_f^2+1)L_g^2\frac{\eta_n^3}{\gamma_n} \right. \\
&\qquad \left. + \frac{2(1+L_h\gamma_n)^2\eta_n^3L_g^2}{\gamma_n^3}\right)\|\tilde{\theta}_n\|^2 \\
&\quad + 2L_g^2(L_f+1)^2\left(L_f^2\frac{\eta_n^3}{\gamma_n} + \gamma_n\eta_n + \frac{(1+L_h\gamma_n)^2\eta_n^3}{\gamma_n^3}\right)\|\tilde{W}_n\|^2 \\
&\overset{(e2)}{\leq} \max(1-2\gamma_n\mu_1, 1-2\eta_n\mu_2)\mathbb{E}\Big[M(\theta_{n+1}, W_{n+1})\Big|\mathcal{F}_n\Big] \\
&\quad + \frac{\eta_n}{\gamma_n}\gamma_n^2\Lambda + (L_h^2+L_f^2+2L_g^2(L_f+1)^2)\gamma_n\eta_n\Big(\|\tilde{\theta}_n\|^2 + \|\tilde{W}_n\|^2\Big) \\
&\quad + 2L_g^2(L_f+1)^2\big(L_f^2+(1+L_h\gamma_n)^2\big)\frac{\eta_n^3}{\gamma_n^3}\Big(\|\tilde{\theta}_n\|^2 + \|\tilde{W}_n\|^2\Big).
\end{aligned} \tag{57}
$$

Since $(n+1)^2\cdot\gamma_n\eta_n = \gamma_0\eta_0(n+1)^{1/3}$ and $(n+1)^2\cdot\frac{\eta_n^3}{\gamma_n^2} = \frac{\eta_0^2}{\gamma_0^2}(n+1)^{1/3}$, multiplying both sides of (57) with $(n+1)^2$, we have

$$
\begin{aligned}
&(n+1)^2\mathbb{E}\Big[M(\theta_{n+1}, W_{n+1})\Big|\mathcal{F}_n\Big] \\
&\leq (n+1)^{1/3}\gamma_0\eta_0\Lambda + n^2\mathbb{E}\Big[M(\theta_n, W_n)\Big|\mathcal{F}_n\Big] \\
&\quad + (L_h^2+L_f^2+2L_g^2(L_f+1)^2)\alpha_0\eta_0(n+1)^{1/3}\Big(\|\tilde{\theta}_n\|^2 + \|\tilde{W}_n\|^2\Big) \\
&\quad + 2L_g^2(L_f+1)^2\big(L_f^2+(1+L_h\gamma_n)^2\big)\frac{\eta_0^3}{\gamma_0^3}(n+1)^{\frac{1}{3}}\Big(\|\tilde{\theta}_n\|^2 + \|\tilde{W}_n\|^2\Big) \\
&\leq (n+1)^{1/3}\gamma_0\eta_0\Lambda + n^2\mathbb{E}\Big[M(\theta_n, W_n)\Big|\mathcal{F}_n\Big] \\
&\quad + (n+1)^{1/3}\left(C_1\left(\|\tilde{\theta}_0\|^2 + |\tilde{W}_0\|^2\right)\right),
\end{aligned} \tag{58}
$$

where $C_1 = (L_h^2+L_f^2+2L_g^2(L_f+1)^2)\alpha_0\eta_0 + 2L_g^2(L_f+1)^2\left(L_f^2+(1+L_h\alpha_0)^2\right)\frac{\eta_0^3}{\alpha_0^3}$. Summing (58) from time step 0 to time step n, we have

$$
\begin{aligned}
&(n+1)^2\mathbb{E}\Big[M(\theta_{n+1}, W_{n+1})\Big|\mathcal{F}_n\Big] \\
&\leq \mathbb{E}\Big[M(\theta_0, W_0)\Big] + (n+1)^{4/3}C_1(\|\tilde{\theta}_0\|^2 + \|\tilde{W}_0\|^2) \\
&\qquad + (n+1)^{4/3}\gamma_0\eta_0\Lambda.
\end{aligned} \tag{59}
$$

Finally, dividing both sides by $(n+1)^2$ yields the results in Theorem 1.

REFERENCES

[1] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[2] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 1994.

[3] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, pp. 287–298, Jan. 1988.

[4] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *J. Appl. Probab.*, vol. 27, no. 3, pp. 637–648, Sep. 1990.

[5] I. M. Verloop, "Asymptotically optimal priority policies for indexable and nonindexable restless bandits," *Ann. Appl. Probab.*, vol. 26, no. 4, pp. 1947–1995, Aug. 2016.

[6] J. Niño-Mora, "Dynamic priority allocation via restless bandit marginal productivity indices," *Top*, vol. 15, no. 2, pp. 161–198, Oct. 2007.

[7] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, no. 4, pp. 1–8, 2010.

[8] A. Gopalan and S. Mannor, "Thompson sampling for learning parameterized Markov decision processes," in *Proc. COLT*, 2015, pp. 861–898.

[9] K. E. Avrachenkov and V. S. Borkar, "Whittle index based Q-learning for restless bandits with average reward," *Automatica*, vol. 139, May 2022, Art. no. 110186.

[10] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, vol. 48. Berlin, Germany: Springer, 2009.

[11] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire, "The role of caching in future communication systems and networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1111–1125, Jun. 2018.

[12] K. Poularakis, G. Iosifidis, A. Argyriou, I. Koutsopoulos, and L. Tassiulas, "Distributed caching algorithms in the realm of layered video streaming," *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 757–770, Apr. 2019.

[13] B. Abolhassani, J. Tadrous, and A. Eryilmaz, "Achieving freshness in single/multi-user caching of dynamic content over the wireless edge," in *Proc. 18th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOPT)*, Jun. 2020, pp. 1–8.

[14] M. Dehghan, L. Massoulié, D. Towsley, D. S. Menasché, and Y. C. Tay, "A utility optimization approach to network cache design," *IEEE/ACM Trans. Netw.*, vol. 27, no. 3, pp. 1013–1027, Jun. 2019.

[15] N. K. Panigrahy, J. Li, and D. Towsley, "Hit rate vs. hit probability based cache utility maximization," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 45, no. 2, pp. 21–23, Oct. 2017.

[16] S. Ioannidis and E. Yeh, "Adaptive caching networks with optimality guarantees," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Sci.*, Jun. 2016.

[17] J. Li et al., "DR-cache: Distributed resilient caching with latency guarantees," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2018, pp. 441–449.

[18] N. Garg, M. Sellathurai, V. Bhatia, B. N. Bharath, and T. Ratnarajah, "Online content popularity prediction and learning in wireless edge caching," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1087–1100, Feb. 2020.

[19] T. Zhao, I.-H. Hou, S. Wang, and K. Chan, "Red/Led: An asymptotically optimal and scalable online algorithm for service caching at the edge," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1857–1870, Aug. 2018.

[20] A. Sadeghi, G. Wang, and G. B. Giannakis, "Deep reinforcement learning for adaptive caching in hierarchical content delivery networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 4, pp. 1024–1033, Dec. 2019.

[21] S. O. Somuyiwa, A. György, and D. Gündüz, "A reinforcement-learning approach to proactive caching in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1331–1344, Jun. 2018.

[22] F. Wang, F. Wang, J. Liu, R. Shea, and L. Sun, "Intelligent video caching at network edge: A multi-agent deep reinforcement learning approach," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Jul. 2020, pp. 2499–2508.

[23] G. Yan and J. Li, "RL-Bélády: A unified learning framework for content caching," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1009–1017.

[24] G. Yan, J. Li, and D. Towsley, "Learning from optimal caching for content delivery," in *Proc. 17th Int. Conf. Emerg. Netw. EXperiments Technol.*, Dec. 2021, pp. 344–358.

[25] M. Larrnaaga, U. Ayesta, and I. M. Verloop, "Dynamic control of birth-and-death restless bandits: Application to resource-allocation problems," *IEEE/ACM Trans. Netw.*, vol. 24, no. 6, pp. 3812–3825, Dec. 2016.

[26] M. Larrañaga, U. Ayesta, and I. M. Verloop, "Index policies for a multi-class queue with convex holding cost and abandonments," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst.*, Jun. 2014, pp. 125–137.

[27] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," in *Proc. IEEE 9th Annu. Conf. Structure Complex. Theory*, Jun. 1994, pp. 318–322.

[28] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1902–1916, Mar. 2013.

[29] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.

[30] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret bounds for restless Markov bandits," in *Proc. ALT*, 2012, pp. 214–228.

[31] Y. H. Jung and A. Tewari, "Regret bounds for Thompson sampling in episodic restless bandit problems," in *Proc. NIPS*, 2019, pp. 1–10.

[32] J. Fu, Y. Nazarathy, S. Moka, and P. G. Taylor, "Towards Q-learning the whittle index for restless bandits," in *Proc. Austral. New Zealand Control Conf. (ANZCC)*, Nov. 2019, pp. 249–254.
[33] S. Wang, L. Huang, and J. Lui, "Restless-UCB, an efficient and low-complexity algorithm for online restless bandits," in *Proc. NIPS*, 2020, pp. 11878–11889.
[34] F. Robledo, V. Borkar, U. Ayesta, and K. Avrachenkov, "QWI: Q-learning with whittle index," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 49, no. 2, pp. 47–50, Jan. 2022.
[35] F. Robledo, V. S. Borkar, U. Ayesta, and K. Avrachenkov, "Tabular and deep learning of whittle index," in *Proc. EWRL*, 2022, pp. 1–16.
[36] G. Xiong, J. Li, and R. Singh, "Reinforcement learning augmented asymptotically optimal index policies for finite-horizon restless bandits," in *Proc. AAAI*, 2022, pp. 8726–8734.
[37] G. Xiong, X. Qin, B. Li, R. Singh, and J. Li, "Index-aware reinforcement learning for adaptive video streaming at the wireless edge," in *Proc. ACM MobiHoc*, 2022, pp. 81–90.
[38] K. Nakhleh and I. Hou, "DeepTOP: Deep threshold-optimal policy for MDPs and RMABs," in *Proc. NIPS*, 2022, pp. 1–13.
[39] G. Xiong, S. Wang, G. Yan, and J. Li, "Reinforcement learning for dynamic dimensioning of cloud caches: A restless bandit approach," *IEEE/ACM Trans. Netw*, vol. 31, no. 5, pp. 2147–2161, Oct. 2023.
[40] G. Xiong, S. Wang, and J. Li, "Learning infinite-horizon average-reward restless multi-action bandits via index awareness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 17911–17925.
[41] X. Wei, J. Liu, Y. Wang, C. Tang, and Y. Hu, "Wireless edge caching based on content similarity in dynamic environments," *J. Syst. Archit.*, vol. 115, May 2021, Art. no. 102000.
[42] F. Baccelli and P. Brémaud, *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, vol. 26. Berlin, Germany: Springer, 2013.
[43] T. X. Vu, L. Lei, S. Vuppala, A. Kalantari, S. Chatzinotas, and B. Ottersten, "Latency minimization for content delivery networks with wireless edge caching," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.
[44] J. D. C. Little, "A proof for the queuing formula: L $= \lambda$W," *Oper. Res.*, vol. 9, no. 3, pp. 383–387, Jun. 1961.
[45] S. K. Singh, V. S. Borkar, and G. S. Kasbekar, "User association in dense mmWave networks as restless bandits," 2021, *arXiv:2107.09153*.
[46] Y.-P. Hsu, E. Modiano, and L. Duan, "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals," *IEEE Trans. Mobile Comput.*, vol. 19, no. 12, pp. 2903–2915, Dec. 2019.
[47] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, no. 2. Nashua, NH, USA: Athena Scientific Belmont, 1995.
[48] J. Abounadi, D. Bertsekas, and V. S. Borkar, "Learning algorithms for Markov decision processes with average cost," *SIAM J. Control Optim.*, vol. 40, no. 3, pp. 681–698, Jan. 2001.
[49] D. Blackwell, "Discrete dynamic programming," *Ann. Math. Statist.*, vol. 33, no. 2, pp. 719–726, Jun. 1962.
[50] C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain, "Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes," in *Proc. ICML*, 2020, pp. 1–11.
[51] A. Biswas, G. Aggarwal, P. Varakantham, and M. Tambe, "Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare," in *Proc. IJCAI*, 2021, pp. 1–10.
[52] T. T. Doan, "Nonlinear two-time-scale stochastic approximation: Convergence and finite-time performance," 2020, *arXiv:2011.01868*.
[53] T. T. Doan and J. Romberg, "Linear two-time-scale stochastic approximation a finite-time analysis," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2019, pp. 399–406.
[54] Z. Chen, S. Zhang, T. T. Doan, J.-P. Clarke, and S. Theja Maguluri, "Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning," 2019, *arXiv:1905.11425*.
[55] H. Gupta, R. Srikant, and L. Ying, "Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning," in *Proc. NIPS*, 2019, pp. 1–10.
[56] R. Bhattacharjee, S. Banerjee, and A. Sinha, "Fundamental limits on the regret of online network-caching," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 4, no. 2, pp. 1–31, 2020.
[57] P. Wu, J. Li, L. Shi, M. Ding, K. Cai, and F. Yang, "Dynamic content update for wireless edge caching via deep reinforcement learning," *IEEE Commun. Lett.*, vol. 23, no. 10, pp. 1773–1777, Oct. 2019.
[58] C. Zhong, M. C. Gursoy, and S. Velipasalar, "Deep reinforcement learning-based edge caching in wireless networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 1, pp. 48–61, Mar. 2020.
[59] G. Ma, Z. Wang, M. Zhang, J. Ye, M. Chen, and W. Zhu, "Understanding performance of edge content caching for mobile video streaming," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1076–1089, May 2017.
[60] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Watch global, cache local: YouTube network traffic at a campus network: Measurements and implications," *Proc. SPIE*, vol. 6818, pp. 35–47, Jan. 2008.
[61] S. A. Lippman, "Semi-Markov decision processes with unbounded rewards," *Manag. Sci.*, vol. 19, no. 7, pp. 717–731, Mar. 1973.

Guojun Xiong (Student Member, IEEE) received the B.S. degree in information science and technology from Sun Yat-sen University, China, in 2015, and the M.S. degree in electrical engineering and computer science from The University of Kansas, Lawrence, KS, USA, in 2020. He is currently pursuing the Ph.D. degree with Stony Brook University. His research interests include wireless communication, networking, optimization and control, and reinforcement learning.

Shufan Wang received the B.S. degree from Nanjing University, China, in June 2017, and the M.S. degree from Binghamton University in September 2022. He is currently pursuing the Ph.D. degree with Stony Brook University. His research interests include online algorithms and reinforcement learning.

Jian Li (Member, IEEE) received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in June 2012, and the Ph.D. degree in computer engineering from Texas A&M University, College Station, TX, USA, in December 2016. He is currently an Assistant Professor of data science with Stony Brook University. Before that, he was an Assistant Professor with Binghamton University, Binghamton, NY, USA, and a Post-Doctoral Researcher with the University of Massachusetts Amherst. His current research interests include the intersection of algorithms for reinforcement learning, federated learning, and stochastic optimization and control, with applications to next-generation networked systems.

Rahul Singh (Member, IEEE) received the B.Tech. degree in electrical engineering from Indian Institute of Technology at Kanpur, Kanpur, India, in 2009, the M.S. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 2011, and the Ph.D. degree from Texas A&M University, College Station, TX, USA, in 2015. He was a Post-Doctoral Scholar with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology, and The Ohio State University. He is currently an Assistant Professor with the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru, India. His research interests include networks and machine learning. His article was a Runner-Up for the Best Paper Award from ACM MobiHoc 2020.