Asymptotically Optimal Downlink Scheduling over Markovian Fading Channels

Wenzhuo Ouyang, Atilla Eryilmaz, and Ness B. Shroff

Abstract

We consider the scheduling problem in downlink wireless networks with heterogeneous, Markov-modulated, ON/OFF channels. It is well-known that the performance of scheduling over fading channels heavily depends on the accuracy of the available Channel State Information (CSI), which is costly to acquire. Thus, we consider the CSI acquisition via a practical ARQ-based feedback mechanism whereby channel states are revealed at the end of only scheduled users' transmissions. In the assumed presence of temporally-correlated channel evolutions, the desired scheduler must optimally balance the *exploitation-exploration trade-off*, whereby it schedules transmissions both to exploit those channels with up-to-date CSI and to explore the current state of those with outdated CSI.

In earlier works, Whittle's Index Policy had been suggested as a low-complexity and high-performance solution to this problem. However, analyzing its performance in the typical scenario of statistically heterogeneous channel state processes has remained elusive and challenging, mainly because of the highly-coupled and complex dynamics it possesses. In this work, we overcome these difficulties to rigorously establish the asymptotic optimality properties of Whittle's Index Policy in the limiting regime of many users. More specifically: (1) we prove the *local optimality* of Whittle's Index Policy, provided that the initial state of the system is within a certain neighborhood of a carefully selected state; (2) we then establish the *global optimality* of Whittle's Index Policy under a recurrence assumption that is verified numerically for the problem at hand. These results establish, for the first time to the best of our knowledge, that Whittle's Index Policy possesses analytically provable optimality characteristics for scheduling over heterogeneous and temporally-correlated channels.

I. Introduction

Channel fluctuation is an intrinsic characteristic of wireless communications. Such a variation calls for allocation of the wireless resources in a dynamic manner, leading to the classic *opportunistic scheduling principle* (e.g., [1], [2]). Under the assumption that the instantaneous channel state information (CSI) is fully available to the scheduler, many efficient opportunistic scheduling algorithms (e.g., [3]-[5]) have been proposed and extensively studied.

More recent works have focused on designing scheduling algorithms under imperfect CSI, where the channel state is modeled as independent and identically distributed (*i.i.d.*) processes across time (e.g., [6], [7]). On the other hand, although the *i.i.d.* channel model brings ease of analysis, it fails to capture the time-correlation of the fading channels [8]. Specifically, it fails to exploit the channel memory, which is a critical resource for making scheduling decisions. However, designing efficient scheduling schemes under time-correlated channels with imperfect CSI is a very challenging problem. The challenge is mainly because of the difficulty in making the classic 'exploitation versus exploration' trade-off, in which a scheduler needs to strike a balance between selecting the channels with up-to-date channel memory that guarantees high immediate gains, or to explore the channels with outdated CSI for more informed decisions and associated future throughput gains.

We consider the downlink scheduling problem where a base station transmits to the users within its transmission range, subject to scheduling constraints. To model the time correlations present over fading channels, we assume that wireless channels evolve as Markov-modulated ON/OFF processes. The channel state information is obtained from ARQ-based feedback, only *after* each scheduled transmission. Nevertheless, due to time correlation, the memory of the past channel state can be used to predict the current channel state *prior to* scheduling decision. Hence, channel memory should be intelligently exploited by the scheduler in order to achieve high throughput performance.

Wenzhuo Ouyang and Atilla Eryilmaz are with the Department of ECE, The Ohio State University (e-mails: ouyangw@ece.osu.edu, eryilmaz@ece.osu.edu). Ness B. Shroff holds a joint appointment in both the Department of ECE and the Department of CSE at The Ohio State University (e-mail: shroff@ece.osu.edu).

In a related work [9], a similar problem is considered under delayed CSI, where it is assumed that perfect CSI is available within a maximum delay, which is in turn smaller than the delay experienced by the ARQ feedback used for collision detection. These assumptions allow the scheduling decisions to be decoupled from CSI acquisition, which leads to the development of centralized as well as distributed schedulers. However, this approach does not use ARQ as a means of acquiring improved channel quality information. In contrast, in our setup the nature of ARQ feedback creates an implicit impact of scheduling decisions on the CSI feedback, which completely transforms the nature of the optimal scheduler design, and therefore requires a different approach. Under the scenario where all the channels have *identical Markov statistics*, round-robin-based algorithms (e.g., [10]-[12]) have been shown to possess optimality properties in throughput performance. However, the round-robin-based algorithms are no longer optimal in *asymmetric scenarios*, e.g., when different channels have different Markov transition statistics, as is naturally the case in typical heterogeneous conditions.

Under the asymmetric scenarios, our downlink scheduling problem is an example of the classic Restless Multi-armed Bandit Problem (RMBP) [13]. Low-complexity Whittle's Index Policies [13] for the downlink scheduling problem have been proposed in [14][15] based on RMBP theory. However, although Whittle's Index Policy can bring significant throughput gains by exploiting the channel memory [15], the analytical characterization of its performance under asymmetric scenarios is very challenging and prohibitively technical. This is because asymmetry leads to a sophisticated interplay of memory evolution among channels with heterogeneous characteristics, which brings a significant challenge to the analysis of Whittle's Index Policy not present in the perfectly symmetric scenario.

For RMBP problems under general scenarios, Whittle's Index Policy has been proven in [16] to be asymptotically optimal as the number of users grows, provided a non-trivial condition, known as Weber's condition, holds. Nonetheless, Weber's condition concerns the global convergence of a non-linear differential equation, which is extremely difficult to verify even numerically in our downlink scheduling scenario.

In this paper, we take significant steps in analyzing the optimality properties of Whittle's Index Policy for the downlink scheduling problem in the presence of channel heterogeneity. Specifically, our contributions are as follows.

- We apply the Whittle's index framework to our downlink scheduling problem and identify the optimal policy for the problem with a relaxed constraint in Section III. This policy, with carefully selected randomization, provides a performance upper bound to Whittle's Index Policy.
- We establish the local optimality of Whittle's Index Policy in the asymptotic regime when the number of users scales in Section V. Specifically, we show that the performance of the index policy can get arbitrarily close to that of the relaxed-constraint optimal policy, provided that the initial state of the system is within a certain neighborhood of a carefully selected state.
- Based on the local optimality result, under a numerically verifiable recurrence assumption, we then establish the global optimality of Whittle's Index Policy in the limiting regime of many users in Section VI.

To the best of our knowledge, our work is the first to give analytical characterization of Whittle's Index Policy for downlink scheduling under channel heterogeneity.

II. System Model and Problem Formulation

A. Downlink Wireless Channel Model

We consider a time-slotted, wireless downlink system with one base station and N users. The wireless channel $C_i[t]$ between base station and user i remains static within each time slot t and evolves stochastically across time slots, independently across users. We adopt the simplest non-trivial model of time-correlated fading channels by considering two-state ON/OFF channels, where the state space of channel i is $\mathcal{S}_i = \{0, 1\}$, with the value of each state representing the transmission rate a channel can support at the state.

One important component of our model is the inclusion of channel heterogeneity that the users will typically experience in real systems. Such asymmetry creates a significant challenge to the design and analysis of optimal scheduling schemes compared to perfectly symmetric channels. To avoid cumbersome notation and unessential technical complications, in this work we model channel asymmetry by considering only *two classes* of channel statistics. Specifically, for all the channels in class k, $k{=}1, 2$, their states evolve according to the same Markov

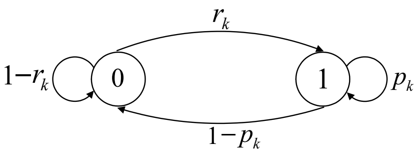

Fig. 1: Two state Markov chain model for channels in class k.

statistics. However, these characteristics differ between classes. The state transition of channels in class k is depicted in Fig. 1, represented by a 2×2 probability transition matrix,

$$\mathbb{P}_k = \begin{bmatrix} p_k & 1-p_k \\ r_k & 1-r_k \end{bmatrix},$$

where

$$p_k := \text{prob}\big(C_i[t]{=}1 \ \big| \ C_i[t{-}1]{=}1\big),$$
$$r_k := \text{prob}\big(C_i[t]{=}1 \ \big| \ C_i[t{-}1]{=}0\big).$$

for channel i in class k. The number of class k channels is $\gamma_k N$, $k \in \{1,2\}$ with γ_k being the *proportion* of channels in class k with respective to the total number N of channels.

We study the scenario where all the Markovian channels are positively correlated, i.e., $p_k > r_k$ for $k{=}1,2$. This assumption, which is commonly made in this domain (e.g., [12], [17]), means that the channel evolution has a positive auto-correlation. Hence, roughly speaking, the channel has a stronger potential to stay in its previous state than jumping to another, which is typical especially in slow fading environment. For ease of exposition, we shall exclude the trivial case when $r_k{=}0$ or $p_k{=}1$, $k=1,2$.

B. Scheduling Model – Belief Value Evolution

We assume that the base station can simultaneously transmit to at most $\alpha N{\in}\mathbb{Z}^+$ users in a time slot without interference, where $\alpha{\in}(0,1]$ stands for the maximum *fraction* of users that can be activated. For example, in a multi-channel communication model, α would correspond to the fraction of all users that can be simultaneously serviced in unit time. However, the scheduler does not know the exact channel state in the current slot when the scheduling decision is made. Instead, the scheduler maintains a *belief value* $\pi_i[t]$ for each channel i, which is defined as the probability of channel i being in the ON state at the beginning of slot t. The accurate channel state is revealed via ACK/NACK feedback from the scheduled users, only at the end of each time slot after the data is transmitted. This accurate channel state feedback is in turn used by the scheduler to update the belief values.

For user i in class k, $k{=}1,2$, let $a_i[t]{\in}\{0,1\}$ indicate whether the user is selected for transmission in slot t. Then, from the definition the belief values, $\pi_i[t]$ evolves as follows,

$$\pi_i[t{+}1]{=}\begin{cases} p_k, & \text{if } a_i[t]{=}1,\ C_i[t]{=}1, \\ r_k, & \text{if } a_i[t]{=}1,\ C_i[t]{=}0, \\ \pi_i[t]p_k{+}(1{-}\pi_i[t])r_k, & \text{if } a_i[t]{=}0. \end{cases} \tag{1}$$

In our setup, belief values are known to be sufficient statistics to represent the past scheduling decisions and feedback (e.g., [11], [18]). In the meanwhile, in our ON/OFF channel model, $\pi_i[t]$ also equals to the expected throughput contributed by channel i if it is scheduled in time slot t.

For a user in class k, $k{=}1,2$, we use $b^k_{c,l}$ to denote its belief value when the most recent observed channel was $c \in \{0,1\}$, and is l slots in the past. From the belief update rule (1), $b^k_{c,l}$ can be calculated as a function of $l{\geq}1$ as,

$$b^k_{0,l}{=}\frac{r_k{-}(p_k-r_k)^l r_k}{1+r_k-p_k},\ \ b^k_{1,l}{=}\frac{r_k{+}(1-p_k)(p_k-r_k)^l}{1+r_k-p_k}.$$

Fig. 2 illustrates the belief value update when a channel stays idle (i.e., $a_i{=}0$). It is clear that if the scheduler is never updated of the state of channel i (in class k), the belief value will converge to its stationary probability of being ON, denoted by the stationary belief value $b^k_s{:=}r_k/(1{+}r_k{-}p_k)$.

The vector $\vec{\boldsymbol{\pi}}[t]{=}(\pi_1[t],\cdots,\pi_N[t])$ denotes the belief values of all channels at the beginning of slot t. We use $\mathcal{B}_k$ to represent the set of the belief values for class k channels, where $\mathcal{B}_k{=}\{b^k_s, b^k_{c,l}, c{\in}\{0,1\}, l{\in}\mathbb{Z}^+\}$. We assume

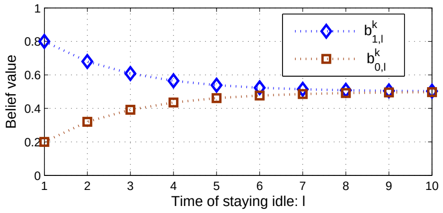

Fig. 2: Belief values update when staying idle, $p_k = 0.8$, $r_k = 0.2$.

that the system starts to operate from slot $t = 0$. At the beginning of slot 0, for each channel the scheduler has either observed its channel state before, or has never been updated of its channel state, i.e., with belief value b_s^k. It is then clear that, based on the belief update rule (1), $\pi_i[t] \in \mathcal{B}_k$ for all $t \geq 0$, i.e., each belief value $\pi_i[t]$ evolves over countably many states.

In the rest of the paper, we shall use 'belief value' and 'belief state' interchangeably.

C. Downlink Scheduling Problem – POMDP Formulation

We consider the broad class U of (possibly non-stationary) scheduling policies that makes a scheduling decision based on the history of observed channel states and scheduling actions. The downlink scheduling problem is then to identify a policy in U that maximizes the infinite horizon, *time average expected throughput*, subject to the constraint on the number of users selected at each time slot. Given the initial state $\vec{\boldsymbol{\pi}}[0]$, the problem is formulated as,

$$\max_{u \in U} \ \liminf_{T \to \infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \sum_{i=1}^{N} \pi_i[t] \cdot a_i^u[t] \Big| \vec{\boldsymbol{\pi}}[0]\Big] \tag{2}$$

$$s.t. \ \ \sum_{i=1}^{N} a_i^u[t] \leq \alpha N, \quad \forall t. \tag{3}$$

where the belief value $\pi_i[t]$ evolves according to rule (1) based on the scheduling decision $a_i^u[t]$ under policy u. Such an objective is standard in literature for Markov Decision Processes under the long term average reward criteria (e.g., [19]). Noting that since the scheduling decisions are made based on incomplete knowledge of channel states, this problem is a Partially Observable Markov Decision Process [18].

This problem is in fact an example of Restless Multiarmed Bandit Problem (RMBP) [13]. For a general RMBP, finding an optimal solution is PSPACE-hard [20]. However, for the downlink scheduling problem at hand, a low-complexity Whittle's Index Policy was proposed in [14][15] based on the RMBP theory that inherently exploits the channel memory when making scheduling decisions. For detailed descriptions of general RMBP and Whittle's Index Policy for downlink scheduling, please refer to [13]-[15].

For the downlink scheduling problem, we note that there is only limited analytical characterization of Whittle's Index Policy, which is restricted in perfectly symmetric scenarios where Whittle's Index Policy takes a special round-robin form [14]. In asymmetric cases, however, the scheduling decision no longer takes the form of round-robin, bringing sophisticated complications in belief value evolutions that are tightly coupled among channels, which significantly complicates the analysis. The main focus of this paper is to analytically characterize the performance of Whittle's Index Policy in the asymmetric case with two classes of channels.

III. Upper Bound on Achievable Throughput

We begin our analysis by characterizing an upper bound to the throughput performance of all feasible downlink scheduling policies that satisfies the constraint (3). The upper bound is obtained from a fictitious policy which is optimal for the downlink scheduling problem under a *relaxed constraint*.

Note here that such relaxation is also a crucial step in the study of the general RMBP problem. Yet, our analysis, being specific to the downlink scheduling problem, has its novelties, as we shall remark on later.

A. Average-Constrained Relaxed Scheduling Problem

We consider an associated relaxed problem of (2)-(3) that only requires an *average number* of users to be activated in the long run, defined as follows

$$\max_{u \in U} \ \liminf_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1}\sum_{i=1}^{N} \pi_i[t]\cdot a_i^u[t] \Big| \vec{\boldsymbol{\pi}}[0]\Big] \tag{4}$$

$$s.t. \ \ \limsup_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1}\sum_{i=1}^{N} a_i^u[t]\Big] \le \alpha N. \tag{5}$$

Note that, contrary to the stringent constraint (3), the relaxed constraint (5) allows the activation of more than α fraction of users in each time slot, provided the long term average fraction does not exceed α. Hence the optimal policy under this relaxed constraint, which we shall identify next, provides a throughput upper bound to any policy that satisfies the stringent constraint.

B. Optimal Policy for the Relaxed Problem

We remark that the relaxed problem is also an important component of Whittle's analysis of general RMBPs [13], in which an optimal policy for the relaxed problem is developed based on the *Whittle's index values*. Following the approach of classic RMBP framework [13], in our downlink scenario, we identify an optimal policy for the relaxed problem based on Whittle's indices.

Specifically, for channels in class k, the Whittle's index value $W_k(\pi)$ is assigned to each belief state $\pi \in \mathcal{B}_k$. These index values intuitively capture the exploitation and exploration value to be gained from scheduling the associated channel when its belief value is π. This characteristic of $W_k(\pi)$ is also illustrated in Section VII-B via numerical investigations. While these index value functions have been expressed in closed form in various cases (see [14][15]), the following two characteristics they possess are primarily significant for our analysis:

- $W_k(\pi)$ monotonically increases with $\pi \in \mathcal{B}_k$.
- $W_k(\pi) \in [0,1]$ for all $\pi \in \mathcal{B}_k$.

In the next proposition, we identify an index-based policy with *appropriate randomization* that is optimal for the relaxed constraint problem. This policy schedules each user based on its own belief value, independently from other users.

Proposition 1. *For the problem under relaxed constraint, there exists an optimal stationary policy ϕ^*, parameterized by the threshold ω^* and a randomization parameter $\rho^* {\in} (0,1]$, such that*

(i) Channel i in class k is scheduled if $W_k(\pi_i[t]) {>} \omega^*$, and stays idle if $W_k(\pi_i[t]) {<} \omega^*$. If $W_k(\pi_i[t]) {=} \omega^*$, it is scheduled with probability ρ^*.

(ii) The parameters ω^* and ρ^* are such that, under policy ϕ^*, the relaxed constraint (5) is strictly satisfied with equality.

Proof: This proof the proposition builds on the RMBP theory [13][14] along with optimization techniques. Details of the proof are given in Appendix A. ■

From now on, we shall denote ϕ^* as the '*Optimal Relaxed Policy*'. For technical purposes, we henceforth assume α is such that $\rho^* {\neq} 1$. Since each α value maps to a unique (ω^*, ρ^*) pair (see Appendix A), only countably many α values correspond to $\rho^* {=} 1$, i.e., achieved by deterministic policies. Therefore, the set of $\alpha {\in} (0,1]$ for which $\rho^* {\neq} 1$ has Lebesgue measure one.

Remarks:

1) Our work is the first to identify the specific form of the optimal policy for the relaxed problem in downlink scheduling. We identify in Proposition 1 that appropriate randomization is essential to guaranteeing the optimality. The randomization is important, because the deterministic policies are insufficient to guarantee optimality to general constrained Markov Decision Processes when both the reward and constraint are in the expected average form [19], and thus unable to provide a throughput upper bound.

2) Our objective function takes a very general form, it is not restricted to the family of stationary policies, nor does it require the existence of the limit (i.e., $\liminf \frac{1}{T}E[\cdot] = \lim \frac{1}{T}E[\cdot]$ in (2) and (4)), whereas the existence

of limits (with different forms) is assumed in previous literatures [13] [14] on Whittle's Index Policy. Such an extension not only requires a non-trivial amount of technical work, but also is important to prove optimality of the stationary Optimal Relaxed Policy over a larger space of possibly non-stationary control strategies.

C. Steady State Distribution of Belief Values

We next present the transition structure of the belief values under Optimal Relaxed Policy, captured in the following lemma. The structure will be critical in the development of our subsequent main results.

Lemma 1. *For each channel in class k, under the Optimal Relaxed Policy, the structure of belief value evolution depends on the threshold ω^* of policy.*
(i) If $\omega^* < W_k(b_s^k)$, then the belief value evolution of each class k channels is positive recurrent with a finite recurrent class.
(ii) If $\omega^* \geq W_k(b_s^k)$, the belief value evolution is transient. With probability 1, ultimately no channel in class k will transmit.

Proof: The proof of this lemma follows from the monotonic structure of belief evolution, as shown in Fig. 2. Details are included in Appendix F. ■

Thus, if $\omega^* \geq \max\{W_1(b_s^1), W_2(b_s^2)\}$, the above analysis reveals that ultimately no user transits, corresponding to the trivial case of $\alpha N = 0$. Also, if ω^* is between $W_1(b_s^1)$ and $W_2(b_s^2)$, the class with the smaller $W_k(b_s^k)$ will eventually transit into a passive mode, hence reducing the system to a well-understood scenario with a single class of channels [10][11]. Thus, here we focus on the heterogeneous case of $\omega^* < W_k(b_s^k), k=1,2$, where the steady-state belief value distribution exists for both classes under the Optimal Relaxed Policy.

D. Upper bound on achievable throughput

The throughput performance of Optimal Relaxed Policy provides an throughput upper bound for all policies under the stringent constraint. The value of such an upper bound clearly depends on the number of users in each class $\gamma_k N$, $k=1,2$, as well as the fraction α of users allowed for activation. Denoting $\boldsymbol{\gamma} = [\gamma_1, \gamma_2]$, we represent the time average expected throughput of the Optimal Relaxed Policy as $\upsilon^N(\boldsymbol{\gamma}, \alpha)$. The following lemma states that, as long as $\boldsymbol{\gamma}$ and α are given, the *per-user* throughput is independent of N.

Lemma 2. *Given $\boldsymbol{\gamma}$ and α, $\frac{\upsilon^N(\boldsymbol{\gamma},\alpha)}{N}$ is independent of N, denoted henceforth as $r(\boldsymbol{\gamma}, \alpha)$.*

Proof: The proof follows from showing that, when the number of users N grows, as long as the proportion of each class of channels stays the same and the fraction α of users activated does not change, the form of Optimal Relaxed Policy does not change. Since each user is scheduled independently, the throughput $\upsilon^N(\boldsymbol{\gamma}, \alpha)$ is proportional to N, establishing the lemma. Details are provided in Appendix C. ■

We hence refer to the $(\boldsymbol{\gamma}, \alpha)$ pair as '*system parameters*'. Therefore $Nr(\boldsymbol{\gamma}, \alpha)$ provides a throughput upper bound to any policy in the same system under the stringent constraint (3). Equivalently, $r(\boldsymbol{\gamma}, \alpha)$ provides a per-user throughput performance upper bound to all policies that satisfies the stringent constraint.

We next describe Whittle's Index Policy for the strictly-constrained problem (2)-(3), and later study the closeness of its performance to the upper bound established here.

IV. Whittle's Index Policy Description

In this section we formally introduce Whittle's Index Policy for solving the stringently-constrained downlink scheduling problem (2)-(3).

A. Whittle's Index Policy

The Optimal Relaxed Policy, along with the Whittle's index values, gives consistent ordering of belief values with respective to the indices. For instance, under the Optimal Relaxed Policy, if it is optimal to schedule one channel, it is then optimal to transmit to other channels with higher index values. So the Whittle's index value gives an intuitive order of how attractive the channel is for scheduling. This intuition leads to Whittle's Index Policy [14] under the stringent constraint on the maximum number of channels that can be scheduled.

Whittle's Index Policy: *At the beginning of each time slot, the channel i in class k is scheduled if its Whittle's index value $W_k(\pi_i)$ is within the top αN index values of all channels in that slot, with arbitrary tie-breaking while assuring a total αN channels being scheduled.*

Whittle's Index Policy is attractive because it has very low complexity, and it was observed via numerical investigations to yield significant throughput performance gains over the scheduling strategies that does not utilize channel memory [15]. The main focus of our work is to analytically understand the approximate or asymptotic optimality of Whittle's Index Policy in asymmetric scenarios.

B. Whittle's Index Policy over Truncated State Space

Recall from Section II that the belief values evolve over a countable state space, also note that if a channel is not scheduled for a long time, its belief value will get arbitrarily close to its stationary belief value. This motivates us to consider a truncated version of the belief value evolution whereby the belief value is set to its steady state if the corresponding channel is not scheduled for a large number, say τ, slots. This mild assumption facilitates more tractable performance analysis of the policy. Thus, if a class k user is not scheduled for τ time slots, its channel state history is entirely forgotten and its belief value will transit to the stationary belief value b_s^k, where the truncation τ is assumed to be very large.

Whittle's Index Policy is then implemented over the truncated belief state, which differs from the non-truncated case merely in the truncated belief value evolution. We believe that, the truncated scenario can provide arbitrarily close approximation to the original system when τ is large. More importantly, as we shall see in the following two sections, Whittle's Index Policy, implemented over the truncated belief state space, achieve asymptotically optimal performance as long as the truncation is sufficiently large.

V. Local Optimality of Whittle's Index policy

In this section, we study the optimality properties of Whittle's Index Policy for downlink scheduling, over a large truncated belief space. This result forms the basis for the subsequent global optimality result in Section VI. We start by introducing a state space over which the local optimality will be established.

A. System State Vector

We define the *system state* $\boldsymbol{Z}^N$ as a vector that represents the proportion of channels in each belief value, over the truncated space when the total number of users is N, i.e., $\boldsymbol{Z}^N = \left[\boldsymbol{Z}^{1,N}, \boldsymbol{Z}^{2,N}\right]$, with

$$\boldsymbol{Z}^{k,N} = [Z_{0,1}^{k,N}, \cdots, Z_{0,\tau}^{k,N}, Z_s^{k,N}, Z_{1,\tau}^{k,N}, \cdots, Z_{1,1}^{k,N}], k = 1, 2.$$

where $Z_{c,l}^{k,N}$ and $Z_s^{k,N}$ respectively denote the *proportion* of channels in the corresponding belief state $b_{c,l}^k$ and b_s^k, with respect to the total number of users N. Hence, each element of $\boldsymbol{Z}^N$ is a multiple of $1/N$ so that $\boldsymbol{Z}^N$ takes values in a lattice with mesh size $1/N$. Noting that the total number of users in each class does not change over time, for any N the system state $\boldsymbol{Z}^N[t] \in \mathcal{Z}$ where

$$\mathcal{Z} := \{\boldsymbol{Z}^N \geq 0 : Z_s^{k,N} + \sum_{c,l} Z_{c,l}^{k,N} = \gamma_k, \; k = 1, 2\}. \tag{6}$$

The system state vector $\boldsymbol{Z}^N[t]$ does not distinguish users with the same belief state, thus its dimension will not scale with N. Therefore, compared with $\vec{\boldsymbol{\pi}}[t]$, it provides a more convenient representation of the system belief state. Furthermore, $\boldsymbol{Z}^N[t]$ fully determines the instantaneous throughput gain in slot t under both Whittle's Index Policy and the Optimal Relaxed Policy (introduced in Proposition 1), because the instantaneous throughput gains under both policies are only determined by the distribution of the channels with different belief values, not their identities.

From Lemma 1 and the subsequent remarks, under the operation of the Optimal Relaxed Policy, the belief state evolution of each channel is positive recurrent with a steady-state distribution. The following lemma also establishes the independence of this steady-state distribution from N, and defines a useful parameter for future use.

Lemma 3. *Given the system parameters $(\boldsymbol{\gamma}, \alpha)$, the system state vector $\boldsymbol{Z}^N[t]$ under the Optimal Relaxed Policy converges in distribution to a random vector, denoted as $\boldsymbol{Z}^N[\infty]$. The distribution of $\boldsymbol{Z}^N[\infty]$ is independent of N with its mean denoted as*

$$\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}} := E\big[\boldsymbol{Z}^N[\infty]\big].$$

Proof: This lemma follows from a similar principle to the one we established in Lemma 2. For details, please refer Appendix D. ■

It is easy to see that $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}} \in \mathcal{Z}$ and the form of $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}$ fully determines the time average throughput of the Optimal Relaxed Policy. Therefore, the vector $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}$ provides an important benchmark for our asymptotic analysis. If, in the long run under Whittle's Index Policy, the system state $\boldsymbol{Z}^N[t]$ stays close to $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}$, it indicates that Whittle's Index Policy will have throughput performance close to that of the Optimal Relaxed Policy – the throughput upper bound. To capture the closeness, we define the δ neighborhood of $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}$ as

$$\Omega_\delta(\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}) = \{\boldsymbol{Z} \in \mathcal{Z} : ||\boldsymbol{Z} - \vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}|| \le \delta\}, \tag{7}$$

for $\delta > 0$, where $||\cdot||$ stands for Euclidean distance. We are now ready to state and prove our first main result regarding a form of local optimality of Whittle's Index Policy.

B. Local Optimality of Whittle's Index Policy

Under the system parameters $(\boldsymbol{\gamma}, \alpha)$, we let $R^N_T(\boldsymbol{\gamma}, \alpha, \boldsymbol{x})$ represent the time average throughput obtained over the time duration $0 \le t < T$ under Whittle's Index Policy, conditioned on the initial system state $\boldsymbol{Z}^N[0] = \boldsymbol{x}$, i.e.,

$$R^N_T(\boldsymbol{\gamma}, \alpha, \boldsymbol{x}) := \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \sum_{i=1}^{N} \pi_i[t] a_i^{ind}[t] \Big| \boldsymbol{Z}^N[0] = \boldsymbol{x}\Big],$$

where $(a_i^{ind}[t])_i$ denotes the scheduling decision vector made by Whittle's Index Policy at time t.

Recall from Lemma 2 that $r(\boldsymbol{\gamma}, \alpha)$ denotes the per-user throughput under the Optimal Relaxed Policy, which serves as an upper bound on Whittle's Index Policy performance. The next proposition characterizes the local convergence property of Whittle's Index Policy performance to $r(\boldsymbol{\gamma}, \alpha)$.

Proposition 2. *Under the system parameters $(\boldsymbol{\gamma}, \alpha)$, there exists a $\delta > 0$ neighborhood $\Omega_\delta(\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}})$ of $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}$ such that, if the initial system state $\boldsymbol{x}$ is within $\Omega_\delta(\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}})$, then*

$$\lim_{T\to\infty} \lim_{m\to\infty} \frac{R^{N_m}_T(\boldsymbol{\gamma}, \alpha, \boldsymbol{x})}{N_m} = r(\boldsymbol{\gamma}, \alpha).$$

where $\{N_m\}_m$ is any increasing sequence of positive integers with αN_m, $\gamma_k N_m \in \mathbb{Z}^+$, for $k = 1, 2$ and all m.

Proof Outline: Here, we give a high level description of the proof for an intuitive understanding, and refer the reader to Appendix E for the rigorous derivation.

• We start by defining a fluid approximation, whereby the discrete-time evolution of $\boldsymbol{Z}^N[t]$ under Whittle's Index Policy is modeled as a deterministic vector $\boldsymbol{z}[t] \in \mathcal{Z}$ that evolves in discrete time over $\mathcal{Z}$ and is independent of N. Under this fluid approximation, the users are no longer unsplittable entities so that the state space of $\boldsymbol{z}[t]$ is no longer restricted to a lattice as it was for $\boldsymbol{Z}^N[t]$. Also, the fluid approximation $\boldsymbol{z}[t]$ evolves in a deterministic manner, in contrast to the stochastic transition of $\boldsymbol{Z}^N[t]$. The evolution of $\boldsymbol{z}[t]$ is defined by a difference equation as a function of the *expected* state change of $\boldsymbol{Z}^N[t]$ under Whittle's Index Policy as follows

$$\boldsymbol{z}[t+1] - \boldsymbol{z}[t]\Big|_{\boldsymbol{z}[t]=\boldsymbol{z}} = E\Big[\boldsymbol{Z}^N[t+1] - \boldsymbol{Z}^N[t] \Big| \boldsymbol{Z}^N[t] = \boldsymbol{z}\Big], \tag{8}$$

where N is any integer for which $\boldsymbol{z}$ is a feasible state.

• We then establish local convergence of the fluid approximation model when $\boldsymbol{z}[0]$ is within a small enough δ neighborhood $\Omega_\delta(\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}})$ of $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}$. We show the convergence by first noting that the differential equation (8) is linear within a wider convex region than $\Omega_\delta(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$. Within this region, we obtain a closed form expression of the right hand side of (8), which enables us to investigate the eigenvalue structure of the linear differential equation. We show that each eigenvalue λ satisfies $|\lambda| < 1$ and apply standard linear system theory to establish the local convergence.

• We then connect the fluid approximation model $\boldsymbol{z}[t]$ to the discrete-time stochastic system state $\boldsymbol{Z}^N[t]$ by using a discrete-time extension of Kurtz's Theorem, which can be interpreted as an extension of the strong law of large numbers to random processes (see [21]). Essentially, it states that, over any finite time duration $[0, T]$, the actual system evolution $\boldsymbol{Z}^N[t]$ can be made arbitrarily close to the above fluid approximation $\boldsymbol{z}[t]$ by increasing the number of channels N sufficiently, with exponential convergence rate.

• The previous convergence result, together with the local convergence result of the fluid evolution $\boldsymbol{z}[t]$ to $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha$, enables us to establish the local convergence of the system state $\boldsymbol{Z}^N[t]$ to $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha$ as the number of users N grows, provided that the initial state $\boldsymbol{Z}^N[0] \in \Omega_\delta(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$. Hence the system state under Whittle's Index Policy will stay close (in a probabilistic sense) to the expectation $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha$ of the system state under the Optimal Relaxed Policy, which, in turn, indicates that the throughput performance of Whittle's Index Policy will approach the throughput upper bound $r(\boldsymbol{\gamma}, \alpha)$, as expressed in the proposition.

We again emphasize that the technical details of the outlined steps are fairly intricate and are moved to Appendix E. ■

Proposition 2 illustrates an interesting local optimality property of Whittle's Index Policy as the number of users N and the time horizon T increases while the system parameters $(\boldsymbol{\gamma}, \alpha)$ stay the same. It indicates that, under Whittle's Index Policy, as long as the initial state $\boldsymbol{Z}^N[0]$ is close enough to $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha$, the average per-user throughput over any finite time duration will get arbitrarily close to the Optimal Relaxed Policy performance as the number of users scales.

Remark: We note that the sequence $\{N_m\}_m$ is used to guarantee that the number of channels in each class, as well as the number of scheduled users, take integer values. In fact, our result can be generalized to all N by slightly perturbing $\boldsymbol{\gamma}$ and α as a function of N but assuring their limits are well-defined.

VI. Global Optimality of Whittle's Index Policy

The above local optimality result heavily relies on the initial state $\boldsymbol{Z}^N[0]$ being close to $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha$, which is difficult to guarantee. In this section, we study the global optimality of the infinite horizon throughput performance of Whittle's Index Policy starting from any initial state. We begin our analysis by presenting the recurrence structure of the system state.

Lemma 4. *Under system parameters $(\boldsymbol{\gamma}, \alpha)$, for any $\epsilon > 0$, if the number of users N is large enough,*
(i) The system state $\boldsymbol{Z}^N[t]$ evolves as an aperiodic Markov chain, in a state space that contains only one recurrent class.
(ii) There exists at least one recurrent state within the ϵ neighborhood $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$ of $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha$.

Proof: We prove this lemma by constructing probability paths from any state to the neighborhood $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$. Details of the proof are included in Appendix F. ■

This lemma states that $\boldsymbol{Z}^N[t]$ will ultimately enter any small neighborhood of $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha$ when N is large enough. Together with Proposition 2, this result shows promise for establishing the global asymptotic optimality of Whittle's Index Policy. This is plausible because once $\boldsymbol{Z}^N[t]$ enters $\Omega_\delta(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$, the performance of Whittle's Index Policy *afterwards* can get very close to its upper bound as N scales, as established in Proposition 2. However, since we consider the infinite horizon time average throughput, this argument would break down if the time it takes for $\boldsymbol{Z}^N[t]$ to enter $\Omega_\delta(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$ also scales up with N. This observation motivates us to introduce a useful assumption, which will later be justified (in Section VII-A) via numerical studies.

Assumption Ψ: For each $\epsilon{>}0$, let $\Gamma_{\boldsymbol{x}}^N(\epsilon)$ represent the first time of reaching $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$ starting from $\boldsymbol{Z}^N[0] = \boldsymbol{x}$, i.e.,

$$\Gamma_{\boldsymbol{x}}^N(\epsilon) = \min\{t : \boldsymbol{Z}^N[t] \in \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha) \big| \boldsymbol{Z}^N[0] = \boldsymbol{x}\}.$$

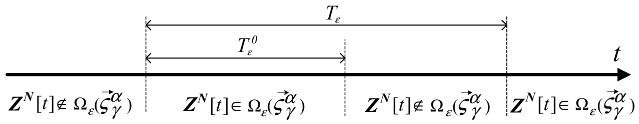
T_ϵ
T_ϵ^0
t
$\boldsymbol{Z}^N[t]\notin\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})$
$\boldsymbol{Z}^N[t]\in\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})$
$\boldsymbol{Z}^N[t]\notin\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})$
$\boldsymbol{Z}^N[t]\in\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})$

Fig. 3: Transition behavior of $\boldsymbol{Z}^N[t]$ in steady state.

Then, we assume that the expected time of reaching $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})$ is bounded uniformly over N and $\boldsymbol{x}$, i.e., there exists $M_\epsilon<\infty$ such that $E\big[\Gamma_{\boldsymbol{x}}^N(\epsilon)\big]\leq M_\epsilon$ for all N and $\boldsymbol{x}$.

Since for each N, $\boldsymbol{Z}^N[t]$ under Whittle's Index Policy is recurrent and aperiodic with a finite state space, there exists a steady-state distribution associated with $\boldsymbol{Z}^N[t]$. As before, we use $\boldsymbol{Z}^N[\infty]$ to denote the associated limiting random vector. The next lemma establishes that, under Assumption Ψ, the distribution of $\boldsymbol{Z}^N[\infty]$ approaches a point-mass at $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha}$ as N scales. Here, again, the sequence $\{N_m\}_m$ is defined in the same way as in Proposition 2.

Lemma 5. *Under Assumption Ψ and system parameters $(\boldsymbol{\gamma},\alpha)$, for any $\epsilon>0$, the steady state probability of $\boldsymbol{Z}^N[t]$ under Whittle's Index Policy satisfies*

$$\lim_{m\to\infty}P\big(\boldsymbol{Z}^{N_m}[\infty]\in\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\big)=1.$$

Proof: The proof utilizes Theorem 6.89 from [21], which builds on the following arguments.

Note that $\epsilon>0$ can be selected to be small enough for the following argument. As depicted in Fig. 3, we let T_ϵ be a random variable denoting, in steady state, the time duration between *consecutive* hitting times into the neighborhood $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})$ from outside of the neighborhood. Let T_ϵ^0 denote the time duration from the time $\boldsymbol{Z}^N[t]$ enters the neighborhood $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})$ from outside until the time it leaves. Hence, the expected proportion of time that $\boldsymbol{Z}^N[t]$ stays outside this neighborhood is $(E[T_\epsilon]-E[T_\epsilon^0])/E[T_\epsilon]$.

We know that the numerator $E[T_\epsilon]-E[T_\epsilon^0]$ is uniformly bounded for all N due to Assumption Ψ. However, as N increases, it is more likely for $\boldsymbol{Z}^N[t]$ to stay within the neighborhood for a long time before exiting it (based on the convergence of fluid approximation model and Kurtz's Theorem in the proof of Proposition 2). Thus, $E[T_\epsilon^0]$, and hence the denominator $E[T_\epsilon]$, grow to infinity as N scales. Therefore, the expected proportion of time spent outside $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})$ vanishes as N scales up, which leads to the statement of the lemma. Details of the proof can be found in Appendix G. ■

Under Whittle's Index Policy with system parameters $(\boldsymbol{\gamma},\alpha)$, we let $R_{\boldsymbol{x}}^N(\boldsymbol{\gamma},\alpha)$ be the achieved infinite horizon, time average throughput, conditioned on the initial system state $\boldsymbol{Z}^N[0]{=}\boldsymbol{x}$, i.e.,

$$R_{\boldsymbol{x}}^N(\boldsymbol{\gamma},\alpha):=\lim_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\sum_{i=1}^{N}\pi_i[t]a_i^{ind}[t]\Big|\boldsymbol{Z}^N[0]=\boldsymbol{x}\Big].$$

From Lemma 5 we know that, in steady-state, the system state $\boldsymbol{Z}^{N_m}[\infty]$ is increasingly concentrated around $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha}$ as m increases, regardless of the initial state $\boldsymbol{x}$. We build on this to establish the global asymptotical optimality of Whittle's Index Policy.

Proposition 3. *Under Assumption Ψ, for any initial system state $\boldsymbol{x}$, we have*

$$\lim_{m\to\infty}\frac{R_{\boldsymbol{x}}^{N_m}(\boldsymbol{\gamma},\alpha)}{N_m}=r(\boldsymbol{\gamma},\alpha).$$

Since $r(\boldsymbol{\gamma},\alpha)$ is an upper bound on the maximum achievable per-user throughput by any policy, this implies that Whittle's Index Policy is optimal in the many user regime.

Proof: We prove this result by decomposing $R_{\boldsymbol{x}}^N(\boldsymbol{\gamma},\alpha)$ as a summation of the expected throughput conditioned on whether the system state is within or outside an arbitrarily small ϵ neighborhood of $\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha}$. Since the latter has diminishing probability according to Lemma 5, the expected throughput of Whittle's Index Policy can get arbitrarily close to that of Optimal Relaxed Policy. Details of the proof are provided in Appendix H. ■

Remarks:

1) We would like to emphasize that the global optimality result is not a straight-forward extension of the local converge result by contrasting Proposition 2 and Proposition 3. Note that in Proposition 2, the time limit is

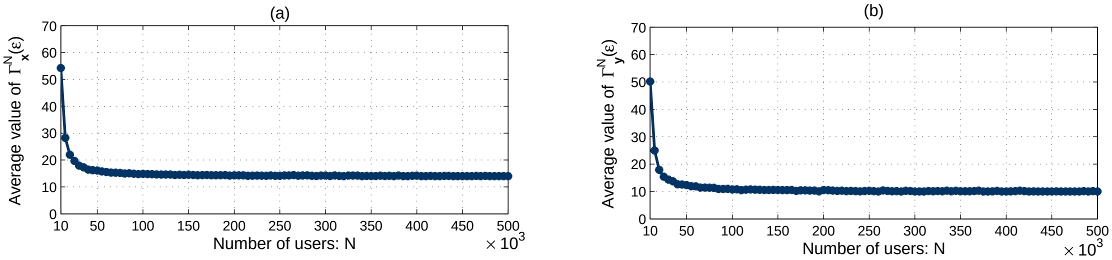

Fig. 4: Average time of hitting $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}^\alpha_{\boldsymbol{\gamma}})$. (a) $\boldsymbol{Z}^N[0]=\boldsymbol{x}$; (b) $\boldsymbol{Z}^N[0]=\boldsymbol{y}$.

outside the limit of the number of users N, where each convergence (with N) is with respective to a *fixed time duration*. However, the order of limit is switched in the global optimality result of Proposition 3, as it states the convergence with N *the infinite horizon* average throughput, which is much stronger and hence is much more challenging to prove.

2) We would like to contrast Assumption Ψ with Weber's condition [16]. For general RMBP problem, Weber's condition leads to the same global asymptotic optimality result. While confirming Weber's condition may be possible in very low-dimensional problems, in our downlink scheduling problem, this requires one to rule out the existence of both closed orbits and chaotic behavior of a high-dimensional non-linear differential equation, which is extremely difficult to check - even numerically. Assumption Ψ, on the other hand, takes a much simpler form, as it is defined over the actual stochastic system and is amenable to easy numerical verification, as is performed in Section VII-A.

VII. Numerical Results

A. Verification and Interpretation of Assumption Ψ

We start by numerically verifying Assumption Ψ. We consider the asymmetric scenario with two classes of channels with system parameters γ=$[0.45, 0.55]$, α=0.6, with p_1=0.9, r_1=0.45, p_2=0.8, r_2=0.3.

We next examine the change of the average hitting time $\Gamma^N_{\boldsymbol{x}}(\epsilon)$, while maintaining α and $\boldsymbol{\gamma}$.

We let $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{Z}$ be initial values of $\boldsymbol{Z}^N[0]$ that are selected to be two extreme points in the state space to exhibit the uniformity of $\Gamma^N_{\boldsymbol{x}}(\epsilon)$ to the initial state. Specifically, state $\boldsymbol{x}$ corresponds to the case when all the users have just observed their channels to be in OFF state, i.e., with belief value $b^k_{0,1}$, $k=1,2$. And $\boldsymbol{y}$ corresponds to the case when all users have no initial observation of their channels state history, i.e., with belief value b^k_s, $k=1,2$.

We examine the average value of hitting time $\Gamma^N_{\boldsymbol{x}}(\epsilon)$ and $\Gamma^N_{\boldsymbol{y}}(\epsilon)$ with a very small neighborhood ϵ=0.005, when the number of users N grows from 10×10^3 to 500×10^3. As indicated in Fig. 4, for both cases, the average time of hitting the ϵ neighborhood first decreases with N, and then *converges* and stays almost the same as N scales up. This is especially intriguing. The rationale behind this phenomenon is as follows. Under Whittle's Index Policy, a total number of αN users are activated at each time slot. Therefore, for relatively small number of users, the amount of probabilistic belief state transitions, as well as the amount of system states in the neighborhood, increases with N, leading to a higher chance of hitting the desired neighborhood $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}^\alpha_{\boldsymbol{\gamma}})$ and smaller value of hitting time. However, the belief update of each user contributes to the $1/N$ change of the system state $\boldsymbol{Z}^N[t]$, which decreases with N. Therefore, as N further increases, the *total amount of transitions* of the system state $\boldsymbol{Z}^N[t]$ due to channel state feedback is roughly $\alpha N \cdot 1/N = \alpha$, which is invariant of N. This result, along with many other numerical experiments we have conducted that lead to the same observation, gives verification to Assumption Ψ.

B. 'Exploitation versus Exploration' Trade-off

In this section, we demonstrate how the Whittle's index value captures the 'exploitation versus exploration' trade-off for our *asymmetric downlink scheduling problem*.

Consider two classes of ON/OFF fading channels with belief value evolutions plotted in Fig. 5(a). Note that both classes have the same stationary distribution $b^k_s=0.5$, $k\in\{1,2\}$ of being at ON state, but channels in class

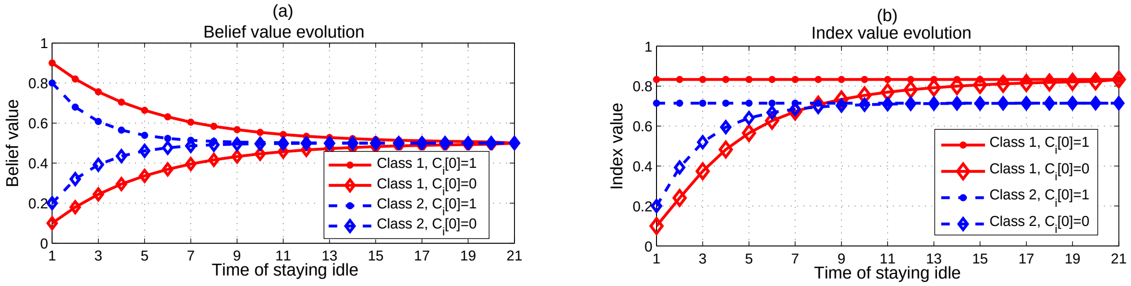

Fig. 5: The evolution of belief value and Whittle's index value. (a) Belief value evolution (b) Whittle's index value evolution.

1 has a higher degree of time correlation, i.e., fades slower, than channels in class 2 since $p_1 > p_2$ and $r_1 < r_2$. The corresponding Whittle index values of the two classes of channels are depicted in Fig. 5(b) as functions of the updated belief value starting from different initial states.

To understand the nature of Whittle's index value, we first consider the case when the channels in both classes are observed to be ON at time 0 and stay passive since then. As indicated in Fig. 5(a) the class 1 channel has a higher belief value than the class 2 channel, hence scheduling the class 1 channel gives a higher immediate throughput than scheduling the class 2 channel. Moreover, once a class 1 channel is scheduled, it is more likely to stay in ON state again, bringing high future gains. Accordingly, the index values in Fig. 5(b) when both state evolutions start from ON states capture that it is more attractive to schedule the class 1 channel because of the advantage in both exploitation and exploration.

On the other hand, when the scheduler has observed channels in both classes to be OFF at time 0, Fig. 5(a) shows that the class 2 channel has a higher belief value than the class 1 channel. However, although the Whittle's index value in Fig. 5(b) of class 2 channel is initially smaller than that of class 1 channel, after a certain amount of delay (around 8 slots in the figure) this order is switched, which is interpreted as follows: initially, since the class 1 channel has smaller belief value than that of the class 2 channel, it is more attractive to exploit the immediate gain brought by the class 2 channel. However, as the passive time grows, as indicated in Fig. 5(a), the difference between immediate gain of both classes diminishes. Then, it becomes more attractive to explore the class 1 channel because its longer memory can bring higher future gains if it turns out to be in ON state.

This investigation reveals the intricate nature of Whittle's index value in capturing the fundamental 'exploration versus exploitation' trade-off. In our scheduling problem with asymmetric channel statistics, such a property of Whittle's Index Policy turns out to be crucial in *achieving asymptotically optimal performance*.

VIII. Conclusion

In this paper, we studied the problem of downlink scheduling over ON/OFF Markovian fading channels in the presence of channel heterogeneity. We consider the scenario where instantaneous channel state information is not perfectly known at the scheduler, but is acquired via a practical ARQ-styled feedback after each scheduled transmission. We analytically characterized the performance of Whittle's Index Policy for downlink scheduling, and proved its local and global asymptotic optimality properties as the number of users scales. Specifically, provided that the initial system state is within a certain region, we established the local optimality of Whittle's Index Policy by investigating the evolution of the system belief state with a fluid approximation. We then established the global asymptotic optimality of Whittle's Index Policy under a recurrence condition, which is suitable for numerical verification. Our results establish that Whittle's Index Policy, which is attractive due to its low-complexity operation, also processes strong asymptotic optimality properties for scheduling over heterogeneous Markovian fading channels.

References

[1] R. Knopp, P. A. Humblet, "Information capacity and power control in single cell multiuser communications," in *IEEE ICC*, 1995.

[2] X. Liu, E. K. P. Chong, N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE JSAC*, 2001.

[3] L. Tassiulas, "Scheduling and performance limits of networks with constantly changing topology," *IEEE Trans. on Inform. Theory*, 1997.

[4] X. Lin, N. B. Shroff, “The impact of imperfect scheduling on cross-Layer congestion control in wireless networks,” *IEEE/ACM Trans. on Networking*, 2006

[5] A. Eryilmaz, R. Srikant, “Fair resource allocation in wireless networks using queue-length based scheduling and congestion control,” *IEEE/ACM Trans. on Networking,* 2007.

[6] M. J. Neely, “Max weight learning algorithms with application to scheduling in unknown environments,” *arXiv:0902.0630,* 2009.

[7] W. Ouyang, S. Murugesan, A. Eryilmaz, N. B. Shroff, “Scheduling with Rate Adaptation under Incomplete Knowledge of Channel/Estimator Statistics,” in *Allerton Conference,* 2010.

[8] D. Tse, P. Viswanath, “*Fundamentals of wireless communication,*” Cambridge University Press, 2005.

[9] L. Ying, S. Shakkottai, “On throughput optimality with delayed network-state information,” *IEEE Trans. on Information Theory,* 2011.

[10] Q. Zhao, B. Krishnamachari, K. Liu, “On myopic sensing for multichannel opportunistic access: Structure, optimality, and performance,” *IEEE Trans. on Wireless Communications,* 2008.

[11] S.H. Ahmad, M. Liu, T. Javidi, Q. Zhao, B. Krishnamachari, “Optimality of myopic sensing in multi-Channel opportunistic access,” *IEEE Trans. on Infomation Theory,* 2009.

[12] C. Li, M. J. Neely, “Exploiting channel memory for multi-user wireless scheduling without channel measurement: capacity regions and algorithms,” in *IEEE WiOpt* 2010.

[13] P. Whittle, “Restless Bandits: Activity allocation in a changing world,“ *Journal of Applied Probability,* 1988.

[14] K. Liu, Q. Zhao, “Indexability of restless bandit problems and optimality of Whittle’s index for dynamic multichannel access,” *IEEE Trans. on Information Theory,* 2008.

[15] W. Ouyang, S. Murugesan, A. Eryilmaz, N. Shroff, “Exploiting channel memory for joint estimation and scheduling in downlink networks,” in *IEEE INFOCOM,* 2011.

[16] R. Weber and G. Weiss, “On an Index Policy for Restless Bandits,” *Journal of Applied Probability,* vol. 27, no. 3, 1990.

[17] S. Murugesan, P. Schniter, N. B. Shroff, “Opportunistic scheduling using ARQ feedback in Multi-Cell Downlink,” in *Asilomar* 2010.

[18] E. J. Sondik, “*The optimal control of partially observable Markov Decision Processes,*” Ph.D. thesis, Stanford University, 1971.

[19] Eitan Altman, “*Constrained Markov Decision Processes,*” Chapman & Hall, 1999.

[20] C. Papadimitriou, J.N. Tsitsiklis “ The complexity of optimal queueing network control,” *Mathematics of Operation Research,* 1999.

[21] A. Shwartz, A. Weiss, “*Large deviation for performance analysis,*” Chapman & Hall, 1994.

[22] P. K. Dutta, “What do discounted optima converge to? A theory of discount rate asymptotics in economic models,“ *Journal of Economic Theory*, vol. 55, pp. 64-94, 1991.

[23] D. P. Bertsekas, “*Nonlinear programming, 2nd edition*“, Belmont: Athena Scientific.

[24] T. G. Kurtz, “Strong approximation theorems for density dependent Markov chains”, *Stochastic Processes and their Applications,* vol. 6, no. 3, pp. 223-240, 1978.

[25] R. A. Horn, “ *Matrix analysis,* ” Cambridge University Press, 1999.

[26] W. J. Rugh, “*Linear system theory,* ” Prentice Hall, 1996

Appendix A
Proof of Proposition 1

A. Lagrangian decomposition - Thresholdability

The constraint (5) can be written in an equivalent form that requires at least $(1-\alpha)N$ channels to be *passive* on average, i.e.,

$$\liminf_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\sum_{i=1}^{N}(1-a_i^u[t])\Big]\geq(1-\alpha)N. \tag{9}$$

Associating a Lagrange multiplier ω to the above constraint (9), we consider the following Lagrangian function $L(u,\omega)$ of the relaxed problem (4)-(5),

$$L(u,\omega)=\liminf_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\sum_{i=1}^{N}\pi_i[t]\cdot a_i^u[t]\Big|\vec{\boldsymbol{\pi}}[0]\Big]+\omega\cdot\liminf_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\sum_{i=1}^{N}(1-a_i^u[t])\Big|\vec{\boldsymbol{\pi}}[0]\Big]-\omega\cdot(1-\alpha)N. \tag{10}$$

The dual function $D(\omega)$ is defined as $D(\omega)=\max_{u\in U}L(u,\omega)$. The following lemma provides a useful upper bound to $D(\omega)$.

Lemma 6.

$$D(\omega)\leq\max_{u\in U}\ \sum_{i=1}^{N}\limsup_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\big[\pi_i[t]\cdot a_i^u[t]+\omega\cdot(1-a_i^u[t])\big]\Big|\vec{\boldsymbol{\pi}}[0]\Big]-\omega\cdot(1-\alpha)N. \tag{11}$$

Proof:

$$\begin{aligned} D(\omega) &\leq \max_{u\in U}\ \liminf_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\sum_{i=1}^{N}\big[\pi_i[t]\cdot a_i^u[t]+\omega\cdot(1-a_i^u[t])\big]\Big|\vec{\boldsymbol{\pi}}[0]\Big]-\omega\cdot(1-\alpha)N \\ &\leq \max_{u\in U}\ \limsup_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\sum_{i=1}^{N}\big[\pi_i[t]\cdot a_i^u[t]+\omega\cdot(1-a_i^u[t])\big]\Big|\vec{\boldsymbol{\pi}}[0]\Big]-\omega\cdot(1-\alpha)N \\ &\leq \max_{u\in U}\ \sum_{i=1}^{N}\limsup_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\big[\pi_i[t]\cdot a_i^u[t]+\omega\cdot(1-a_i^u[t])\big]\Big|\vec{\boldsymbol{\pi}}[0]\Big]-\omega\cdot(1-\alpha)N, \end{aligned}$$

where the first and the last inequality follows from the superadditiviey and subadditivity of limit superior and limit inferior, respectively. ■

Consider the unconstrained optimization problem in the upper bound (11), it can be viewed as a composition of N independent *ω-subsidy problems* interpreted as follows. For each channel i at belief state π_i, it will receive a reward π_i when it activates, otherwise it will receive a subsidy ω for passivity. Here, for each channel, its reward only depends on the transmissions of its own and independent of decisions for other channels. Hence, the optimization problem in (11) can be decomposed into N *ω-subsidy problems*. For channel i, the ω-subsidy problem is expressed as follows,

$$V_i(\omega)=\max_{u\in U_i}\ \limsup_{T\to\infty}\frac{1}{T}E\Big[\sum_{t=0}^{T-1}\big[\pi_i[t]\cdot a_i^u[t]+\omega\cdot(1-a_i^u[t])\big]\Big|\pi_i[0]\Big] \tag{12}$$

where U_i denotes the set of scheduling decisions that activate and idle the channel according to the observed channel history. An important property for each ω-subsidy problem is *thresholdability*, given in the following lemma.

Lemma 7. *The optimal policy for the ω-subsidy problem (12) is a threshold-based policy. Specifically, for each channel i in class k, there exists a threshold value $\theta_k(\omega)\in[0,1]$ such that it is optimal to transmit when its current belief value $\pi_i[t]>\theta_k(\omega)$, and to stay passive when $\pi_i[t]<\theta_k(\omega)$, with tie breaking arbitrarily at $\pi_i[t]=\theta_k(\omega)$.*

Proof: The thresholdability has been proved in [14] assuming a different form of objective function than (12). In fact, thresholdability holds for the general optimization problem of (12) as well, explained in details below.

It was shown in [14] that a stationary threshold-based policy u_β^* with threshold value $\theta_k(\beta,\omega)$ is optimal for the β-discounted ω-subsidy problem

$$\max_{u\in U_i}\ E\Big[\sum_{t=0}^{\infty}\beta^t\big[\pi_i[t]\cdot a_i^u[t]+\omega\cdot(1-a_i^u[t])\big]\Big|\pi_i[0]\Big]. \tag{13}$$

for channels in class k, where $\beta\in(0,1)$ is the discount factor. The optimal policy u_β^* for (13) activates the channels with belief values greater than $\theta_k(\beta,\omega)$ and idles the channels whose belief values are smaller than $\theta_k(\beta,\omega)$, with tie breaking arbitrarily at $\theta_k(\beta,\omega)$.

From Dutta's paper [22], we know that if a *value boundedness condition* holds for the discounted problem (13), and if a family of optimal policy $\{u_\beta^*\}$ converges to certain limit ϕ as $\beta\to1$, then ϕ is optimal for the v-subsidy average reward problem (12) that is defined with respective to limit superior.

Indeed, it was shown in [14] that as $\beta\to1$, $\theta_k(\omega)=\lim_{\beta\to1}\theta_k(\beta,\omega)$ exists and the value boundedness condition holds for the discounted problem (13). Therefore the threshold-based policy is optimal for the problem (12). ■

In the ω-subsidy problem, we let $\vec{\boldsymbol{\theta}}(\omega)=\{\theta_k(\omega),k=1,2\}$ denote the optimal threshold-based policy for the system. Because of the simple form of the threshold-based policy, we have the following lemma.

Lemma 8. *Given a Lagrange multiplier $\omega\geq0$, the threshold-based policy $\vec{\boldsymbol{\theta}}(\omega)$ achieves the maximum value of the Lagrange function $L(u,\omega)$, i.e.,*

$$D(\omega)=L\big(\vec{\boldsymbol{\theta}}(\omega),\omega\big). \tag{14}$$

Proof: Case (1). Suppose for channels in class k, $\theta_k(\omega) \geq b_s^k$. Due to the monotonicity behavior of belief evolution, channels with initial belief $\pi_i[0] < \theta_k(v)$ always stay idle. Channels with initial belief $\pi_i[0] \geq \theta_k(\omega)$ will be activated until its channel turns out to be 0 and then remain idle henceforth. Therefore with probability 1, all channels will stay in the idle mode (see proof of Lemma 1 for detailed description). Therefore, the following limit inferior will coincide with limit superior and can be calculated,

$$\liminf_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \big[\pi_i[t] \cdot a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t] \Big| \pi_i[0]\Big] = \limsup_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \big[\pi_i[t] \cdot a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t] \Big| \pi_i[0]\Big] = 0, \tag{15}$$

$$\liminf_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} (1 - a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t]) \Big| \pi_i[0]\Big] = \limsup_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} (1 - a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t]) \Big| \pi_i[0]\Big] = 1, \tag{16}$$

Case (2). Suppose for channels in class k, $\theta^k(\omega) < b_s^k$. From belief value evolution in Fig. 2, the belief values of each channel evolves as a positive recurrent Markov Chain (again, see proof of Lemma 1 for detailed description). Therefore, the limit inferior and limit superior coincides,

$$\liminf_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \big[\pi_i[t] \cdot a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t] \Big| \pi_i[0]\Big] = \limsup_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \big[\pi_i[t] \cdot a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t] \Big| \pi_i[0]\Big], \tag{17}$$

$$\liminf_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} (1 - a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t]) \Big| \pi_i[0]\Big] = \limsup_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} (1 - a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t]) \Big| \pi_i[0]\Big], \tag{18}$$

From equation (15)-(18), as well as equation (10), we have,

$$\begin{aligned} L\big(\vec{\boldsymbol{\theta}}(\omega), \omega\big) &= \sum_{i=1}^{N} \lim_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \big[\pi_i[t] a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t] \Big| \pi_i[0]\Big] + \sum_{i=1}^{N} \lim_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \omega(1 - a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t]) \Big| \pi_i[0]\Big] - \omega \cdot (1-\alpha) N \\ &= \sum_{i=1}^{N} \lim_{T\to\infty} \frac{1}{T} E\Big[\sum_{t=0}^{T-1} \big[\pi_i[t] \cdot a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t] + \omega \cdot (1 - a_i^{\vec{\boldsymbol{\theta}}(\omega)}[t])\big] \Big| \pi_i[0]\Big] - \omega \cdot (1-\alpha) N \\ &= \sum_{i=1}^{N} V_i(\omega) - \omega \cdot (1-\alpha) N \\ &\geq D(\omega), \end{aligned}$$

where the last inequality follows from Lemma 6. Because we also know $L\big(\vec{\boldsymbol{\theta}}(\omega), \omega\big) \leq D(\omega)$ since $D(\omega) = \max_{u \in U} L(u, \omega)$, we have $D(\omega) = L\big(\vec{\boldsymbol{\theta}}(\omega), \omega\big)$. ■

B. The ω-subsidy problem: Indexability

For each channel in class k, let $\mathcal{I}_k(\omega) \subseteq \mathcal{B}_k$ be the set of belief states for which, under threshold-based policy $\vec{\boldsymbol{\theta}}(\omega)$, it is optimal to stay idle. From the thresholdability property, it is clear that $\mathcal{I}_k(\omega)$ includes all the belief values in $\mathcal{B}_k$ no greater than $\theta_k(\omega)$.

For class-k channels, we let $a_\omega^k(\pi)$ denote the optimal decision at belief value $\pi \in \mathcal{B}^k$ under subsidy ω. Following the definition in [13], the Whittle's index value $W_k(\pi)$, $\pi \in \mathcal{B}^k$, is given by the infimum value of subsidy ω for which it is equally optimal activate or idle at belief π, i.e.,

$$W_k(\pi) = \inf\{\omega : a_\omega^k(\pi) = \{0, 1\}\}. \tag{19}$$

The Whittle's Indexability condition, specific to the scheduling problem, is defined as follows.

Whittle's Indexability condition: The downlink scheduling problem is Whittle Indexable if, as ω increases from $-\infty$ to ∞, the set $\mathcal{I}_k(\omega)$ monotonically increases from $\emptyset$ to $\mathcal{B}_k$.

It was proved in [14] that the idle set $\mathcal{I}_k(\omega)$ indeed monotonically increases from $\emptyset$ to $\mathcal{B}_k$ as ω increases from $-\infty$ to ∞. Therefore, the downlink scheduling problem is Whittle indexable, recorded below.

Indexability Theorem. The downlink scheduling problem is Whittle indexable.

It can be observed that, from Indexability condition, the Index value $W_k(\pi) \in [0,1]$ and $W_k(\pi)$ monotonically increases with $\pi \in \mathcal{B}^k$. The next lemma gives the closed form expression of the index values.

Lemma 9. *The closed form expression of Whittle's index values is given as follows,*

$$W_k(\pi) = \begin{cases} \frac{(b^k_{0,l} - b^k_{0,l+1})(l+1) + b^k_{0,l+1}}{1 - p_k + (b^k_{0,l} - b^k_{0,l+1})l + b^k_{0,l+1}} & \textit{if } p_k \leq \pi = b^k_{0,l} < b^k_s \\ \frac{r_k}{(1-p_k)(1+r_k-p_k)+r_k} & \textit{if } b^k_s \leq \pi \leq p_k \end{cases} \tag{20}$$

Proof: The derivation of the Whittle's Index value is included in Appendix K. We remark that the Index expression $W_k(\pi)$ is constant when $b^k_s \leq \pi \leq p_k$, which differs from the indices derived in [14]. Such a difference is due to the definition (19) of the index value and is explained in detail in Appendix K. ■

With the definition of index value and the established indexability condition, the optimal threshold-based policy can be implemented in a more efficient manner, characterized in Lemma 10. Here instead of maintaining different threshold values $\theta_i(\omega)$ for each ω, the scheduler simply compares the index value with ω to decide weather to transmit on the channel.

Lemma 10. *Under the ω-subsidized problem, at each time slot, for the i^{th} channel in class k, it is optimal to transmit when $W_k(\pi_i) > \omega$, and to stay idle when $W_k(\pi_i) < \omega$, with tie breaking arbitrarily if $W_k(\pi_i) = \omega$.*

Proof: If $W_k(\pi_i) > \omega$, from definition (19) of the Index value, $W_k(\pi_i)$ is the minimum subsidy required for the belief value π_i to be within the idle set. Since $W_k(\pi_i)$ is higher than the actual subsidy ω, it is optimal to activate the channel at subsidy ω.

If $W_k(\pi_i) < \omega$, similarly we know the subsidy ω is higher than the minimum subsidy value $W_k(\pi_i)$ such that it is optimal to stay idle at π_i. Hence, from Indexability, it is optimal to stay idle at π_i at subsidy ω.

If $W_k(\pi_i) = \omega$, from the above definition (19) of Index value, it is equally optimal to activate or idle at π_i. ■

C. Optimal policy for the relaxed problem

We have thus far seen from Lemma 10 that the dual function $D(\omega)$ can be achieved by a threshold-based policy implemented over the index values. We now proceed to identify the optimal policy for the original relaxed problem (4)-(5).

Let $\phi(\omega, \rho)$ denote the policy where the channels with the index value greater than ω activate, channels with the index value smaller than ω remain idle, and the channels with index value ω activate with probability ρ.

Lemma 11. *Given α, there exists a unique pair (ω^*, ρ^*) such that, under policy $\phi(\omega^*, \rho^*)$,*

$$\lim_{T \to \infty} \frac{1}{T} E\Big[\sum_{t=1}^{T} \sum_{i=1}^{N} a_i^{\phi(\omega^*, \rho^*)}[t]\Big] = \alpha N. \tag{21}$$

Proof: For a single channel i in class k, consider the policy where the channel activates if its belief value $\pi_i > b^k$, stays idle when $\pi_i < b^k$, and activates with probability ρ when $\pi_i = b^k$, for some belief value b^k. From the belief value evolution we can calculate the expected time of activion, denoted by $A^k(b^k, \rho)$,

$$A^k(b^k, \rho) = \begin{cases} 1 - \frac{(1-p_k)(h-\rho)}{\rho b^k_{0,h} + (1-\rho) b^k_{0,h+1} + (1-p_k)(h+1-\rho)} & \text{if } b^k = b^k_{0,h}, \\ 0 & \text{if } \pi \geq b^k_s. \end{cases} \tag{22}$$

It is clear from its expression that, given b^k, $A^k(b^k, \rho)$ is continuous with ρ. Also we have $A^k(b^k_{0,h}, 0) = A^k(b^k_{0,h+1}, 1)$. In addition, some simple algebra reveals that, given $b^k_{0,h}$, $A^k(b^k_{0,h}, \rho)$ strictly increases with ρ. Therefore, since $A^k(b^k_{0,h}, 0) = A^k(b^k_{0,h+1}, 1)$, given ρ $A^k(b^k, \rho)$ monotonically decreases with $b^k \in \mathcal{B}_k$.

Also, one can observe from (22) that, given ρ, $\lim_{h \to \infty} A^k(b^k_{0,h}, \rho) = 0$ and $A^k(b^k_{0,1}, 1) = 1$. Hence by appropriately choosing b^k and ρ, $A^k(b^k, \rho)$ can achieve any value within $[0, 1]$.

Recall from the definition of indexability that the index value $W_k(b^k)$ monotonically increases with $b^k \in \Pi^k$, $k = 1, 2$. It follows from the above analysis that, as ω increases, under policy $\phi(\omega, 1)$, the fraction of activation

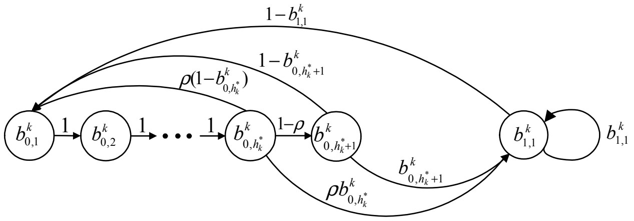

Fig. 6: Belief value transition in steady state when $\omega^* = W_k(b^k_{0,h^*_k})$

time for each user strictly decreases from 1 to 0. Therefore, there exists an unique $(\omega^*, \rho^*$ pair, such that the policy $\phi(\omega^*, \rho^*)$ strictly satisfies activation constraint. ■

We now consider the relaxed optimization problem (4)-(5) and the policy $\phi(\omega^*, \rho^*)$ as specified in the previous lemma. Clearly, the policy $\phi(\omega^*, \rho^*)$ is primal feasible and the lagrange multiplier ω^* is dual feasible. From Lemma 10, $\phi(\omega^*, \rho^*)$ is optimal for the ω^*-subsidy problem and hence $D(\omega^*) = L(\phi(\omega^*, \rho^*), \omega^*)$. Furthermore, according to (21), $\phi(\omega^*, \rho^*)$ activates αN users on average, and thus the complementary slackness condition holds for the primal-dual pair $\big(\phi(\omega^*, \rho^*), \omega^*\big)$. From the optimality condition for primal-dual optimal solution [23], $\big(\phi(\omega^*, \rho^*), \omega^*\big)$ is an optimal primal-dual pair. Therefore $\omega^* \in \arg\min_\omega D(\omega)$ and $\phi(\omega^*, \rho^*)$ is the optimal solution to the relaxed problem. Letting ϕ^* represent $\phi(\omega^*, \rho^*)$, we thus have proved Proposition 1.

Appendix B
Proof of Lemma 1

(i) First consider the scenario where $\omega^* < W_k(b^k_s)$ and suppose $\omega^* = W_k(b^k_{0,h^*_k})$ for the belief state b^k_{0,h^*_k}. If the belief value of a channel is above b^k_{0,h^*_k} at the beginning of a slot, the channel will be activated. According to the belief value evolution rule (1), in the next slot its belief value will either be p_k or r_k, depending on the underlying channel state revealed at the end of a slot. Clearly, the belief evolution in this case is positive recurrent within a finite state space, i.e., the belief state can only take the values $p_k, r_k, b^k_{0,2}, \cdots, b^k_{0,h^*_k+1}$. On the other hand, if the belief value is below b^k_{0,h^*_k}, the channel remains idle and will activate once its belief value exceeds b^k_{0,h^*_k}. Fig. 6 illustrates the belief evolution in steady state under this scenario.

(ii) Consider the scenario where $\omega^* \geq W_k(b^k_s)$. In this case, a channel is activated if its index value is above ω^*. After transmission, if the channel is observed to be in OFF state, its belief value will transit to r_k and stays idle until its index value crosses ω^*. Since $\omega^* \geq W_k(b^k_s)$, it is clear from the belief value evolution (see Fig. 2) that, starting from r_k, the belief value will always be smaller than b^k_s. Hence the channel will stay idle at all times. On the other hand, if the channel is observed to be in ON state after transmission, the belief value will transit to p_k and the channel will keep on transmitting until the underlying channel turns out to be in OFF state. Since we assumed $p_k < 1$, the channel will ultimately be in OFF state and its belief value will transit to r_k and stays in idle mode ever since. Therefore eventually no channel in class k will be scheduled and the belief values will keep transit toward, but never reach, the steady state belief value b^k_s.

Appendix C
Proof of Lemma 2

Consider two systems with different total number of users but identical α and $\boldsymbol{\gamma}$. Suppose the first system has N_1 total number of users while the second system has N_2 number of users. For the first system with N_1 total number of users, suppose the policy ϕ^*, specified in Proposition 1, is optimal for the relaxed-constraint problem. Therefore from the proof of Proposition 1, ϕ^* is optimal for each individual ω^*-subsidized problem (12). For each channel in class k, we let $A^k_{\phi^*}$ denote the expected fraction of time of activatoin, which is expressed specifically in equation (22). Then, according to Proposition 1(ii), the expected number of activated users satisfies

$$\gamma_1 N_1 \cdot A^1_{\phi^*} + \gamma_2 N_1 \cdot A^2_{\phi^*} = \alpha N_1. \tag{23}$$

Now apply the same policy ϕ^* when the total number of users is N_2. Since ϕ^* schedules each channel independently, $A^1_{\phi^*}$ and $A^2_{\phi^*}$ does not change in this scenario. Therefore, the expected number of activated users is expressed as

$$\gamma_1 N_2 \cdot A^1_{\phi^*} + \gamma_2 N_2 \cdot A^2_{\phi^*} = \frac{N_2}{N_1}\left[\gamma_1 N_1 \cdot A^1_{\phi^*} + \gamma_2 N_1 \cdot A^2_{\phi^*}\right] = \alpha N_2, \tag{24}$$

hence the complementary slackness condition for the relaxed-constraint problem is also satisfied under ϕ^* and ω^*, when the total number of users is N_2. In this case, since ϕ^* is still optimal for each individual ω^*-subsidized problem and both ϕ^* and ω^* are feasible, (ϕ^*, ω^*) is a primal-dual feasible pair when the total number of users is N_2, by the same argument as in the proof of Proposition 1.

Therefore, fixing system parameters $(\boldsymbol{\gamma}, \alpha)$, for different number N of users, the policy ϕ^* is always optimal. Since the policy ϕ^* schedules each channel independently, we let $\upsilon_k(\boldsymbol{\gamma}, \alpha)$ denote the expected reward contributed by each channel in class k. Hence we have

$$\upsilon^N(\boldsymbol{\gamma}, \alpha) = N\gamma_1\upsilon_1(\boldsymbol{\gamma}, \alpha) + N\gamma_2\upsilon_2(\boldsymbol{\gamma}, \alpha).$$

Therefore the per-user throughput is

$$\frac{\upsilon^N(\boldsymbol{\gamma}, \alpha)}{N} = \gamma_1\upsilon_1(\boldsymbol{\gamma}, \alpha) + \gamma_2\upsilon_2(\boldsymbol{\gamma}, \alpha),$$

which is independent of N. Hence the lemma is proven.

APPENDIX D
PROOF OF LEMMA 3

Given system parameters $(\boldsymbol{\gamma}, \alpha)$, we know from the proof of Lemma 2 that the form of the Optimal Relaxed Policy, denoted by ϕ^*, does not change with the number N of users. Since ϕ^* schedules each channel independently, we let vector $\boldsymbol{\varepsilon}^k = [\varepsilon^k_{0,1}, \cdots, \varepsilon^k_{0,\tau}, \varepsilon^k_s, \varepsilon^k_{1,\tau}, \cdots, \varepsilon^k_{1,1}]$ denote the steady state distribution of the belief value of users in class k under ϕ^*, with $\varepsilon^k_s + \sum_{c,h} \varepsilon^k_{c,h} = 1$. Therefore,

$$E[\boldsymbol{Z}^N(\infty)] = \frac{1}{N}[\gamma_1 N \boldsymbol{\varepsilon}^1, \gamma_2 N \boldsymbol{\varepsilon}^2] = [\gamma_1 \boldsymbol{\varepsilon}^1, \gamma_2 \boldsymbol{\varepsilon}^2].$$

Since ϕ^* is independent of N, $\boldsymbol{\varepsilon}^k$ is independent of N for $k = 1, 2$. Therefore $E[\boldsymbol{Z}^N(\infty)]$ is independent of the user number N, which proves the lemma.

APPENDIX E
PROOF OF PROPOSITION 2

A. Notations

We shall denote the i^{th} element of $\boldsymbol{Z}^N[t]$ as $Z^N_i[t]$, and let β_i denote the corresponding belief value. The index value corresponding to β_i is denoted as w_i. In this proof, since we are fixing the system parameters $(\boldsymbol{\gamma}, \alpha)$, we shall drop the suffixes α and $\boldsymbol{\gamma}$ to denote $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}$ as $\vec{\boldsymbol{\zeta}}$.

For ease of exposition, in this proof we assume $W_k(b^2_{0,h^*_2-1}) < W_1(b^1_{0,h^*_1}) = \omega^* < W_k(b^2_{0,h^*_2})$. Hence, in the optimal relaxed problem, channels in class 1 are activated when their belief values are above b^1_{0,h^*_1} and stay idle if their belief values are below b^1_{0,h^*_1}, and activates with probability $\rho^* \in (0, 1)$ at b^1_{0,h^*_1}. For channels in class 2, they are activated when their belief values no smaller than b^2_{0,h^*_2} and stay idle otherwise.

B. Transition properties of the system state

We first investigate the belief transition structure of the system state $\boldsymbol{Z}^N[t]$ under the Whittle's Index Policy. It is clear that $\boldsymbol{Z}^N[t]$ evolves as a Markov Chain. We define the *expected drift* $\nabla \boldsymbol{Z}^N[t]$ associated with the transition of $\boldsymbol{Z}^N[t]$ as follows,

$$\nabla \boldsymbol{Z}^N[t] = E\big[\boldsymbol{Z}^N[t+1] - \boldsymbol{Z}^N[t] \,\big|\, \boldsymbol{Z}^N[t]\big]. \tag{25}$$

For a channel with belief value β_i, we let $q^0_{i,j}$ and $q^1_{i,j}$ be the probability that its belief state changes to state β_j under the idle and transmission actions, respectively. For example, if β_i corresponds to belief value $b^1_{0,l}$, then $q^0_{i,i+1} = 1$ if the channel stays idle, otherwise $q^1_{i,1} = 1 - b^1_{0,l}$ and $q^1_{i,2\tau+1} = b^1_{0,l}$, which corresponds to the probability of observed channel being 0 or 1, respectively. Under the Whittle's Index Policy, we let $g_i(\boldsymbol{z})$ be the fraction of users in belief value β_i that are activated, which is expressed as,

$$g_i(\boldsymbol{z}) = \begin{cases} \min\Big\{\Big[\frac{\alpha - \sum_{w_j > w_i} z_j > z_i}{z_i}\Big]^+, 1\Big\}, & \text{if } z_i \neq 0, \\ 1, & \text{if } z_i = 0 \text{ and } \alpha - \sum_{w_j > w_i} z_j > 0, \\ 0, & \text{if } z_i = 0 \text{ and } \alpha - \sum_{w_j > w_i} z_j \leq 0. \end{cases} \tag{26}$$

where $[\cdot] = \max\{0, \cdot\}$. We use $q_{i,j}(\boldsymbol{z})$ to denote the probability that the belief value of a channel transit from β_i to β_j under system state $\boldsymbol{z}$. Then $q_{ij}(\boldsymbol{z})$ is expressed as

$$q_{ij}(\boldsymbol{z}) = g_i(\boldsymbol{z}) q^1_{ij} + \big(1 - g_i(\boldsymbol{z})\big) q^0_{ij}. \tag{27}$$

We shall let $\boldsymbol{e}_{ii} = \vec{0}$, and let $\boldsymbol{e}_{ij}, i \neq j$ be a vector that has -1 at the i^{th} element, $+1$ at the j^{th} element, and 0 at all other elements. Hence if a user changes its belief state from β_i to β_j, the corresponding change of the system state $\boldsymbol{Z}^N[t]$ is in the direction of $\boldsymbol{e}_{ij}$ with magnitude $1/N$. Therefore, $\nabla \boldsymbol{Z}^N[t]$ is a composition of expected changes in each direction $\boldsymbol{e}_{ij}$. Suppose $\boldsymbol{Z}^N[t] = \boldsymbol{z}$, since the expected amount of change of $\boldsymbol{Z}^N[t]$ in direction $\boldsymbol{e}_{ij}$ is $z_i[t] q_{ij}(\boldsymbol{z}[t])$, the expected drift $\nabla \boldsymbol{Z}^N[t]$ can then be written as,

$$\nabla \boldsymbol{Z}^N[t]\Big|_{\boldsymbol{Z}^N[t] = \boldsymbol{z}} = \sum_{i,j} z_i q_{ij}(\boldsymbol{z}) \cdot \boldsymbol{e}_{ij} := Q(\boldsymbol{z})\boldsymbol{z}, \tag{28}$$

where the $(i,j)^{th}$ element of matrix $Q(\boldsymbol{z})$ is expressed as

$$Q_{ij}(\boldsymbol{z}) = \begin{cases} -\sum_{j \neq i} q_{ij}(\boldsymbol{z}) & \text{for } i = j, \\ q_{ji}(\boldsymbol{z}) & \text{for } i \neq j. \end{cases} \tag{29}$$

Note that, although the system state $\boldsymbol{z}$ can only take values on a lattice that depends on N, the matrix function $Q_{ij}(\boldsymbol{z})$ is defined over more general space $\mathcal{Z}$. Based on this, we proceed to define a fluid approximation model.

C. Fluid Approximation Model

We consider a fluid approximation model $\boldsymbol{z}[t]$, which is defined by the following difference equation

$$\boldsymbol{z}[t+1] - \boldsymbol{z}[t] = Q(\boldsymbol{z}[t])\boldsymbol{z}[t]. \tag{30}$$

Note that the right-hand-side is completely determined by equation (26)-(29), as a function of $\boldsymbol{z}[t]$ and is independent of N. We denote $\boldsymbol{z}[t]$ as the 'fluid approximation model' because $\boldsymbol{z}[t]$ is no longer restricted to take values on the lattice as with the case of the original system state $\boldsymbol{Z}^N[t]$, and $\boldsymbol{z}[t]$ evolves in the direction of the *expected change* of the system state [1]. Recall that the set $\mathcal{Z}$ is defined in equation (6), we proceed with the following lemma.

Lemma 12. *If $\boldsymbol{z}[0] \in \mathcal{Z}$, then $\boldsymbol{z}[t] \in \mathcal{Z}$ for all $t \geq 0$.*

[1] Note that by 'fluid' we mean fluid in users/channels instead of fluid with respective to time.

Proof: Since from (28) we have

$$\boldsymbol{z}[t+1]-\boldsymbol{z}[t]\Big|_{\boldsymbol{z}[t]=\boldsymbol{z}} = Q(\boldsymbol{z}[t])\boldsymbol{z} = \sum_{i,j} z_i[t]q_{ij}(\boldsymbol{z}[t])\cdot \boldsymbol{e}_{ij}.$$

Note that the belief values of a channel can only evolve within the belief states of class of the channel, hence for class 1,

$$\begin{aligned}\sum_{i=1}^{2\tau+1} z_i[t+1]-\sum_{i=1}^{2\tau+1} z_i[t] &= \vec{\mathbf{1}}^T\cdot \sum_{1\le i,j\le 2\tau+1} z_i[t]q_{ij}(\boldsymbol{z}[t])\boldsymbol{e}_{ij}\\ &= \sum_{1\le i,j\le 2\tau+1} z_i[t]q_{ij}(\boldsymbol{z}[t])\cdot(1-1)\\ &= 0.\end{aligned}$$

where $\vec{\mathbf{1}}$ is a vector with 1 in each element. Similar result holds for class 2. Since $z[0]{\in}\mathcal{Z}$, we have

$$\sum_{i=1}^{2\tau+1} z_i[t]\equiv\gamma_1,\ \sum_{i=2\tau+2}^{2(2\tau+1)} z_i[t]\equiv\gamma_2,\ \forall t\ge 0.$$

Also equation (28)-(30) indicates that $z_i[t]{\ge}0$ for all $t{\ge}0$ if $\boldsymbol{z}[0]\in\mathcal{Z}$. Therefore $\boldsymbol{z}[t]{\in}\mathcal{Z}$ for all $t\ge 0$, establishing the lemma. ■

Lemma 13. *The vector $\vec{\boldsymbol{\zeta}}$ is the unique fixed point of the fluid approximation model, i.e., for all $\boldsymbol{z}\in\mathcal{Z}$, $Q(\boldsymbol{z})\boldsymbol{z}=0$ if and only if $\boldsymbol{z}=\vec{\boldsymbol{\zeta}}$.*

Proof: The proof follows from a similar line of [16]. Note that, under the Optimal Relaxed Policy, $\vec{\boldsymbol{\zeta}} = E\big[\boldsymbol{Z}^N(\infty)\big]$ and α fraction of channels are activated on average. Therefore, in the fluid approximation model, we have $\boldsymbol{z}[t+1]-\boldsymbol{z}[t]\big|_{\boldsymbol{z}[t]=\vec{\boldsymbol{\zeta}}}=0$, i.e., $Q(\vec{\boldsymbol{\zeta}})\vec{\boldsymbol{\zeta}}=0$.

Now suppose there exists another fixed point $\vec{\boldsymbol{\zeta}}_0\in\mathcal{Z}$ such that $\vec{\boldsymbol{\zeta}}_0\neq\vec{\boldsymbol{\zeta}}$ and $Q(\vec{\boldsymbol{\zeta}}_0)\vec{\boldsymbol{\zeta}}_0=0$. Then $\vec{\boldsymbol{\zeta}}_0$ corresponds to the stationary distribution of the system state under another policy $\phi(\omega_0,\rho_0)$ with threshold parameter ω_0 and randomization factor ρ_0. Furthermore, under $\phi(\omega_0,\rho_0)$, the expected fraction of activated channels equals to α. However, this contradicts with Lemma 11, which states that (ω^*,ρ^*) is the unique parameter pairs that strictly satisfies the average constraint of activation. Therefore, the fixed point $\vec{\boldsymbol{\zeta}}$ is unique. ■

D. Convergence of the Fluid Limit Model

Define the region $\mathcal{J}_{\omega^*}\subseteq\mathcal{Z}$ as the set of $\boldsymbol{z}\in\mathcal{Z}$ such that, under the Whittle's Index Policy defined in Section IV, the channel is activated if and only if its index value is no smaller than ω^*, which is the threshold for the Optimal Relaxed Policy defined in Proposition 1. This means that, at system state $z\in\mathcal{J}_{\omega^*}$, all channels with index value higher than ω^* are scheduled, and the channels with index value smaller than ω^* stay idle, while the channels at index value ω^* are scheduled with certain randomization. Specifically, $\mathcal{J}_{\omega^*}=\{\boldsymbol{z}{\in}\mathcal{Z}: \sum_{i:w_i>\omega^*} z_i<\alpha,\ \sum_{i:w_i\ge\omega^*} z_i\ge\alpha.\}$.

The following lemma characterizes the linearity property of the fluid approximation model in J_{ω^*}.

Lemma 14. *(i) The vector $\vec{\boldsymbol{\zeta}}\in\mathcal{J}_{\omega^*}$.*

(ii) The fluid difference equation (30) is linear within the region $\mathcal{J}_{\omega^*}$, i.e., there exist matrix $\boldsymbol{Q}^*$ and vector $\boldsymbol{a}^*$ such that

$$\boldsymbol{z}[t+1]-\boldsymbol{z}[t]=Q^*\cdot\boldsymbol{z}[t]+\boldsymbol{a}^*,\quad \textit{for all}\ \ \boldsymbol{z}[t]\in\mathcal{J}_{\omega^*}. \tag{31}$$

Proof: (i) The vector $\vec{\boldsymbol{\zeta}}\in\mathcal{J}_{\omega^*}$ because, if $\boldsymbol{z}[t]=\vec{\boldsymbol{\zeta}}$, we have $\sum_{i:w_i\ge\omega^*} g_i(\boldsymbol{z}[t])z_i[t]=\alpha$.

(ii) Recall that, at the beginning of the section, we have assumed $\omega^* = W_1(b^1_{0,h^*_1})$ for the belief value b^1_{0,h^*_1} of class-1 channel. The difference equation (30) becomes,

$$\begin{aligned} \boldsymbol{z}[t+1] - \boldsymbol{z}[t]\Big|_{\boldsymbol{z}[t]=\boldsymbol{z}} &= \sum_{i,j:i\neq h^*_1} z_i q_{ij}(\boldsymbol{z}) \cdot \boldsymbol{e}_{ij} + z_{h^*_1} \sum_j q_{h^*_1 j}(\boldsymbol{z}) \cdot \boldsymbol{e}_{h^*_1 j} \\ &= \sum_{i,j:i\neq h^*_1} z_i q_{ij}(\boldsymbol{z}) \cdot \boldsymbol{e}_{ij} + z_{h^*_1} \sum_j \left[g_{h^*_1}(\boldsymbol{z}) q^1_{h^*_1 j} + [1 - g_{h^*_1}(\boldsymbol{z})] q^0_{h^*_1 j} \right] \cdot \boldsymbol{e}_{h^*_1 j} \\ &= \sum_{i,j:i\neq h^*_1} z_i q_{ij}(\boldsymbol{z}) \cdot \boldsymbol{e}_{ij} + z_{h^*_1} \sum_j q^0_{h^*_1 j} \cdot \boldsymbol{e}_{h^*_1 j} + g_{h^*_1}(\boldsymbol{z}) z_{h^*_1} \sum_j \left[q^1_{h^*_1 j} - q^0_{h^*_1 j} \right] \cdot \boldsymbol{e}_{h^*_1 j}. \end{aligned} \tag{32}$$

where the second equality is from (27).

Since the total fraction of users activated is α, we have

$$g_{h^*_1}(\boldsymbol{z}) z_{h^*_1} = \alpha - \sum_{w_i > \omega^*} z_i, \tag{33}$$

Substituting the expression (33) back in (32), and noting that $q_{ij}(\boldsymbol{z}), i{\neq}h^*_1$ stays constant for $\boldsymbol{z}{\in}\mathcal{J}_{\omega^*}$ (since the threshold ω^* for activation does not change for $\boldsymbol{z} \in \mathcal{J}_{\omega^*}$), the linearity property holds. ■

From Lemma 12 we know that $\boldsymbol{z}[t] \in \mathcal{Z}$ for all $t \geq 0$, i.e.,

$$\sum_{i=1}^{2\tau+1} z_i = \gamma_1, \quad \sum_{i=2\tau+2}^{2(2\tau+1)} z_i = \gamma_2. \tag{34}$$

Taking note of Lemma 12, instead of using a $2(2\tau+1)$ dimensional vector $\boldsymbol{z}$, it suffices to represent the system state by a $2 \cdot 2\tau$ dimension vector $\tilde{\boldsymbol{z}}$, i.e.,

$$\tilde{\boldsymbol{z}} = \left[z_1, \cdots, z_{h^*_1 - 1}, z_{h^*_1 + 1}, \cdots, z_{2\tau + h^*_2 - 1}, z_{2\tau + h^*_2 + 1}, \cdots z_{2(2\tau+1)} \right].$$

in which elements $z_{h^*_1}$ and $z_{2\tau+h^*_2}$ are eliminated from $\boldsymbol{z}$. The transition of $\tilde{\boldsymbol{z}}[t]$, when $\boldsymbol{z}[t] \in \mathcal{J}_{\omega^*}$, is obtained by substituting the relationship (34) in the difference equation (32) and eliminate the elements $z_{h^*_1}$ and $z_{2\tau+h^*_2}$, i.e.,

$$\tilde{\boldsymbol{z}}[t+1] - \tilde{\boldsymbol{z}}[t] = U^* \cdot \tilde{\boldsymbol{z}}[t] + \boldsymbol{b}^*., \tag{35}$$

where the matrix U^* and vector $\boldsymbol{b}^*$ are obtained after the substitution. The next key lemma captures the eigen structure of matrix U^*.

Lemma 15. *Each eigen value λ of U^* satisfies $\left|\lambda + 1\right| < 1$.*

Proof: The proof is based on explicit study of matrix U^* and is given in Appendix I. ■

This lemma leads to the local convergence of $\boldsymbol{z}[t]$.

Lemma 16. *There exists a positive constant σ such that, if the initial state $\boldsymbol{z}[0] = \boldsymbol{x}$ of the fluid approximation model is within the σ neighborhood $\Omega_\sigma(\vec{\boldsymbol{\zeta}})$ of $\vec{\boldsymbol{\zeta}}$, where $\Omega_\sigma(\vec{\boldsymbol{\zeta}}) \subseteq J_{\omega^*}$, then*

(i) $\boldsymbol{z}[t] \in \mathcal{J}_{\omega^*}$ for all $t \geq 0$; *(ii) $\boldsymbol{z}[t] \to \vec{\boldsymbol{\zeta}}$ as $t \to \infty$.*

Proof: Corresponding to $\vec{\boldsymbol{\zeta}}$, we let $\tilde{\boldsymbol{\zeta}}$ represent the stationary expectation of vector $\tilde{\boldsymbol{z}}[t]$. Therefore, from Lemma 13,

$$U^* \cdot \tilde{\boldsymbol{\zeta}} + \boldsymbol{b}^* = 0. \tag{36}$$

Substituting (36) in equation (35), we have

$$\tilde{\boldsymbol{z}}[t] - \tilde{\boldsymbol{\zeta}} = (U^* + I)^t (\boldsymbol{x} - \tilde{\boldsymbol{\zeta}}). \tag{37}$$

Since we have assumed that $\rho^* \neq 1$, there exists a σ_0 neighborhood $\Omega_{\sigma_0}(\vec{\boldsymbol{\zeta}})$ with $\Omega_{\sigma_0}(\vec{\boldsymbol{\zeta}}) \subseteq J_{\omega^*}$. Correspondingly, there is a neighborhood of $\tilde{\boldsymbol{\zeta}}$ for which $\tilde{\boldsymbol{z}}[t]$ evolution is linear and is described by (37). From Lemma 16, each eigen value λ of $(U^* + I)$ satisfies $\left|\lambda\right| < 1$. According to the stability theory of linear systems [26], $\tilde{\boldsymbol{z}}[t]$ converges to $\tilde{\boldsymbol{\zeta}}$ if the initial state is close enough to $\tilde{\boldsymbol{\zeta}}$.

Therefore, there exists a $\sigma < \sigma_0$ neighborhood of $\vec{\boldsymbol{\zeta}}$ for which if the initial state $\boldsymbol{x} \in \Omega_\sigma(\vec{\boldsymbol{\zeta}})$, $\boldsymbol{z}[t] \in \mathcal{J}_{\omega^*}$ and $\boldsymbol{z}[t] \to \vec{\boldsymbol{\zeta}}$ as $t \to \infty$. ■

E. Convergence of the system state

The fluid approximation model provides a good estimate for the system state evolution when the number of users is large, captured in the following proposition, which can be viewed as a *discrete-time version* of Kurtz theorem [24] applied to our problem.

Proposition 4. *There exists a neighborhood $\Omega_\delta(\vec{\zeta})$ of $\vec{\zeta}$ such that if $\boldsymbol{Z}^N[0]{=}\boldsymbol{z}[0]{=}\boldsymbol{x} \in \Omega_\delta(\vec{\zeta})$, then for any $\mu > 0$ and finite time horizon T there exists positive constants C_1 and C_2 such that*

$$P_{\boldsymbol{x}}\Big(\sup_{0\leq t<T} \big\|\boldsymbol{Z}^N[t]-\boldsymbol{z}[t]\big\| \geq \mu\Big) \leq C_1 \exp(-NC_2),$$

where $\delta < \sigma$, and $P_{\boldsymbol{x}}$ denotes the probability conditioned on the initial state $\boldsymbol{Z}^N[0] = \boldsymbol{x}$. Furthermore, C_1 and C_2 are independent of $\boldsymbol{x}$ and N.

Proof: Consider the random variable $\boldsymbol{Z}^N[t+1]$ given $\boldsymbol{Z}^N[t] = \boldsymbol{z}$,

$$\boldsymbol{Z}^N[t+1] = \boldsymbol{Z}^N[t] + \sum_{i,j=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z}) \cdot \boldsymbol{e}_{ij}}{N}, \tag{38}$$

where $\eta_{ij}^h(\boldsymbol{z})$ is an indicator function representing whether the belief value of the h^{th} user in belief value β_i transits to belief value β_j at the next time slot. Note that, given $\boldsymbol{Z}^N[t] = \boldsymbol{z}$, the scheduling action for users in belief state β_i is independent of N because the scheduling decision only depends on the belief state distribution $\boldsymbol{z}$. As N increases and $\boldsymbol{z}$ stays unchanged, more users is in belief state β_i and the contribution of each channel to the transition of $\boldsymbol{Z}^N$ scales down with N. From the law of large numbers, if the number of users scales up while z_i is kept the same, we have

$$\lim_{N\to\infty} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z})}{N} = \lim_{N\to\infty} \frac{Nz_i}{N} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z})}{Nz_i} = z_i q_{ij}(\boldsymbol{z}) \quad \text{almost surely.}$$

Lemma 17. *There exists a neighborhood $\Omega_\varepsilon(\vec{\zeta})$ of $\vec{\zeta}$ such that, if $Z^N[t] = \boldsymbol{z} \in \Omega_\varepsilon(\boldsymbol{\zeta})$, there exist c_1 and c_2 for which $Z^N[t+1]$ satisfies*

$$P\Big(\big\|\boldsymbol{Z}^N[t+1] - \big(I + Q(\boldsymbol{z})\big)\boldsymbol{z}\big\| \geq \mu \Big| Z^N[t] = \boldsymbol{z}\Big) \leq c_1 \exp(-Nc_2),$$

where c_1 and c_2 are independent of $\boldsymbol{z}$ and N.

Proof: Let $\vec{\mathbf{1}}_i$ be a vector with 1 at the i^{th} position. From (38),

$$\begin{aligned}
&\boldsymbol{Z}^N[t+1] - \big(I + Q(\boldsymbol{z})\big)\boldsymbol{z} \\
=& \sum_{i,j=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z})}{N} \cdot \boldsymbol{e}_{ij} - Q(\boldsymbol{z})\boldsymbol{z} \\
=& \sum_{i,j=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z})}{N} \cdot \boldsymbol{e}_{ij} - \sum_{i,j=1}^{2(2\tau+1)} z_i q_{ij}(\boldsymbol{z}) \cdot \boldsymbol{e}_{ij} \\
=& \sum_{i,j=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z})}{N} \cdot \big(\vec{\mathbf{1}}_j - \vec{\mathbf{1}}_i\big) - \sum_{i,j=1}^{2(2\tau+1)} z_i q_{ij}(\boldsymbol{z}) \cdot \big(\vec{\mathbf{1}}_j - \vec{\mathbf{1}}_i\big) \\
=& \Big[\sum_{i,j=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z})}{N} \cdot \vec{\mathbf{1}}_j - \sum_{i,j=1}^{2(2\tau+1)} z_i q_{ij}(\boldsymbol{z}) \cdot \vec{\mathbf{1}}_j\Big] - \Big[\sum_{i,j=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z})}{N} \cdot \vec{\mathbf{1}}_i - \sum_{i,j=1}^{2(2\tau+1)} z_i q_{ij}(\boldsymbol{z}) \cdot \vec{\mathbf{1}}_i\Big].
\end{aligned}$$

Note that

$$\begin{aligned}\sum_{i,j=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i} \eta_{ij}^h(\boldsymbol{z})}{N} \cdot \vec{\mathbf{1}}_i - \sum_{i,j=1}^{2(2\tau+1)} z_i q_{ij}(\boldsymbol{z}) \cdot \vec{\mathbf{1}}_i &= \sum_{i=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i}\sum_{j=1}^{2(2\tau+1)} \eta_{ij}^h(\boldsymbol{z})}{N} \cdot \vec{\mathbf{1}}_i - \sum_{i=1}^{2(2\tau+1)} z_i \sum_{j=1}^{2(2\tau+1)} q_{ij}(\boldsymbol{z}) \cdot \vec{\mathbf{1}}_i \\ &= \sum_{i=1}^{2(2\tau+1)} z_i \vec{\mathbf{1}}_i - \sum_{i=1}^{2(2\tau+1)} z_i \vec{\mathbf{1}}_i \\ &= 0.\end{aligned}$$

Therefore

$$\boldsymbol{Z}^N[t+1] - \big(I + Q(\boldsymbol{z})\big)\boldsymbol{z} = \sum_{i,j=1}^{2(2\tau+1)} \frac{\sum_{h=1}^{Nz_i} \big(\eta_{ij}^h(\boldsymbol{z}) - q_{ij}(\boldsymbol{z})\big)}{N} \cdot \vec{\mathbf{1}}_j. \tag{39}$$

Note that once a user is activated, its belief value will only transit to p_k or r_k, therefore $\eta_{ij}^h(\boldsymbol{z}) \neq 0$ only for $j \in \Theta := \{1, 2\tau+1, 2\tau+2, 2(2\tau+1)\}$. Also note that for those channels that stay idle, there is no randomness associated with its belief transition, i.e., for them $\eta_{ij}^h(\boldsymbol{z}) = q_{ij}(\boldsymbol{z}) \in \{0, 1\}$. Therefore the randomness is only associated with the channels which are activated, i.e., those with index value no smaller than ω^*. Hence, (39) becomes

$$\boldsymbol{Z}^N[t+1] - \big(I + Q(\boldsymbol{z})\big)\boldsymbol{z} = \sum_{j\in\Theta} \sum_{i\in\Pi_j(\boldsymbol{z})} \frac{\sum_{h=1}^{Ng_i(\boldsymbol{z})z_i} \big(\eta_{ij}^h(\boldsymbol{z}) - q_{ij}(\boldsymbol{z})\big)}{N} \cdot \vec{\mathbf{1}}_j,$$

where the summation $\sum_{h=1}^{Ng_i(\boldsymbol{z})z_i}(\cdot)$ is over all the channels in belief state β_i that are activated, and $\Pi_j(\boldsymbol{z})$ is the set of belief values in which channels are scheduled within the class that corresponds to belief $j \in \Theta$, i.e.,

$$\Pi_j(\boldsymbol{z}) := \begin{cases} \{1 \leq i \leq 2\tau+1 : g_i(\boldsymbol{z}) > 0\} & \text{if } j = 1, 2\tau+1, \\ \{(2\tau+1)+1 \leq i \leq 2(2\tau+1) : g_i(\boldsymbol{z}) > 0\} & \text{if } j = 2\tau+2, 2(2\tau+1). \end{cases}$$

For each $j \in \Theta$, we have

$$\begin{aligned} P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \big(I+Q(\boldsymbol{z})\big)\boldsymbol{z}\big\| \geq \mu \Big| Z^N[t] = \boldsymbol{z}\Big) &= P\Big(\big\|\sum_{j\in\Theta}\sum_{i\in\Pi_j(\boldsymbol{z})} \frac{\sum_{h=1}^{g_i(\boldsymbol{z})Nz_i} \big(\eta_{ij}^h(\boldsymbol{z}) - q_{ij}(\boldsymbol{z})\big)}{N} \cdot \vec{\mathbf{1}}_j\big\| > \mu\Big) \\ &\leq \sum_{j\in\Theta} P\Big(\Big|\sum_{i\in\Pi_j(\boldsymbol{z})} \sum_{h=1}^{g_i(\boldsymbol{z})Nz_i} \frac{\eta_{ij}^h(\boldsymbol{z}) - q_{ij}(\boldsymbol{z})}{N}\Big| > \frac{\mu}{4}\Big), \end{aligned} \tag{40}$$

where the last inequality holds because $\big|\Theta\big| = 4$ and also from union bound. Specifically, the union bound holds since

$$\{\Big\|\sum_{j\in\Theta}\sum_{i\in\Pi_j(\boldsymbol{z})} \frac{\sum_{h=1}^{g_i(\boldsymbol{z})Nz_i} \big(\eta_{ij}^h(\boldsymbol{z}) - q_{ij}(\boldsymbol{z})\big)}{N} \cdot \vec{\mathbf{1}}_j\Big\| > \mu\} \subseteq \bigcup_{j\in\Theta} \{\Big|\sum_{i\in\Pi_j(\boldsymbol{z})} \sum_{h=1}^{g_i(\boldsymbol{z})Nz_i} \frac{\eta_{ij}^h(\boldsymbol{z}) - q_{ij}(\boldsymbol{z})}{N}\Big| > \frac{\mu}{4}\}$$

From an extension of Chebychoff's inequality (See Excercise 1.8 in [21]) we have that, for each $j \in \Theta$, there exists a positive continuous function $f_j(\mu)$, which does not depend on $\boldsymbol{z}$ and N, with

$$P\Big(\Big|\sum_{i\in\Pi_j(\boldsymbol{z})} \sum_{h=1}^{g_i(\boldsymbol{z})Nz_i} \frac{\eta_{ij}^h(\boldsymbol{z}) - q_{ij}(\boldsymbol{z})}{N}\Big| > \frac{\mu}{4}\Big) < \exp\big(-f_j(\mu) \sum_{i\in\Pi_j(\boldsymbol{z})} g_i(\boldsymbol{z})Nz_i\big). \tag{41}$$

Let α_j be the fraction of channels activated, under the *steady state* of Optimal Relaxed Policy, in the class corresponding to belief value β_j, i.e.,

$$\alpha_j = \begin{cases} \sum_{i=1}^{2\tau+1} g_i(\boldsymbol{\zeta})\zeta_i & \text{if j=1, } 2\tau+1, \\ \sum_{i=2\tau+2}^{2(2\tau+1)} g_i(\boldsymbol{\zeta})\zeta_i & \text{if j=}2\tau+2,\ 2(2\tau+1). \end{cases} \tag{42}$$

For any $0 < \ell < \min\{\alpha_j, j \in \Theta\}$, there exists a neighborhood $\Omega_\varepsilon(\vec{\boldsymbol{\zeta}})$ such that for all $\boldsymbol{z} \in \Omega_\varepsilon(\vec{\boldsymbol{\zeta}})$,

$$\sum_{i \in \Pi_j(\boldsymbol{z})} g_i(\boldsymbol{z}) z_i \geq \alpha_j - \ell, \quad j \in \Theta, \tag{43}$$

which essentially means, under system state $\boldsymbol{z} \in \Omega_\varepsilon(\vec{\boldsymbol{\zeta}})$, the fraction of activated channels in each class will stay close to the case when system state is actually $\boldsymbol{\zeta}$. Let $f(\mu) = \min\{f_j(\mu)(\alpha_j - \ell), j \in \Theta\}$, then from (40)-(43),

$$\begin{aligned} P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \big(I + Q(\boldsymbol{z})\big)\boldsymbol{z}\big\| \geq \mu \Big| Z^N[t] = \boldsymbol{z}\Big) &\leq \sum_{j \in \Theta} P\Big(\Big| \sum_{i \in \Pi_j} \sum_{h=1}^{g_i(\boldsymbol{z})Nz_i} \frac{\eta_{ij}^h(\boldsymbol{z}) - q_{ij}(\boldsymbol{z})}{N}\Big| > \frac{\mu}{4}\Big) \\ &\leq 4\exp(-f(\mu)N). \end{aligned}$$

It is clear from the proof that $f(\mu)$ does not depend on $\boldsymbol{z}$ or N. Letting $c_1 = 4$ and $c_2 = f(\mu)$, the lemma thus holds. ■

Lemma 18. *There exists a neighborhood $\Omega_\delta(\boldsymbol{\zeta})$ of $\vec{\boldsymbol{\zeta}}$ for which, if $Z^N[0] = \boldsymbol{x} \in \Omega_\delta(\boldsymbol{\zeta})$, for any $t \geq 1$, there exist c_1^t and c_2^t with*

$$P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| \geq \mu\Big) \leq c_1^t \exp(-Nc_2^t),$$

where c_1^t and c_2^t are independent of $\boldsymbol{x}$ and N.

Proof: Recall that σ and ε are defined in Lemma 16 and Lemma 17, respectively. We let $\delta < \min\{\sigma, \varepsilon\}$ be such that, if $\boldsymbol{z}[0] \in \Omega_\delta(\boldsymbol{\zeta})$, $\boldsymbol{z}[t] \in \Omega_{\varepsilon-\rho}(\boldsymbol{\zeta})$ for all $t \geq 1$ where ρ is a constant with $0 \leq \rho \leq \varepsilon$ and satisfies

$$\big\|\big(Q(\boldsymbol{x}) + I\big)\boldsymbol{x} - \big(Q(\boldsymbol{y}) + I\big)\boldsymbol{y}\big\| \leq \nu, \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{Z} \text{ with } \|\boldsymbol{x} - \boldsymbol{y}\| \leq \rho. \tag{44}$$

for positive constant $\nu < \mu$. We proceed to prove this statement by induction.

For $t = 1$, if $\boldsymbol{x} \in \Omega_\delta(\boldsymbol{\zeta})$, from Lemma 17, there exist $c_1^1 > 0$ and $c_2^1 > 0$,

$$P_{\boldsymbol{x}}(\big\|\boldsymbol{Z}^N[1] - \boldsymbol{z}[1]\big\| \geq \mu\Big) = P_{\boldsymbol{x}}(\big\|\boldsymbol{Z}^N[1] - \big(I + Q(\boldsymbol{x})\big)\boldsymbol{x}\big\| \geq \mu\Big) \leq c_1^1 \exp(-c_2^1 N).$$

Suppose the statement is true at $t \geq 1$, then there exist d_1^t and d_2^t for which,

$$\begin{aligned} &P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \boldsymbol{z}[t+1]\big\| \geq \mu\Big) \\ =&P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| \geq \rho\Big) P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \boldsymbol{z}[t+1]\big\| \geq \mu \Big| \big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| \geq \rho\Big) \\ &\qquad + P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| < \rho\Big) P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \boldsymbol{z}[t+1]\big\| \geq \mu \Big| \big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| < \rho\Big) \\ \leq& d_1^t \exp(-d_2^t N) + P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \boldsymbol{z}[t+1]\big\| \geq \mu \Big| \big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| < \rho\Big) \end{aligned} \tag{45}$$

Now consider the second term in (45),

$$\begin{aligned} &P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \boldsymbol{z}[t+1]\big\| \geq \mu \Big| \ \big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| < \rho\Big) \\ =&P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \big(I + Q(\boldsymbol{Z}^N[t])\big)\boldsymbol{Z}^N[t] + \big(I + Q(\boldsymbol{Z}^N[t])\big)\boldsymbol{Z}^N[t] - \boldsymbol{z}[t+1]\big\| \geq \mu \Big| \ \big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| < \rho\Big) \\ \leq&P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \big(I+Q(\boldsymbol{Z}^N[t])\big)\boldsymbol{Z}^N[t]\big\| + \big\|\big(I+Q(\boldsymbol{Z}^N[t])\big)\boldsymbol{Z}^N[t] - \big(I+Q(\boldsymbol{z}[t])\big)\boldsymbol{z}[t]\big\| \geq \mu \Big| \ \big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| < \rho\Big) \\ \leq&P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \big(I+Q(\boldsymbol{Z}^N[t])\big)\boldsymbol{Z}^N[t]\big\| \geq \mu - \nu \Big| \ \big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| < \rho\Big). \\ =& \sum_{\boldsymbol{z} \in \Omega_\rho(\boldsymbol{z}[t])} P_{\boldsymbol{x}}\big(\boldsymbol{Z}^N[t] = \boldsymbol{z} \Big| \boldsymbol{Z}^N[t] \in \Omega_\rho(\boldsymbol{z}[t])\big) P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \big(I+Q(\boldsymbol{z})\big)\boldsymbol{z}\big\| \geq \mu - \nu \Big| \ \boldsymbol{Z}^N[t] = \boldsymbol{z}\Big) \end{aligned} \tag{46}$$

where the first inequality follows from triangle inequality, and the second inequality is from relationship (44).

Since $\Omega_\rho(\boldsymbol{z}[t]) \subseteq \Omega_\varepsilon(\boldsymbol{\zeta})$, from Lemma 17, for $\boldsymbol{z} \in \Omega_\rho(\boldsymbol{z}[t])$, there exist positive constants c_1 and c_2, that do not depend on $\boldsymbol{z}$ or N, with

$$P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \big(I+Q(\boldsymbol{z})\big)\boldsymbol{z}\big\| \geq \mu - \nu\Big|\ \boldsymbol{Z}^N[t] = \boldsymbol{z}\Big) \leq c_1 \exp(-c_2 N). \tag{47}$$

Substituting (47) to (46), we have

$$P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \boldsymbol{z}[t+1]\big\| \geq \mu\Big|\ \big\|\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]\big\| < \rho\Big) \leq c_1 \exp(-c_2 N). \tag{48}$$

Hence from Equation (45) and (48), there exist constants $c_1^{t+1} > 0$ and $c_2^{t+1} > 0$ that do not depend on $\boldsymbol{z}$ and N with

$$P_{\boldsymbol{x}}\Big(\big\|\boldsymbol{Z}^N[t+1] - \boldsymbol{z}[t+1]\big\| \geq \mu\Big) \leq c_1^{t+1} \exp(-N c_2^{t+1}).$$

By induction, the lemma holds. ■

Note that from union bound,

$$P_{\boldsymbol{x}}\Big(\sup_{0\leq t<T} ||\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]|| \geq \mu\Big) \leq \sum_{t=0}^{T-1} P_{\boldsymbol{x}}\Big(||\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]|| \geq \mu\Big). \tag{49}$$

Therefore, from Lemma 18, over finite time horizon T, there exist positive constants C_1 and C_2, which do not depend on $\boldsymbol{x}$ and N, such that

$$P_{\boldsymbol{x}}\Big(\sup_{0\leq t<T} ||\boldsymbol{Z}^N[t] - \boldsymbol{z}[t]|| \geq \mu\Big) \leq C_1 \exp(-N C_2),$$

which concludes the proof of Proposition 4. ■

According to Proposition 4 we have just established, the system state $\boldsymbol{Z}^N[t]$ behaves very close to the fluid approximation model $\boldsymbol{z}[t]$ when the number of users N is large. Since we have shown the convergence of $\boldsymbol{z}[t]$ to $\vec{\boldsymbol{\zeta}}$ within $\Omega_\sigma(\vec{\boldsymbol{\zeta}})$ in Lemma 16, we are ready to establish the local convergence of the system state $\boldsymbol{Z}^N[t]$ to $\vec{\boldsymbol{\zeta}}$.

Lemma 19. *If $\boldsymbol{Z}^N[0] = \boldsymbol{x} \in \Omega_\delta(\vec{\boldsymbol{\zeta}})$, then for any $\mu > 0$ there exists a time T_0 such that for each $T > T_0$, there exist positive constants s_1 and s_2 with,*

$$P_{\boldsymbol{x}}\Big(\sup_{T_0\leq t<T} ||\boldsymbol{Z}^N[t] - \vec{\boldsymbol{\zeta}}|| \geq \mu\Big) \leq s_1 \exp(-N s_2).$$

Proof: We let $0 < \nu < \mu$. Noting that $\delta < \sigma$, from Lemma 16 we have, given $\boldsymbol{z}[0] = \boldsymbol{x} \in \Omega_\delta(\vec{\boldsymbol{\zeta}})$, there exists T_0 such that for all $t \geq T_0$.

$$\big\|\boldsymbol{z}[t] - \vec{\boldsymbol{\zeta}}\big\| \leq \nu.$$

From Proposition 4 we know that there exist positive constants s_1 and s_2 such that,

$$\begin{aligned}
P_{\boldsymbol{x}}\Big(\sup_{T_0\leq t<T} \big\|\boldsymbol{Z}^N[t]-\vec{\boldsymbol{\zeta}}\big\| \geq \mu\Big) &\leq P_{\boldsymbol{x}}(\sup_{T_0\leq t<T} \big\|\boldsymbol{Z}^N[t]-\boldsymbol{z}[t]\big\|+\big\|\boldsymbol{z}[t]-\vec{\boldsymbol{\zeta}}\big\| \geq \mu)\\
&\leq P_{\boldsymbol{x}}(\sup_{T_0\leq t<T} \big\|\boldsymbol{Z}^N[t]-\boldsymbol{z}[t]\big\| \geq \mu-\nu)\\
&\leq P_{\boldsymbol{x}}(\sup_{0\leq t<T} \big\|\boldsymbol{Z}^N[t]-\boldsymbol{z}[t]\big\| \geq \mu-\nu)\\
&\leq s_1 \exp(-N s_2).
\end{aligned}$$

Hence the lemma holds. ■

The previous lemma allows us to establish the local convergence result. Let $v : \mathcal{Z} \to \mathcal{R}$ be a mapping such that $v(\boldsymbol{z})$ represents the per-user average throughput under system state $\boldsymbol{z}$. Therefore, $Nv(\boldsymbol{Z}^N[t])$ is the immediate reward at time t and we also have $r(\boldsymbol{\gamma}, \alpha) = v(\vec{\boldsymbol{\zeta}})$.

For $\ell > 0$, we let $\mu > 0$ be such that for any $\boldsymbol{x} \in \mathcal{Z}$, if $\|\boldsymbol{x} - \vec{\boldsymbol{\zeta}}\| < \mu$, then

$$|v(\boldsymbol{x}) - v(\vec{\boldsymbol{\zeta}})| < \ell. \tag{50}$$

Note that the per-user instantaneous throughput $v(\boldsymbol{z}) \leq 1$. Therefore,

$$\begin{aligned}
\Big|\frac{R_T^{N_m}(\boldsymbol{\gamma},\alpha,\boldsymbol{x})}{N_m}-r(\boldsymbol{\gamma},\alpha)\Big| &= \Big|\frac{1}{N_mT}E\big[\sum_{t=0}^{T-1}N_m v(Z^{N_m}[t])\big]-r(\boldsymbol{\gamma},\alpha)\Big| \\
&= \Big|\frac{1}{T}\sum_{t=0}^{T_0-1}E\big[v(Z^{N_m}[t])-v(\vec{\boldsymbol{\zeta}})\big]+\frac{1}{T}\sum_{t=T_0}^{T-1}E\big[v(Z^{N_m}[t])-v(\vec{\boldsymbol{\zeta}})\big]\Big| \\
&\leq \Big|\frac{1}{T}\sum_{t=0}^{T_0-1}E\big[v(Z^{N_m}[t])-v(\vec{\boldsymbol{\zeta}})\big]\Big|+\Big|\frac{1}{T}\sum_{t=T_0}^{T-1}E\big[v(Z^{N_m}[t])-v(\vec{\boldsymbol{\zeta}})\big]\Big| \\
&\leq \frac{T_0}{T}+\frac{1}{T}\sum_{t=T_0}^{T-1}E\big[\big|v(Z^{N_m}[t])-v(\vec{\boldsymbol{\zeta}})\big|\big]. \qquad (51)
\end{aligned}$$

Letting A_{N_m} be the event $\{\sup_{T_0\leq t\leq T}||\boldsymbol{Z}^{N_m}[t]-\vec{\boldsymbol{\zeta}}||\geq\mu\}$, we proceed to bound the second term in (51),

$$\begin{aligned}
&\frac{1}{T}\sum_{t=T_0}^{T-1}E\Big[\big|v(Z^{N_m}[t])-v(\vec{\boldsymbol{\zeta}})\big|\Big] \\
=&P_{\vec{\boldsymbol{x}}}(A_{N_m})\frac{1}{T}\sum_{t=T_0}^{T-1}E\Big[\big|v(Z^{N_m}[t])-v(\vec{\boldsymbol{\zeta}})\big|\Big|A_{N_m}\Big]+\big(1-P_{\vec{\boldsymbol{x}}}(A_{N_m})\big)\frac{1}{T}\sum_{t=T_0}^{T-1}E\Big[\big|v(Z^{N_m}[t])-v(\vec{\boldsymbol{\zeta}})\big|\Big|\bar{A}_{N_m}\Big] \\
\leq&P_{\vec{\boldsymbol{x}}}(A_{N_m})+(1-P_{\vec{\boldsymbol{x}}}(A_{N_m}))\ell \\
=&P_{\vec{\boldsymbol{x}}}(A_{N_m})(1-\ell)+\ell.
\end{aligned}$$

where the inequality if from the fact $v(\boldsymbol{z})\leq 1$ and the relation (50).

According to Lemma 19, when $\boldsymbol{x}\in\Omega_\delta(\vec{\boldsymbol{\zeta}})$, we have $\lim_{m\to\infty}P_{\vec{\boldsymbol{x}}}(A_{N_m})=0$, therefore,

$$\lim_{m\to\infty}\Big|\frac{R_T^{N_m}(\boldsymbol{\gamma},\alpha,\boldsymbol{x})}{N_m}-r(\boldsymbol{\gamma},\alpha)\Big|\leq\frac{T_0}{T}+\ell.$$

Since ℓ can be arbitrarily small, we have

$$\lim_{m\to\infty}|\frac{R_T^{N_m}(\boldsymbol{\gamma},\alpha,\boldsymbol{x})}{N_m}-r(\boldsymbol{\gamma},\alpha)|\leq\frac{T_0}{T}.$$

Hence, taking limit with T in both sides,

$$\lim_{T\to\infty}\lim_{m\to\infty}\frac{R_T^{N_m}(\boldsymbol{\gamma},\alpha,\boldsymbol{x})}{N_m}=r(\boldsymbol{\gamma},\alpha).$$

We have thus proved Proposition 2.

Appendix F
Proof of Lemma 4

(i) Here we prove the Markov chain has one unique class by stating that, starting from any state, there exists a possibility to reach a particular state, and hence there is only one class of recurrent state.

Case (1). Suppose $\alpha\leq\gamma_1$. Starting from any initial state $\boldsymbol{Z}^N[0]$, the following transition can occur: whenever the channels in class 1 are activated, their states are observed to be in ON state, and whenever channels in class 2 are activated, they are revealed to be in OFF state. Then after a long enough time duration t_1, α fraction of channels, which are in class 1, will be in belief value p_1, and other channels will have stationary belief value π_s. Hence the system state will be $\boldsymbol{Z}^N[t_1]=[\boldsymbol{Z}^{1,N}[t_1],\boldsymbol{Z}^{2,N}[t_1]]$ (defined in Section V-A) with $Z_{1,1}^{1,N}[t_1]=\alpha$, $Z_s^{1,N}[t_1]=\gamma_1-\alpha$, $Z_s^{2,N}[t_1]=\gamma_2$, and with 0 in all other positions.

Case (2). Suppose $\alpha>\gamma_1$. Starting from any initial state $\boldsymbol{Z}^N[0]$, consider the following transition path. Within the first period of time slots, $0\leq t\leq t_0$, whenever users in class 1 are activated, they turn out to be in state 1, and whenever users in class 2 are activated, they turn out to be in state 0. Then if t_0 is long enough, $\boldsymbol{Z}^{1,N}[t_0]$

is such that $Z^{1,N}_{1,1}[t_0] = \gamma_1$, with zero in all other elements. In the second period, $t_0 \leq t \leq t_1$, whenever users in class 1 are activated, it will remain in state 1, and whenever users in class 2 are activated, it turns out to be in state 1 as well. Then after long enough of time until t_1, $\boldsymbol{Z}^N[t_1] = [\boldsymbol{Z}^{1,N}[t_1], \boldsymbol{Z}^{2,N}[t_1]]$ with $Z^{1,N}_{1,1}[t_1] = \gamma_1$, $Z^{2,N}_{1,1}[t_1] = \alpha - \gamma_1$, and $Z^{2,N}_{s}[t_1] = 1 - \alpha$, with zero in all other elements .

Since the state space of the Markov Chain $\boldsymbol{Z}^N[t]$ is finite, there is at least one recurrent class. As we have seen in the above cases that, starting from all states, $\boldsymbol{Z}^N[t]$ can reach a particular state. Therefore there can only be one recurrent state. We shall henceforth denote this particular state as $\boldsymbol{Z}^N_p$. It is also clear from the proof that the Markov chain is aperiodic because of the possible self-transition in state $\boldsymbol{Z}^N_p$.

(ii) Similar to the proof of Proposition 2, in this part, we drop the suffix α and $\boldsymbol{\gamma}$ in the notation $\vec{\boldsymbol{\zeta}}^{\alpha}_{\boldsymbol{\gamma}}$, and we assume $W_k(b^2_{0,h_2^*-1}) < W_1(b^1_{0,h_1^*}) = \omega^* < W_k(b^2_{0,h_2^*})$. Recall that from the expression 20 of Whittle's index value that $W_k(\pi) = W_k(b^k_s)$ for $\pi \in \mathcal{B}_k$, $\pi \geq b^k_s$, $k = 1, 2$. We first characterize the structure of $\vec{\boldsymbol{\zeta}}$. From the description in Lemma 1 we know that the non-zero elements of $\vec{\boldsymbol{\zeta}}$ are

$$\zeta^1_0 := \zeta^1_{0,1} = \zeta^1_{0,2} = \cdots = \zeta^1_{0,h_1^*}, \quad \zeta^1_{0,h_1^*+1} = (1-\rho^*)\zeta^1_{0,h_1^*}, \quad \zeta^1_{1,1} = 1 - \sum_{h=1}^{h_1^*+1} \zeta^1_{0,h},$$
$$\zeta^2_0 := \zeta^2_{0,1} = \zeta^2_{0,2} = \cdots = \zeta^2_{0,h_2^*-1} = \zeta^2_{0,h_2^*}, \qquad \zeta^2_{1,1} = 1 - \sum_{h=1}^{h_2^*} \zeta^2_{0,h}.$$

We shall proceed to construct a path from the state $\boldsymbol{Z}^N_p$ to an arbitrary neighborhood of $\vec{\boldsymbol{\zeta}}$. For ease of exposition, in the proof we no longer consider the channels as unsplittable entities. Instead, the transition in the each stages deals with belief state evolution of certain *fraction* of users. As we shall see, under this assumption, we can construct a transition path of $\boldsymbol{Z}^N[t]$ under the Whittle's Index Policy, that transits from $\boldsymbol{Z}^N_p$ to the *exact* value $\vec{\boldsymbol{\zeta}}$. Although the identified path may not be feasible in reality for small value of N, but as the number of users N increases, we can find a transition path, which operates each user as unsplittable entities, that is arbitrarily close to this identified path, and thus can ultimately get arbitrarily close to any neighborhood of $\vec{\boldsymbol{\zeta}}$.

Note that when $\boldsymbol{Z}^N[t_1] = \boldsymbol{Z}^N_p$, $\boldsymbol{Z}^N[t_1] = \left[\boldsymbol{Z}^{1,N}[t_1], \boldsymbol{Z}^{2,N}[t_1]\right]$, where

$$Z^{1,N}_{1,1}[t_1] + Z^{1,N}_{s}[t_1] = \gamma_1, \quad \text{and} \quad Z^{2,N}_{1,1}[t_1] + Z^{2,N}_{s}[t_1] = \gamma_2.$$

In the following construction we shall assume that belief values are updated at the end of each slot when the actual channel states are revealed.

Case (1). Suppose $h_1^* \geq h_2^*$ and $W_k(b^1_s) \geq W_k(b^2_s)$. We shall denote $h'_1 = \max\{l : W^1(b^1_{0,l}) \leq W^2(b^2_s)\}$. In this case, The path is constructed with the stages below, starting from state $\boldsymbol{Z}^N[t_1] = \boldsymbol{Z}^N_p$.

Stage 1.1. In the first slot, among the α fraction activated channels, $\alpha - \zeta^1_{0,h_1^*+1}$ amount remains in ON state, and ζ^1_{0,h_1^*+1} amount turn out in OFF state and are in class 1. Hence the end of this slot, $\boldsymbol{Z}^N = [\boldsymbol{Z}^{1,N}, \boldsymbol{Z}^{2,N}]$ has the following non-zero elements

$$Z^{1,N}_{0,1} = \zeta^1_{0,h_1^*+1}, \quad Z^{1,N}_{1,1} + Z^{1,N}_{s} = \gamma_1 - \zeta^1_{0,h_1^*+1}, \quad Z^{2,N}_{1,1} + Z^{2,N}_{s} = \gamma_2.$$

Stage 1.2. In each of the next h_1^* slots, $\alpha - \zeta^1_0$ amount in the activated channels turn out in ON state, and ζ^1_0 amount of them turn out to be in OFF state and are in class 1. So at the end of the last slot of this stage, the non-zero elements of the system state $\boldsymbol{Z}^N = [\boldsymbol{Z}^{1,N}, \boldsymbol{Z}^{2,N}]$ satisfies

$$Z^{1,N}_{0,1} = Z^{1,N}_{0,2} = \cdots = Z^{1,N}_{0,h_1^*} = \zeta^1_0, \; Z^{1,N}_{0,h_1^*+1} = \zeta^1_{0,h_1^*+1}, \; Z^{1,N}_{1,1} = \zeta^1_{1,1}, \; Z^{2,N}_{1,1} + Z^{2,N}_{s} = \gamma_2.$$

Stage 2. In the next few slots, all activated channels turn out to be in state 1. This stage goes on for $h'_1 - h_1^*$ slots, until those channels that reach belief state $b^1_{0,1}$ at the end of stage 1.1 are in belief state b^1_{0,h'_1+1}. Then by the end of the last slot of this stage, the non-zero elements of the system state $\boldsymbol{Z}^N$ satisfies

$$Z^{1,N}_{0,h'_1-h_1^*+1} = \cdots = Z^{1,N}_{0,h'_1} = \zeta^1_0, \; Z^{1,N}_{0,h'_1+1} = \zeta^1_{0,h_1^*+1}, \; Z^{1,N}_{1,1} = \zeta^1_{1,1}, \; Z^{2,N}_{1,1} + Z^{2,N}_{s} = \gamma_2.$$

Stage 3. In each of the following slots, among all channel activated, only those in belief state $b^1_{0,h_1'+1}$ turn out to be in OFF state. This stage goes on until those channels that transit to belief state $b^1_{0,h_1'}$ in stage 2 reaches belief state $b^1_{0,h_1^*-h_2^*+1}$. Hence by the end of the final slot of this stage,

$$Z^{1,N}_{0,1} = \cdots = Z^{1,N}_{0,h_1^*-h_2^*} = \zeta^1_0, Z^{1,N}_{0,h_1^*-h_2^*+1} = \zeta^1_{0,h_1^*+1}, Z^{1,N}_{0,h_1'-h_2^*+2} = \cdots = Z^{1,N}_{0,h_1'+1} = \zeta^1_0, Z^{2,N}_{1,1} + Z^{2,N}_s = \gamma_2.$$

Stage 4. In each of the next h_2^* slots, among all users activated, those in belief state $b^1_{0,h_1'+1}$ turn out to be in OFF state, and ζ^2_0 amount of activated channels in class 2 turn out in OFF state. Then by the end of the final slot in this stage, the system state will be $\boldsymbol{Z}^N = \vec{\boldsymbol{\zeta}}$, i.e.,

$$\begin{aligned} Z^{1,N}_{0,1} =& Z^{1,N}_{0,2} = \cdots = Z^{1,N}_{0,h_1^*} = \zeta^1_0, \quad Z^{1,N}_{0,h_1^*+1} = \zeta^1_{0,h_1^*+1}, \quad Z^{1,N}_{1,1} = \zeta^1_{1,1} \\ Z^{2,N}_{0,1} =& Z^{2,N}_{0,2} = \cdots = Z^{2,N}_{0,h_2^*-1} = Z^{2,N}_{0,h_2^*} = \zeta^2_0, \quad Z^{2,N}_{1,1} = \zeta^2_{1,1}. \end{aligned}$$

Case (2). Suppose $W_k(b^1_s) \geq W_k(b^2_s)$ and $h_1^* \leq h_2^*$. We shall let $h_1' = \max\{l : W^1(b^1_{0,l}) \leq W^2(b^2_s)\}$ and $d = \lfloor h_2^*/(h_1'+1) \rfloor$. Starting from state $\boldsymbol{Z}^N[t_1] = \boldsymbol{Z}^N_p$, the path is constructed with the stages below, where stage 1.1 and 1.2 are the same with the previous case.

Stage 1.1. In the first slot, among the α fraction of activated channels, only ζ^1_{0,h_1^*+1} amount turn out in OFF state and they are in class 1. Therefore at the end of this slot, $\boldsymbol{Z}^N = [\boldsymbol{Z}^{1,N}, \boldsymbol{Z}^{2,N}]$ with non-zero elements being

$$Z^{1,N}_{0,1} = \zeta^1_{0,h_1^*+1}, \quad Z^{1,N}_{1,1} + Z^{1,N}_s = \gamma_1 - \zeta^1_{0,h_1^*+1}, \quad Z^{2,N}_{1,1} + Z^{2,N}_s = \gamma_2.$$

Stage 1.2. In each of the next h_1^* slots, $\alpha - \zeta^1_0$ amount of activated channels are in state '1', and ζ^1_0 amount are in OFF state and are in class 1. Hence at the end of the last slot of this stage, the non-zero elements of $\boldsymbol{Z}^N = [\boldsymbol{Z}^{1,N}, \boldsymbol{Z}^{2,N}]$ satisfies

$$Z^{1,N}_{0,1} = Z^{1,N}_{0,2} = \cdots = Z^{1,N}_{0,h_1^*} = \zeta^1_0,\ Z^{1,N}_{0,h_1^*+1} = \zeta^1_{0,h_1^*+1},\ Z^{1,N}_{1,1} = \zeta^1_{1,1},\ Z^{2,N}_{1,1} + Z^{2,N}_s = \gamma_2.$$

Letting t_2 be the slot right after stage 1.2, the path proceeds as follows.

Stage 2.
(1) From slot t_2 to slot $t_2 + h_1' - h_1^* - 1$, all activated channels in class 1 turn out to be in state 1. Hence at the end of slot $t_2 + h_1' - h_1^* - 1$, the channels that reach belief state b^1_{0,h_1^*+1} at the end of stage 1.2 are in belief state $b^1_{0,h_1'+1}$. Next, from slot $t_2 + h_1' - h_1^*$ to slot $t_2 + (d+1)(h_1'+1) - 1$, among the activated channels in class 1, only those in belief state $b^1_{0,h_1'+1}$ turn out to be in OFF state. Therefore, at the end of slot $t_2 + (d+1)(h_1'+1) - 1$, the system state vector $\boldsymbol{Z}^{1,N}$ that correspond to class-1 channels is

$$Z^{1,N}_{0,1} = Z^{1,N}_{0,2} = \cdots = Z^{1,N}_{0,h_1^*} = \zeta^1_0, \quad Z^{1,N}_{0,h_1^*+1} = \zeta^1_{0,h_1^*+1}, \quad Z^{1,N}_{1,1} = \zeta^1_{1,1}.$$

(2) In the meanwhile, from slot $t_2 + (d+1)(h_1'+1) - h_2^* - 1$ to slot $t_2 + (d+1)(h_1'+1) - 1$, among the activated channels in class 2, ζ^2_0 amount turn out to be in OFF state. Hence by the end of slot $t_2 + (d+1)(h_1'+1) - 1$, the vector $\boldsymbol{Z}^{1,N}$ that correspond to class-2 channels is

$$Z^{2,N}_{0,1} = Z^{2,N}_{0,2} = \cdots = Z^{2,N}_{0,h_2^*-1} = Z^{2,N}_{0,h_2^*} = \zeta^2_0, \quad Z^{2,N}_{1,1} = \zeta^2_{1,1}.$$

Therefore, at the end of slot $t_2 + (d+1)(h_1'+1) - 1$, $\boldsymbol{Z}^N = \vec{\boldsymbol{\zeta}}$.

Appendix G
Proof of Lemma 5

The proof is a discrete-time version of the proof of Theorem 6.89 from [21]. We first present a lemma which is an extension of Lemma 19.

Lemma 20. *There is a neighborhood $\Omega_\vartheta(\vec{\boldsymbol{\zeta}}^\alpha_\gamma)$ of $\vec{\boldsymbol{\zeta}}^\alpha_\gamma$, with $\vartheta < \delta$, for which if $\boldsymbol{Z}^N[0] = \boldsymbol{x} \in \Omega_\vartheta(\vec{\boldsymbol{\zeta}}^\alpha_\gamma)$, then for any $\mu > 0$ and time T, there exist positive constants ρ_1 and ρ_2 with,*

$$P_{\boldsymbol{x}}\Big(\sup_{0 \leq t < T} \|\boldsymbol{Z}^N[t] - \vec{\boldsymbol{\zeta}}^\alpha_\gamma\| \geq \mu \Big) \leq \rho_1 \exp(-N\rho_2)$$

where ρ_1 *and* ρ_2 *are independent of* $\boldsymbol{x}$ *and* N.

Proof: Note that we have established, in Lemma 16, the local convergence of the fluid approximation model $\boldsymbol{z}[t]$ in a neighborhood $\Omega_\sigma(\vec{\boldsymbol{\zeta}}_\gamma^\alpha)$. We let $\nu < \mu$ and let $\vartheta < \delta$ (recall that δ is defined in proposition 4 with $\delta < \sigma$) be such that if $\boldsymbol{z}[0] \in \Omega_\vartheta(\vec{\boldsymbol{\zeta}}_\gamma^\alpha)$, then

$$\|\boldsymbol{z}[t] - \vec{\boldsymbol{\zeta}}_\gamma^\alpha\| \le \nu, \quad \forall t \ge 0.$$

From Proposition 4, there exist positive constants ρ_1 and ρ_2 with,

$$\begin{aligned}P_{\boldsymbol{x}}\Big(\sup_{0\le t<T}\big\|\boldsymbol{Z}^N[t]-\vec{\boldsymbol{\zeta}}\big\| \ge \mu\Big) \le & P_{\boldsymbol{x}}(\sup_{0\le t<T}\big\|\boldsymbol{Z}^N[t]-\boldsymbol{z}[t]\big\|+\big\|\boldsymbol{z}[t]-\vec{\boldsymbol{\zeta}}\big\| \ge \mu)\\ \le & P_{\boldsymbol{x}}(\sup_{0\le t<T}\big\|\boldsymbol{Z}^N[t]-\boldsymbol{z}[t]\big\| \ge \mu-\nu)\\ \le & P_{\boldsymbol{x}}(\sup_{0\le t<T}\big\|\boldsymbol{Z}^N[t]-\boldsymbol{z}[t]\big\| \ge \mu-\nu)\\ \le & \rho_1\exp(-N\rho_2),\end{aligned}$$

which proves the lemma. ■

We let $\epsilon_s < \vartheta$ be such that if $\boldsymbol{z}[0] \in \Omega_{\epsilon_s}(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$, then $\boldsymbol{z}[t] \in \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$ for $t \ge 0$.

We let ϱ_{2n}^N, $n = 0, 1, \cdots$ be the time slots of *consecutive* hitting times into the neighborhood $\Omega_{\epsilon_s}(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$ from *outside* of the neighborhood when the total number of users is N. Similarly, we let ϱ_{2n+1}^N, $n = 0, 1, \cdots$ denote the time slots of *exiting* the neighborhood $\Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$ from inside of the neighborhood, when the total number of users is N. Hence $\boldsymbol{y}_n = \boldsymbol{Z}^N[\varrho_n^N]$, $n = 0, 1, \cdots$ evolves as a Markov chain. In steady state,

$$P\big(\boldsymbol{Z}^N[\infty] \notin \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)\big) \le \frac{E[\varrho_{2n+2}^N - \varrho_{2n+1}^N]}{E[\varrho_{2n+2}^N - \varrho_{2n}^N]} = \frac{E[\varrho_{2n+2}^N - \varrho_{2n+1}^N]}{E[\varrho_{2n+2}^N - \varrho_{2n+1}^N] + E[\varrho_{2n+1}^N - \varrho_{2n}^N]}. \tag{52}$$

We let $T_\epsilon(N)$ denote the random variable $\varrho_{2n+1}^N - \varrho_{2n}^N$. For any constant $K > 0$, we have

$$\begin{aligned}E[T_\epsilon(N)] &= \sum_{t=1}^{\infty} t \cdot P(T_\epsilon(N) = t)\\ &\ge 2K \cdot P(T_\epsilon(N) \ge 2K)\\ &= 2K \cdot P_{\boldsymbol{Z}^N[\varrho_{2n+1}^N]}\Big(\sup_{\varrho_{2n+1}^N \le t < \varrho_{2n+1+2K}^N}\|\boldsymbol{Z}^N[t] - \vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha\| \le \epsilon\Big).\end{aligned} \tag{53}$$

Note that

$$\begin{aligned}&P_{\boldsymbol{Z}^N[\varrho_{2n+1}^N]}\Big(\sup_{\varrho_{2n+1}^N \le t < \varrho_{2n+1}^N+2K}\|\boldsymbol{Z}^N[t] - \vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha\| > \epsilon\Big)\\ =& \sum_{\boldsymbol{z}\in\Omega_{\epsilon_s}(\vec{\boldsymbol{\zeta}}_\gamma^\alpha)} P\big(\boldsymbol{Z}^N(\varrho_1^N) = \boldsymbol{z}\big)P_{\boldsymbol{z}}\Big(\sup_{0\le t<2K}\|\boldsymbol{Z}^N[t] - \vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha\| > \epsilon\Big).\end{aligned} \tag{54}$$

Since $\epsilon_s < \vartheta$, from Lemma 20, there exist positive constants ς_1 and ς_2 such that for any $\boldsymbol{z} \in \Omega_{\epsilon_s}(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha)$,

$$P_{\boldsymbol{z}}\Big(\sup_{0\le t<2K}\|\boldsymbol{Z}^N[t] - \vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha\| > \epsilon\Big) \le \varsigma_1\exp(-\varsigma_2 N). \tag{55}$$

Substitute (55) in (54) we have

$$P_{\boldsymbol{Z}^N[\varrho_{2n+1}^N]}\Big(\sup_{\varrho_{2n+1}^N \le t < \varrho_{2n+1}^N+2K}\|\boldsymbol{Z}^N[t] - \vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha\| > \epsilon\Big) \le \varsigma_1\exp(-\varsigma_2 N).$$

Therefore, $P_{\boldsymbol{Z}^{N_m}[\varrho_{2n+1}^{N_m}]}\Big(\sup_{\varrho_{2n+1}^{N_m} \le t < \varrho_{2n+1+2K}^{N_m}}\|\boldsymbol{Z}^{N_m}[t] - \vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^\alpha\| \le \epsilon\Big) \to 1$ as $m \to \infty$. From (53), if m is large enough, we have

$$E[\varrho_1^{N_m} - \varrho_0^{N_m}] \ge K.$$

Since K can be arbitrarily large, $\lim_{m\to\infty} E[\varrho_1^{N_m} - \varrho_0^{N_m}] = \infty$. Since from Assumption Ψ we know $E[\varrho_2^{N_m} - \varrho_1^{N_m}] \leq M_{\epsilon_s}$, thus from equation (52),

$$\lim_{m\to\infty} P\big(\boldsymbol{Z}^{N_m}[\infty] \notin \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\big) = 0,$$

which concludes the proof.

Appendix H
Proof of Proposition 3

For any $\ell > 0$, let $\epsilon > 0$ be such that for $\boldsymbol{x} \in \mathcal{Z}$, if $||\boldsymbol{x} - \vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})|| < \epsilon$, then

$$|v(\boldsymbol{x}) - r(\boldsymbol{\gamma}, \alpha)| < \ell.$$

Consider fixed N_m, for $\forall \ell > 0$ denote event $E_{N_m} = \{\boldsymbol{Z}^{N_m}[\infty] \in \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\}$, then

$$\begin{aligned}
&\Big|\frac{R_{\boldsymbol{x}}^{N_m}(\boldsymbol{\gamma}, \alpha)}{N_m} - r(\boldsymbol{\gamma}, \alpha)\Big| \\
\leq& E\Big[\Big|v(Z^{N_m}[\infty]) - v(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\Big|\Big] \\
=& P\big(E_{N_m}\big) E\Big[\Big|v(Z^{N_m}[\infty]) - v(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\Big|\Big| E_{N_m}\Big] + P\big(\bar{E}_{N_m}\big) E\Big[\Big|v(Z^{N_m}[\infty]) - v(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\Big|\Big| \bar{E}_{N_m}\Big] \\
\leq& P\big(\boldsymbol{Z}^{N_m}[\infty] \in \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\big) \cdot \ell + P\big(\boldsymbol{Z}^{N_m}[\infty] \notin \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\big).
\end{aligned} \tag{56}$$

Apply Lemma 5 to (56) we have

$$\lim_{m\to\infty} \Big|\frac{R_{\boldsymbol{x}}^{N_m}(\boldsymbol{\gamma}, \alpha)}{N_m} - r(\boldsymbol{\gamma}, \alpha)\Big| \leq \lim_{m\to\infty} \Big[P\big(\boldsymbol{Z}^{N_m}[\infty] \in \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\big) \cdot \ell + P\big(\boldsymbol{Z}^{N_m}[\infty] \notin \Omega_\epsilon(\vec{\boldsymbol{\zeta}}_{\boldsymbol{\gamma}}^{\alpha})\big)\Big] = \ell.$$

Since ℓ can be arbitrary,

$$\lim_{m\to\infty} \frac{R_{\boldsymbol{x}}^{N_m}(\boldsymbol{\gamma}, \alpha)}{N_m} = r(\boldsymbol{\gamma}, \alpha),$$

which proves the proposition.

Appendix I
Proof of Lemma 15

After some calculation, the matrix U^* takes the form

$$U^* = \begin{bmatrix} \tilde{Q}^1(\boldsymbol{z}) & B \\ 0 & \tilde{Q}^2(\boldsymbol{z}) \end{bmatrix}.$$

where matrix B is expressed as

$$B = \begin{bmatrix}
0 & \cdots & 0 & b_{0,h_1^*}^1 - 1 & b_{0,h_1^*}^1 - 1 & \cdots & b_{0,h_1^*}^1 - 1 \\
\vdots & & \vdots & & & & \\
0 & \cdots & 0 & 1 & 1 & \cdots & 1 \\
\vdots & & \vdots & & & & \\
0 & \cdots & 0 & -b_{0,h_1^*}^1 & -b_{0,h_1^*}^1 & \cdots & -b_{0,h_1^*}^1
\end{bmatrix}$$

in which only the first, last and $h_1^* + 1^{th}$ row have non-zero elements, and for each row, non-zero terms start at the h_2^{*th} element.

The matrices $\tilde{Q}^1(\boldsymbol{z})$ and $\tilde{Q}^1(\boldsymbol{z})$ are expressed as,

$$\tilde{Q}^1(\boldsymbol{z}) = \begin{bmatrix} -1 & 0 & \cdots & 0 & b^1_{0,h_1^*} - b^1_{0,h_1^*+1} & b^1_{0,h_1^*} - b^1_{0,h_1^*+2} & \cdots & b^1_{0,h_1^*} - p_1 \\ 1 & -1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & 1 & -1 & & & & \\ -1 & \cdots & -1 & -1 & -1 & & & \\ & & & & & -1 & & \\ & & & & & & \ddots & \\ & & & & b^1_{0,h_1^*+1} - b^1_{0,h_1^*} & b^1_{0,h_1^*+1} - b^1_{0,h_1^*} & \cdots & -(1-p_1) - b^1_{0,h_1^*} \end{bmatrix}$$

$$\tilde{Q}^2(\boldsymbol{z}) = \begin{bmatrix} -1 & 0 & \cdots & 0 & 1-b^2_{0,h_2^*} & 1 - b^2_{0,h_2^*+1} & \cdots & 1 - p_2 \\ 1 & -1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & 1 & -1 & & & & \\ -1 & \cdots & -1 & -1 & -2 & -1 & \cdots & -1 \\ & & & & & -1 & & \\ & & & & & & \ddots & \\ & & & & b^2_{0,h_2^*} & b^2_{0,h_2^*+1} & \cdots & -(1-p_2) \end{bmatrix}.$$

We need the following lemma to proceed.

Lemma 21. *For any $l \in \mathbb{Z}^+$,*

$$(1-p_1) + b^1_{0,l} > (l-1)(b^1_{0,l+1} - b^1_{0,l}).$$

Proof: The proof is moved to Appendix J. ■

With this lemma, we proceed to characterize the eigen values of matrix $\boldsymbol{U}^*$, which are given by the solution to equation $\det(\boldsymbol{U}^* - \lambda I) = 0$, where

$$\det(\boldsymbol{U}^* - \lambda I) = \det \begin{bmatrix} \tilde{Q}^1(\boldsymbol{z}) - \lambda I & B \\ & \tilde{Q}^2(\boldsymbol{z}) - \lambda I \end{bmatrix} = \det \begin{bmatrix} \tilde{Q}^1(\boldsymbol{z}) - \lambda I & 0 \\ & \tilde{Q}^2(\boldsymbol{z}) - \lambda I \end{bmatrix},$$

where the second equality is from the property of block matrices. Therefore, we have

$$\det(\boldsymbol{U}^* - \lambda I) = \det(\tilde{Q}^1(\boldsymbol{z}) - \lambda I) \det(\tilde{Q}^2(\boldsymbol{z}) - \lambda I).$$

(1) We first study the characteristic polynomial $\det(\tilde{Q}^1(\boldsymbol{z}) - \lambda I)$. After some algebra we have

$$\begin{aligned} \det(\tilde{Q}^1(\boldsymbol{z}) - \lambda I) = (1+\lambda)^{2\tau - h_1^*} \Big[& [\lambda + (1-p_1) + b^1_{0,h_1^*}](1+\lambda)^{h_1^*-1} - \\ & (b^1_{0,h_1^*+1} - b^1_{0,h_1^*}) \big[1 + (1+\lambda) + (1+\lambda)^2 + \cdots + (1+\lambda)^{h_1^*-2}\big]\Big] \\ \triangleq (1+\lambda)^{2\tau - h_1^*} \chi_1(\lambda). & \end{aligned}$$

where

$$\chi_1(\lambda) = [\lambda + (1-p_1) + b^1_{0,h_1^*}](1+\lambda)^{h_1^*-1} - (b^1_{0,h_1^*+1} - b^1_{0,h_1^*}) \big[1 + (1+\lambda) + (1+\lambda)^2 + \cdots + (1+\lambda)^{h_1^*-2}\big].$$

Consider the equation $\chi_1(\lambda) = 0$, i.e.,

$$[\lambda + (1-p_1) + b^1_{0,h_1^*}](1+\lambda)^{h_1^*-1} = (b^1_{0,h_1^*+1} - b^1_{0,h_1^*}) \big[1 + (1+\lambda) + (1+\lambda)^2 + \cdots + (1+\lambda)^{h_1^*-2}\big]. \tag{57}$$

Clearly, matrix $\tilde{Q}^1(\boldsymbol{z})$ has eigen value -1 of multiplicity $2\tau - h_1^*$. Let λ be any other eigen value of $\tilde{Q}^1(\boldsymbol{z})$, we proceed to show that $\big|\lambda + 1\big| < 1$.

We prove this by contradiction, suppose λ is such that $\big|\lambda+1\big| \geq 1$. Then taking modulus of the left hand side of equation (57) we have

$$\begin{aligned}
\big|[\lambda{+}(1{-}p_1){+}b^1_{0,h_1^*}](1{+}\lambda)^{h_1^*-1}\big| &= \big|\lambda+1-p+b^1_{0,h_1^*}\big| \cdot \big|1+\lambda\big|^{h_1^*-1} \\
&\geq \left(\big|\lambda+1\big| + \big|-p_1+b^1_{0,h_1^*}\big|\right)\big|1+\lambda\big|^{h_1^*-1} \\
&\geq \Big(1-p_1+b^1_{0,h_1^*}\Big)\big|1+\lambda\big|^{h_1^*-1},
\end{aligned}$$

where the first equality is from triangle inequality. Applying Lemma 21 we have,

$$\begin{aligned}
&\Big(1-p_1+b^1_{0,h_1^*}\Big)\big|1+\lambda\big|^{h_1^*-1} \\
>&(h_1^*-1)(b^1_{0,h_1^*+1}-b^1_{0,h_1^*})\cdot\big|1+\lambda\big|^{h_1^*-1} \\
>&(b^1_{0,h_1^*+1}-b^1_{0,h_1^*})\big[1+\big|1+\lambda\big|+\cdots+\big|1+\lambda\big|^{h_1^*-2}\big] \\
\geq&(b^1_{0,h_1^*+1}-b^1_{0,h_1^*})\big|1{+}(1{+}\lambda){+}\cdots{+}(1{+}\lambda)^{h_1^*-2}\big|.
\end{aligned} \tag{58}$$

where the first inequality is from Lemma 21, and the second inequality is from the fact that $\big|\lambda+1\big|{>}1$, and the last inequality comes from Triangle Inequality. Note that inequality (58) contradicts (57). Therefore each eigen values of matrix $\tilde{Q}^1(\boldsymbol{z})$ must satisfy $\big|\lambda+1\big| < 1$.

(2) We then study the characteristic polynomial $\det(\tilde{Q}^2(\boldsymbol{z})-\lambda I)$. We derive that

$$\begin{aligned}
&\det(\tilde{Q}^2(\boldsymbol{z})-\lambda I) \\
=&(1+\lambda)^{2\tau-h_2^*}\Big[\big[(1-p_2)+(1-b^2_{0,h_2^*})\lambda\big]\Big[1+(1+\lambda)+\cdots+(1+\lambda)^{h_2^*-3}\Big]+ \\
&\qquad\qquad (1+\lambda)^{h_2^*-2}\Big[\big[(1-p_2)+\lambda\big](2+\lambda)+b^2_{0,h_2^*}\Big]\Big] \\
\triangleq&(1+\lambda)^{2\tau-h_2^*}\cdot\chi_2(\lambda),
\end{aligned} \tag{59}$$

where

$$\chi_2(\lambda){=}\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]\Big[1{+}(1{+}\lambda){+}\cdots{+}(1{+}\lambda)^{h_2^*-3}\Big]{+}(1{+}\lambda)^{h_2^*-2}\Big[\big[(1{-}p_2){+}\lambda\big](2{+}\lambda){+}b^2_{0,h_2^*}\Big]$$

and consider

$$\begin{aligned}
\lambda\cdot\chi_2(\lambda)&{=}\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]\lambda\Big[1{+}(1{+}\lambda){+}\cdots{+}(1{+}\lambda)^{h_2^*-3}\Big]{+}(1{+}\lambda)^{h_2^*-2}\lambda\Big[\big[(1{-}p_2){+}\lambda\big](2{+}\lambda){+}b^2_{0,h_2^*}\Big] \\
&{=}\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big](1{+}\lambda{-}1)\Big[1{+}(1{+}\lambda){+}\cdots{+}(1{+}\lambda)^{h_2^*-3}\Big]{+}(1{+}\lambda)^{h_2^*-2}\lambda\Big[\big[(1{-}p_2){+}\lambda\big](2{+}\lambda){+}b^2_{0,h_2^*}\Big] \\
&{=}\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]\big[(1{+}\lambda)^{h_2^*-2}-1\big]{+}(1{+}\lambda)^{h_2^*-2}\lambda\Big[\big[(1{-}p_2){+}\lambda\big](2{+}\lambda){+}b^2_{0,h_2^*}\Big] \\
&=-\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]{+}(1{+}\lambda)^{h_2^*-2}\Big[\lambda\big[(1{-}p_2){+}\lambda\big](2{+}\lambda){+}b^2_{0,h_2^*}\lambda{+}\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]\Big] \\
&=-\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]{+}(1{+}\lambda)^{h_2^*-2}\Big[\lambda\Big[\big[(1{-}p_2){+}\lambda\big](2{+}\lambda){+}1\Big]+(1{-}p_2)\Big] \\
&=-\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]{+}(1{+}\lambda)^{h_2^*-2}\Big[\lambda\Big[(1{-}p_2)(2{+}\lambda){+}(\lambda+1)^2\Big]+(1{-}p_2)\Big] \\
&=-\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]{+}(1{+}\lambda)^{h_2^*-2}\Big[(1{-}p_2)(1{+}\lambda)^2{+}\lambda(\lambda+1)^2)\Big] \\
&=-\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]{+}(1{+}\lambda)^{h_2^*-2}\Big[(1{-}p_2{+}\lambda)(\lambda+1)^2\Big] \\
&=-\big[(1{-}p_2){+}(1{-}b^2_{0,h_2^*})\lambda\big]{+}(1{+}\lambda)^{h_2^*}(1{-}p_2{+}\lambda)
\end{aligned} \tag{60}$$

It is clear from equation (59) that matrix $\tilde{Q}^2(\boldsymbol{z})$ has eigen value -1 with multiplicity $2\tau-h_2^*$. Let λ be any eigen value of $\tilde{Q}^2(\boldsymbol{z})$, we first show the following lemma.

Lemma 22. *Let λ be any eigen value of $\tilde{Q}^2(\boldsymbol{z})$, then $-2 < Re(\lambda) < 0$.*

Proof: 1) Suppose $\tilde{Q}^2(\boldsymbol{z})$ has an eigen value of 0, then, from (59), $\chi_2(0)=0$. However,

$$\begin{aligned}\chi_2(0) &= (1{-}p_2)(h_2^*{-}2){+}2(1{-}p_2){+}b_{0,h_2^*}^2\\ &= h_2^*(1{-}p_2){+}b_{0,h_2^*}^2\\ &\neq 0,\end{aligned}$$

leading to a contradiction. Hence $\tilde{Q}^2(\boldsymbol{z})$ does not have 0 eigen value.

2) Suppose the equation $\chi_2(\lambda)=0$ has a root $\lambda^*=a+bi$ with $a>0$, or $a\leq -2$, or being purely imaginary with $a=0, b\neq 0$. Hence from equation (60),

$$(1{-}p_2){+}(1{-}b_{0,h_2^*}^2)\lambda^*{=}(1{+}\lambda^*)^{h_2^*}(1{-}p_2{+}\lambda^*) \tag{61}$$

Consider the modulus of the right hand side,

$$\begin{aligned}|(1{+}a+bi)^{h_2^*}|\cdot|1-p_2+a+bi| &> |1-p_2+a+bi|\\ &> |1-p_2+(1{-}b_{0,h_2^*}^2)(a+bi)|\\ &= |1-p_2+(1{-}b_{0,h_2^*}^2)\lambda^*|.\end{aligned}$$

The above expression contradicts the previous equation (61).

From 1) and 2) we conclude that $\chi_2(\lambda)=0$ can only have solution with real part within $(-2,0)$. Therefore all eigen values of matrix $\tilde{Q}^2(\boldsymbol{z})$ have real part within $(-2,0)$. ■

We proceed to show that each eigen value λ of $\tilde{Q}^2(\boldsymbol{z})$ needs to satisfy $\big|\lambda+1\big|<1$.

Suppose the equation $\chi_2(\lambda)=0$ has a root λ with $\big|\lambda+1\big|\geq 1$, then from equation (60),

$$(1{-}p_2){+}(1{-}b_{0,h_2^*}^2)\lambda{=}(1{+}\lambda)^{h_2^*}(1{-}p_2{+}\lambda) \tag{62}$$

We let $1+\lambda=x+yi$ where $x,y\in\mathbb{R}$, from the previous lemma we know that $|x|<1$. Some derivation shows that

$$\begin{aligned}&|(1{-}p_2{+}\lambda)|^2-|(1{-}p_2){+}(1{-}b_{0,h_2^*}^2)\lambda|^2\\ =&|1+\lambda|^2(2-b_{0,h_2^*}^2)b_{0,h_2^*}^2-2xb_{0,h_2^*}^2(1-p_2-b_{0,h_2^*}^2)+b_{0,h_2^*}^2(2p_2-b_{0,h_2^*}^2)\\ >&|x|(2-b_{0,h_2^*}^2)b_{0,h_2^*}^2-2|x|b_{0,h_2^*}^2(1-p_2-b_{0,h_2^*}^2)+|x|b_{0,h_2^*}^2(2p_2-b_{0,h_2^*}^2)\\ =&|x|b_{0,h_2^*}^2\Big[(2-b_{0,h_2^*}^2)-2(1-p_2-b_{0,h_2^*}^2)+(2p_2-b_{0,h_2^*}^2)\Big]\\ =&0.\end{aligned}$$

where the first inequality is from the assumption that $|1+\lambda|\geq 1$ and the fact that $|x|<1$. Therefore

$$\begin{aligned}|(1{-}p_2{+}\lambda)(1+\lambda)^{h_2^*}| &\geq |(1{-}p_2{+}\lambda)|\\ &> |(1{-}p_2){+}(1{-}b_{0,h_2^*}^2)\lambda|.\end{aligned}$$

The above expression contradicts the equation (62). Hence it can not be $\big|\lambda+1\big|\geq 1$. Therefore, each eigen value λ of $\boldsymbol{U}^*$ satisfies $\big|\lambda+1\big|<1$, concluding the proof.

Appendix J
Proof of Lemma 21

Proof: From the belief value evolution (1) we know

$$b_{0,l}^1=\frac{r_1-r_1(p_1-r_1)^l}{1+r_1-p_1}, \qquad b_{0,l+1}^1-b_{0,l}^1=r_1(p_1-r_1)^l.$$

Therefore

$$\begin{aligned}
&(1-p_1)+\pi^1_{0,l}-(l-1)(\pi^1_{0,l+1}-\pi^1_{0,l})\\
=&(1-p_1)+\frac{r_1-r_1(p_1-r_1)^l}{1+r_1-p_1}-(l-1)r_1(p_1-r_1)^l\\
=&(1-p_1)+\frac{r_1-r_1(p_1-r_1)^l}{1+r_1-p_1}-(l-1)r_1(p_1-r_1)^l\\
=&(1-p_1)+r_1\Big[\frac{1-(p_1-r_1)^l}{1+r_1-p_1}-(l-1)(p_1-r_1)^l\Big]\\
=&(1-p_1)+r_1\Big[\frac{1+(l-1)(p_1-r_1)^{l+1}-l(p_1-r_1)^l}{1+r_1-p_1}\Big]\\
=&(1-p_1)+r_1\Big[\frac{1+(l-1)(p_1-r_1)^l(p_1-r_1-1)-(p_1-r_1)^l}{1+r_1-p_1}\Big]\\
=&(1-p_1)+r_1\Big[\frac{(l-1)(p_1-r_1)^l(p_1-r_1-1)-(p_1-r_1-1)(1+(p_1-r_1)+\cdots+(p_1-r_1)^{l-1})}{1+r_1-p_1}\Big]\\
=&(1-p_1)+r_1\Big[(1+(p_1-r_1)+\cdots+(p_1-r_1)^{l-1})-(l-1)(p_1-r_1)^l\Big]
\end{aligned}\tag{63}$$

Since $(p_1-r_1)^j \geq (p_1-r_1)^l$ for $l=1,\cdots,j-1$, therefore from equation (63),

$$(1-p_1)+\pi^1_{0,l}-(l-1)(\pi^1_{0,l+1}-\pi^1_{0,l}) \geq (1-p_1)+r_1>0,$$

which proves the lemma. ■

APPENDIX K
DERIVATION OF INDEX VALUES

Here we derive the Whittle's indices according to Definition (19), by studying the relationship between the threshold value and the subsidy value.

(Case 1) $\pi = b^k_{0,l} < b^k_s$. We let $V(\omega, b^k_{0,l})$ denote the reward-plus-subsidy for the ω-subsidy problem when the threshold for activation is at $b^k_{0,l}$, i.e., the channel transmits when the belief is no smaller than $b^k_{0,l}$ and stays idle otherwise. Some algebra (of studying the steady state belief transition) shows that

$$V(\omega, b^k_{0,l}) = \frac{b^k_{0,l}+\omega(1-p_k)(l-1)}{b^k_{0,l}+(1-p_k)(l)}. \tag{64}$$

From the definition (19) of the Whittle's index value, it is equally optimal to activate or idle the channel with the belief value $b^k_{0,l}$ at the subsidy value $W_k(b^k_{0,l})$. From thresholdability, the belief value $b^k_{0,l}$ is at the boundary of the idle set $\mathcal{I}^k(W_k(b^k_{0,l}))$. Therefore the reward obtained by setting the threshold for activation at $b^k_{0,l}$ equals that with threshold $b^k_{0,l+1}$, i.e.,

$$V(W_k(b^k_{0,l}), b^k_{0,l}) = V(W_k(b^k_{0,l}), b^k_{0,l+1}),$$

where $V(W_k(b^k_{0,l}), b^k_{0,l})$ represents the reward corresponding to $a^*_{W_k(b^k_{0,l})}(b^k_{0,l}) = 1$, and $V(W_k(b^k_{0,l}), b^k_{0,l+1})$ represents the reward corresponding to $a^*_{W_k(b^k_{0,l})}(b^k_{0,l}) = 0$.

Substitute expression (64) in the previous relationship leads to the expression of the Whittle's index value,

$$W_k(b^k_{0,l}) = \frac{(b^k_{0,l}-b^k_{0,l+1})(l+1)+b^k_{0,l+1}}{1-p_k+(b^k_{0,l}-b^k_{0,l+1})l+b^k_{0,l+1}}, \tag{65}$$

which is the same as in [14].

(Case 2) $\pi \geq b^k_s$. In this case, we first present the following claim. This claim states that, if the threshold for activation is above b^k_s, then it is optimal to always stay idle.

Claim 1. *If $\theta_k(\omega) \geq b^k_s$, then $\mathcal{I}^k(\omega) = \mathcal{B}^k$.*

This claim is indeed true because, from Lemma 1, if $\theta_k(\omega) \geq b_s^k$, then eventually all users will be idle, hence it is optimal to always stay idle. Hence, for all belief states $\pi \geq b_s^k$, their Whittle's index value, according to the definition, equals to the infimum subsidy value for which the channel always staying idle. Note that $W_k(\pi)$ monotonically increases with π for $\pi < b_s^k$, therefore,

$$W_k(\pi) = \lim_{l \to \infty} W_k(b_{0,l}^k).$$

From (65) we get

$$W_k(\pi) = \frac{r_k}{(1-p_k)(1+r_k-p_k)+r_k}.$$