Technical Report 147

Augmenting Max-Weight with Explicit Learning for Wireless Scheduling with Switching Costs

Research Supervisor
Sanjay Shakkottai

Project Title: V2X Spectrum Resource Allocation for Sensing and Communications

September 2018

Data-Supported Transportation Operations & Planning Center (D-STOP)

A Tier 1 USDOT University Transportation Center at The University of Texas at Austin

D-STOP is a collaborative initiative by researchers at the Center for Transportation Research and the Wireless Networking and Communications Group at The University of Texas at Austin.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
D-STOP/2018/147		

4. Title and Subtitle	5. Report Date
Augmenting Max-Weight with Explicit Learning for Wireless Scheduling with Switching Costs	September 2018
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Subhashini Krishnasamy, Akhil P T, Ari Arapostathis, Rajesh Sundaresan, and Sanjay Shakkottai,	Report 147

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Data-Supported Transportation Operations & Planning Center (D-STOP) The University of Texas at Austin 3925 W. Braker Lane, 4th Floor Austin, Texas 78759	11. Contract or Grant No. DTRT13-G-UTC58

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Data-Supported Transportation Operations & Planning Center (D-STOP) The University of Texas at Austin 3925 W. Braker Lane, 4th Floor Austin, Texas 787591	14. Sponsoring Agency Code

15. Supplementary Notes

Supported by a grant from the U.S. Department of Transportation, University Transportation Centers Program.
Project Title: V2X Spectrum Resource Allocation for Sensing and Communications

16. Abstract

In small-cell wireless networks where users are connected to multiple base stations (BSs), it is often advantageous to switch off dynamically a subset of BSs to minimize energy costs. We consider two types of energy cost: (i) the cost of maintaining a BS in the active state, and (ii) the cost of switching a BS from the active state to inactive state. The problem is to operate the network at the lowest possible energy cost (sum of activation and switching costs) subject to queue stability. In this setting, the traditional approach — a Max-Weight algorithm along with a Lyapunov-based stability argument — does not suffice to show queue stability, essentially due to the temporal co-evolution between channel scheduling and the BS activation decisions induced by the switching cost. Instead, we develop a learning and BS activation algorithm with slow temporal dynamics, and a Max-Weight based channel scheduler that has fast temporal dynamics. We show using convergence of time- inhomogeneous Markov chains, that the co-evolving dynamics of learning, BS activation and queue lengths lead to near optimal average energy costs along with queue stability.

17. Key Words	18. Distribution Statement
wireless scheduling, base-station activation, energy minimization	No restrictions. This document is available to the public through NTIS (http://www.ntis.gov): National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	22	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgements

The authors recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers.

Augmenting Max-Weight with Explicit Learning for Wireless Scheduling with Switching Costs

Subhashini Krishnasamy, Akhil P T, Ari Arapostathis, *Fellow, IEEE,* Rajesh Sundaresan, *Senior Member, IEEE,* and Sanjay Shakkottai, *Fellow, IEEE*

***Abstract*—In small-cell wireless networks where users are connected to multiple base stations (BSs), it is often advantageous to switch off dynamically a subset of BSs to minimize energy costs. We consider two types of energy cost: (i) the cost of maintaining a BS in the active state, and (ii) the cost of switching a BS from the active state to inactive state. The problem is to operate the network at the lowest possible energy cost (sum of activation and switching costs) subject to queue stability. In this setting, the traditional approach — a Max-Weight algorithm along with a Lyapunov-based stability argument — does not suffice to show queue stability, essentially due to the temporal co-evolution between channel scheduling and the BS activation decisions induced by the switching cost. Instead, we develop a learning and BS activation algorithm with slow temporal dynamics, and a Max-Weight based channel scheduler that has fast temporal dynamics. We show using convergence of time-inhomogeneous Markov chains, that the co-evolving dynamics of learning, BS activation and queue lengths lead to near optimal average energy costs along with queue stability.**

***Index Terms*—wireless scheduling, base-station activation, energy minimization**

I. Introduction

Due to the tremendous increase in demand for data traffic, modern cellular networks have taken the densification route to support peak traffic demand [2]. While increasing the density of base-stations gives greater spectral efficiency, it also results in increased costs of operating and maintaining the deployed base-stations. Rising energy cost is a cause for concern, not only from an environmental perspective, but also from an economic perspective for network operators as it constitutes a significant portion of the operational expenditure. To address this challenge, latest research aims to design energy efficient networks that balance the trade-off between spectral efficiency, energy efficiency and user QoS requirements [3], [4].

Studies reveal that base-stations contribute to more than half of the energy consumption in cellular networks [5], [6]. Although dense deployment of base-stations are useful in meeting demand in peak traffic hours, they regularly have excess capacity during off-peak hours [4], [7]. A fruitful way to conserve power is, therefore, to dynamically switch off under-utilized base-stations. Even in networks that do not have fluctuations in traffic load, switching base-stations dynamically is a useful way to reduce power consumption while meeting the network traffic demand. For this purpose, modern cellular standards incorporate protocols that include *sleep* and *active* modes for base-stations. The sleep mode allows for selectively switching under-utilized base-stations to low energy consumption modes. This includes completely switching off base-stations or switching off only certain components.

Consider a time-slotted multi base-station (BS) cellular network where subsets of BSs can be dynamically activated. Since turning off BSs could adversely impact the performance perceived by users, it is important to consider the underlying energy vs. performance trade-off in designing BS activation policies. In this paper, we study the joint problem of dynamically selecting the BS activation sets and user rate allocation depending on the network load. We take into account two types of overheads involved in implementing different activation modes in the BSs.

(i) Activation cost occurs due to maintaining a BS in the active state. This includes energy spent on main power supply, air conditioning, transceivers and signal processing [7]. Surveys show that a dominant part of the energy consumption of an active base-station is due to static factors that do not have dependencies with traffic load intensities [4], [8]. Therefore, an active BS consumes almost the same energy irrespective of the amount of traffic it serves. Typically, the operation cost (including energy consumption) in the sleep state is much lower than that in the active state since it requires only minimal maintenance signaling [6].

(ii) Switching cost is the penalty due to switching a BS from active state to sleep state or vice-versa. This factors in the signaling overhead (control signaling to users, signaling over the backhaul to other BSs and/or the BS controller), state-migration processing, and switching energy consumption associated with dynamically changing the BS modes [7].

Further, switching between these states typically cannot occur instantaneously. Due to the hysteresis time involved in migrating between the active and sleep states, BS switching can be done only at a slower time-scale than that of channel scheduling [9], [10].

Main Contributions

We formulate the problem in a (stochastic) network cost minimization framework. The task is to select the set of active BSs in every time-slot, and then based on the instantaneous

S. Krishnasamy, A. Arapostathis and S. Shakkottai are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: subhashini.kb@utexas.edu).

P. T. Akhil is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India.

R. Sundaresan is with the Department of Electrical Communication Engineering and with The Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore 560012, India.

A shorter version of this paper appears in the Proceedings of IEEE Conference on Computer Communications (IEEE Infocom 2017) [1].

channel state for the activated BSs, choose a feasible allocation of rates to users. Our aim is to minimize the total network cost (sum of activation and switching costs) subject to stability of the user queues at the BSs.

While BS switching can be used to reduce energy costs both when the traffic load is dynamic and static, we consider the static case in this paper. Specifically, we assume that the incoming traffic for each user to a BS is independent and identically distributed (i.i.d.) with fixed rates. In this stationary setting, the task is to find the right way to activate and de-activate BSs so as to serve the incoming load while minimizing the energy cost. This is challenging especially because the energy cost includes the cost of switching the BSs from one state to the other. In practice, one could model the non-stationary setting as one with *regime changes*. One could then separately apply the main findings of our i.i.d. traffic load study to each regime. Our simulation studies described later in the paper suggest the modifications needed for application of our findings to settings with regime changes.

Insufficiency of the standard Lyapunov technique: Such stochastic network resource allocation problems typically adopt greedy primal dual algorithms along with virtual-queues to accommodate resource constraints [11], [12], [13]. To ensure stability, this technique crucially relies on achieving negative Lyapunov drift in some fixed number of time-slots. In our problem, unlike in the traditional setting, such an approach cannot be applied because the rates available for allocation in a time-slot is correlated with the network state in the previous time-slot. See Section IV-D.1 for more details.

To circumvent difficulties introduced through this co-evolution, we propose an approach that uses queue-lengths for channel scheduling at a fast time-scale, but explicitly uses arrival and channel statistics (using learning via an explore-exploit learning policy) for activation set scheduling at a slower time-scale. Our main contributions are as follows.

1) **Static-split Activation + Max-Weight Channel Scheduling:** We propose a solution that explicitly controls the time-scale separation between BS activation and rate allocation decisions. At BS switching instants (which occurs at a slow time-scale), the strategy uses a static-split rule (time-sharing) which is pre-computed using the explicit knowledge of the arrival and channel statistics for selecting the activation state. This activation algorithm is combined with a queue-length based Max-Weight algorithm for rate allocation (applied at the fast time-scale of channel variability). We show that the joint dynamics of these two algorithms lead to stability; further, the choice of parameters for the algorithm enables us to achieve an average network cost that is arbitrarily close to the optimal cost.
2) **Learning algorithm with provable guarantees:** In the setting where the arrival and channel statistics are not known, we propose an *explore-exploit* policy that estimates arrival and channel statistics in the explore phase, and uses the estimated statistics for activation decisions in the exploit phase (this phase includes BS switching at a slow time-scale). This is combined with a Max-Weight based rate allocation rule restricted to the activated BSs (at a fast time-scale). We prove that this joint learning-cum-scheduling algorithm can ensure queue stability while achieving close to optimal network cost.
3) **Convergence bounds for time-inhomogeneous Markov chains:** In the course of proving the theoretical guarantees for our algorithm, we derive useful technical results on convergence of time-inhomogeneous Markov chains. More specifically, we derive explicit convergence bounds for the marginal distribution of a finite-state time-inhomogeneous Markov chain whose transition probability matrices at each time-step are arbitrary (but small) perturbations of a given stochastic matrix. We believe that these bounds are useful not only in this specific problem, but are of independent interest.

To summarize then, our approach can be viewed as an algorithmically engineered separation of time-scales for only the activation set dynamics, while adapting to the channel variability for the queue dynamics. Such an engineering of time-scales leads to coupled fast-slow dynamics, the 'fast' due to opportunistic channel allocation and packet queue evolution with Max-Weight, and the 'slow' due to infrequent base-station switching using learned statistics. Through a novel Lyapunov technique for convergent time-inhomogeneous Markov chains, we show that we can achieve queue stability while operating at a near-optimal network cost.

Related Work

While mobile networks have been traditionally designed with the objective of optimizing spectral efficiency, design of energy efficient networks has been of recent interest. A survey of various techniques proposed to reduce operational costs and carbon footprint can be found in [4], [14], [3], [6]. The survey in [6] specially focuses on sleep mode techniques in BSs.

Various techniques have been proposed to exploit BS sleep mode to reduce energy consumption in different settings. Most of them aim to minimize energy consumption while guaranteeing minimum user QoS requirements. For example, [15], [7], [16] consider inelastic traffic and consider outage probability or blocking probability as metrics for measuring QoS. In [9], the problem is formulated as a utility optimization problem with the constraint that the minimum rate demand should be satisfied. But they do not explicitly evaluate the performance of their algorithm with respect to user QoS. The authors in [17], [18] model a single BS scenario with elastic traffic as an M/G/1 vacation queue and characterize the impact of sleeping on mean user delay and energy consumption. In [10], the authors consider the multi BS setting with Poisson arrivals and delay constraint at each BS.

Most papers that study BS switching use models that ignore switching cost. Nonetheless, a few papers acknowledge the importance of avoiding frequent switching. For example, Oh et al. [19] implement a hysteresis time for switching in their algorithm although they do not consider it in their theoretical analysis. Gou et al. [18] also study hysteresis sleeping schemes which enforce a minimum sleeping time. In [9] and [10], it is ensured that interval between switching times are large

enough to avoid overhead due to transient network states. Finally Jie et al. [7] consider BS sleeping strategies which explicitly incorporate switching cost in the model (but they do not consider packet queue dynamics). They emphasize that frequent switching should be avoided considering its effect on signaling overhead, device lifetime and switching energy consumption, and also note that incorporating switching cost introduces time correlation in the system dynamics.

Finally, this paper builds on the rich MaxWeight literature for opportunistic scheduling [20], [21], [22]. The literature has considered many aspects of utility maximization and tail performance [23], [12], [24], partial channel information [25], [26], and heterogeneous and inconsistent network information [27]; we refer to [13] for a comprehensive survey. Most related among these are the studies with partial information and two-stage decision making [25], [26], [28], with MaxWeight averaged through an appropriate conditional expectation for first-stage decision making, and the usual MaxWeight rule for the second stage, and with the proofs of stability shown using a Lyapunov argument. Our work differs in that the switching stemming from base-station activation does not directly permit a standard Lyapunov argument to hold (see Section IV-D.1 for additional discussion); thus we use explicit learning in the first stage, followed by the usual MaxWeight for the second stage. Our proof technique also substantially differs, as our first stage arguments are based on an analysis of time-inhomogeneous Markov Chains.

Notation: Important notation for the problem setting can be found in Table I. For any two vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and scalar a, $\mathbf{v}_1 \cdot \mathbf{v}_2$ denotes the dot product between the two vectors and $\mathbf{v}_1 + a = \mathbf{v}_1 + a\mathbf{1}$.

II. System Model

We consider a time-slotted cellular network with n users and M base-stations (BS) indexed by $u = 1, \ldots, n$ and $m = 1, \ldots, M$ respectively. Users can possibly be connected to multiple BSs. It is assumed that the user-BS association does not vary with time.

A. Arrival and Channel Model

Data packets destined for a user u arrive at a connected BS m as a bounded (at most $\bar{\mathsf{A}}$ packets in any time-slot), i.i.d. process $\{A_{m,u}(t)\}_{t\geq 1}$ with rate $\mathbb{E}[A_{m,u}(t)] = \lambda_{m,u}$. Arrivals get queued if they are not immediately transmitted. Let $Q_{m,u}(t)$ represent the queue-length of user u at BS m at the beginning of time-slot t.

The channel between the BSs and their associated users is also time-varying and i.i.d across time (but can be correlated across links), which we represent by the network channel-state process $\{H(t)\}_{t>0}$. At any time t, $H(t)$ can take values from a finite set $\mathcal{H}$ with probability mass function given by $\boldsymbol{\mu}$. Let $\bar{\mathsf{R}}$ be the maximum number of packets that can be transmitted over any link in a single time-slot. We consider an abstract model for interference by working with the set $\mathcal{R}(\mathbf{1}, h) \subset \{0, 1, \ldots, \bar{R}\}^{M\times n}$ defined as the set of all possible rate vectors (the number of packets that can be transmitted in a time-slot) achievable by non-randomized scheduling rules in a single time-slot, given that the channel state in that time-slot is h. Since the number of packets that can be transmitted per link is upper bounded by $\bar{\mathsf{R}}$, $\mathcal{R}(\mathbf{1}, h)$ has finite cardinality. For concrete examples of interference models, we refer the reader to [11, Ch. 2].

B. Resource Allocation

At any time-slot t, the scheduler has to make two types of allocation decisions:

BS Activation: Each BS can be scheduled to be in one of the two states, *ON* (*active* mode) and *OFF* (*sleep* mode). Packet transmissions can be scheduled only from BSs in the *ON* state. The cost of switching a BS from *ON* in the previous time-slot to *OFF* in the current time-slot is given by C_0 and the cost of maintaining a BS in the *ON* state in the current time-slot is given by C_1. The activation state at time t is denoted by $\mathbf{J}(t) = (J_m(t))_{m\in[M]}$, where $J_m(t) := \mathbb{1}\{\text{BS } m \text{ is } ON \text{ at time } t\}$. We also denote the set of all possible activation states, $\{0,1\}^M$, by $\mathcal{J}$. The total cost of operation, which we refer to as the *network cost*, at time t is the sum of switching and activation cost and is given by

$$C(t) := \mathsf{C}_0 \|(\mathbf{J}(t-1) - \mathbf{J}(t))^+\|_1 + \mathsf{C}_1 \|\mathbf{J}(t)\|_1. \quad (1)$$

It is assumed that the current network channel-state $H(t)$ is unavailable to the scheduler at the time of making activation decisions.

Rate Allocation: The network channel-state is observed after the BSs are switched *ON* and before the packets are scheduled for transmission. Moreover, only the part of the channel state restricted to the activated BSs, which we denote by $H(t)|_{\mathbf{J}(t)}$, can be observed. For any $j \in \mathcal{J}, h \in \mathcal{H}$, let $\mathcal{R}(j,h) \subset \{0,1,\ldots,\bar{R}\}^{M\times n}$ denote the set of all possible service rate vectors that can be allocated when the activation set is j and the channel state is h. A more precise definition of $\mathcal{R}(j,h)$ is as follows. For any $j \in \mathcal{J}$, $\mathbf{r} \in \mathbb{R}^{M\times n}$, let the product $\mathbf{r} \circ j$ be an $\mathbb{R}^{M\times n}$ matrix defined as

$$(\mathbf{r} \circ j)_{m,u} = \begin{cases} r_{m,u} & \text{if } j_m = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Also for any set $\mathcal{R} \subset \mathbb{R}^{M\times n}$, define $\mathcal{R} \circ j := \{\mathbf{r} \circ j : \mathbf{r} \in \mathcal{R}\}$. We assume that (i) a BS that is merely switched *ON* but not transmitting packets does not cause any interference in the network, and (ii) $\mathcal{R}(\mathbf{1}, h) \circ j \subseteq \mathcal{R}(\mathbf{1}, h)$ for any $j \in \mathcal{J}$. Based on these assumptions, we define $\mathcal{R}(j,h) := \mathcal{R}(\mathbf{1}, h) \circ j$. This means that $\mathcal{R}(j', h) \subseteq \mathcal{R}(j, h)$ for any $j', j \in \mathcal{J}$ such that $j' \leq j$, and $\mathcal{R}(\mathbf{1}, h)$ contains all possible rate vectors when the channel state is h for any BS activation set. Given the channel observation $H(t)|_{\mathbf{J}(t)}$, the scheduler allocates a rate vector $\mathbf{S}(t) = (S_{m,u}(t))_{m\in[M],u\in[n]}$ from the set $\mathcal{R}(\mathbf{J}(t), H(t))$ for packet transmission. This allows for draining of $S_{m,u}(t)$ packets from user u's queue at BS m for all $u \in [n]$ and $m \in [M]$.

Thus the resource allocation decision in any time-slot t is given by the tuple $(\mathbf{J}(t), \mathbf{S}(t))$. The sequence of operations in any time-slot can, thus, be summarized as follows: (i) Arrivals, (ii) BS Activation-Deactivation, (iii) Channel Observation, (iv) Rate Allocation, and (v) Packet Transmissions.

TABLE I
GENERAL NOTATION

Symbol	Description
n	Number of users
M	Number of BSs
$[l]$	The set $\{1, 2, \ldots, l\}$ for an integer l.
$A_{m,u}(t)$	Arrival for user u at BS m at time t
$\bar{\mathsf{A}}$	Maximum number of arrivals to any queue in a time-slot
$\boldsymbol{\lambda}$	Average arrival rate vector
$H(t)$	Channel state at time t
$\mathcal{H}$	Set of all possible channel states
$\boldsymbol{\mu}$	Probability mass function of channel state
$\bar{\mathsf{R}}$	Maximum service rate to any queue in a time-slot
$h\|_j$	Channel state h restricted to the activated BSs in j
$\mathcal{R}(j,h) \subseteq \mathbb{R}^{M\times n}$	Set of all possible rate vectors for activation vector j and channel state h
$\mathbf{J}(t) = (J_m(t))$	Activation vector at time t
$\mathcal{J}$	Set of all possible activation states
$\mathbf{S}(t) = (S_{m,u}(t))$	Rate allocation at time t
C_1	Cost of operating a BS in ON state
C_0	Cost of switching a BS from ON to OFF state
$C(t)$	Network cost at time t
$Q_{m,u}(t)$	Queue of user u at BS m at the beginning of time-slot t
$\mathcal{P}_l$	Set of all probability (row) vectors in $\mathbb{R}^l$
$\mathcal{P}_l^2$	Set of all stochastic matrices in $\mathbb{R}^{l\times l}$
$\mathcal{W}_l$	Set of all stochastic matrices in $\mathbb{R}^{l\times l}$ with a single ergodic class
$\mathbf{1}_l$	All 1's Column vector of size l
$\mathbf{I}_l$	Identity matrix of size l

C. Model Extensions

Some of the assumptions in the model above are made for ease of exposition and can be extended in the following ways:
(i) Network Cost: We assume that the cost of operating a BS in the *OFF* state (sleep mode) is zero. However, it is easy to include an additional parameter, say C_1', which denotes the cost of a BS in the *OFF* state. Similarly, for switching cost, although we consider only the cost of switching a BS from *ON* to *OFF* state, we can also include the cost of switching from *OFF* to *ON* state (say C_0'). The analysis in this chapter can then be extended by defining the network cost as

$$C(t) = \mathsf{C}_0\|(\mathbf{J}(t-1) - \mathbf{J}(t))^+\|_1 + \mathsf{C}_1\|\mathbf{J}(t)\|_1 + \mathsf{C}_0'\|(\mathbf{J}(t) - \mathbf{J}(t-1))^+\|_1 + \mathsf{C}_1'\left(M - \|\mathbf{J}(t)\|_1\right)$$

instead of (1).
(ii) Switching Hysteresis Time: While our system allows switching decisions in every time-slot, we will see that the key to our approach is a slowing of activation set switching dynamics. Specifically, on average our algorithm switches activation states once every $1/\epsilon_s$ timeslots, where ϵ_s is a tunable parameter. Additionally, it is easy to incorporate "hard constraints" on the hysteresis time by restricting the frequency of switching decisions to, say once in every L time-slots (for some constant L). This avoids the problem of switching too frequently and gives a method to implement time-scale separation between the channel allocation decisions and BS activation decisions. While our current algorithm has inter-switching times i.i.d. geometric with mean $1/\epsilon_s$, it is easy to allow other distributions that have bounded means with some independence conditions (independent of each other and also the arrivals and the channel). We skip details in the proofs for notational clarity.

III. Optimization Framework

For any $t \in \mathbb{N}$, let $\mathcal{F}_t = \left(\mathbf{A}(l), \mathbf{J}(l), H(l)|_{\mathbf{J}(l)}, \mathbf{S}(l)\right)_{l=1}^{t-1}$. A policy is given by a (possibly random) sequence of resource allocation decisions $(\mathbf{J}(t), \mathbf{S}(t))_{t>0}$ where, at any time t, the decision may depend on the information from random variables observed in the past but not the future, i.e., BS activation may depend on $\mathcal{F}_t$ and rate allocation on $\left(\mathcal{F}_t, \mathbf{J}(t), H(t)|_{\mathbf{J}(t)}\right)$. Let $(\mathbf{J}(t-1), \mathbf{Q}(t))$ be the network state at time t. The rationale behind this choice of network state is to construct policies that provide control over switching costs.

Notation: We use $\mathbb{P}_\varphi[\cdot]$ and $\mathbb{E}_\varphi[\cdot]$ to denote probabilities and expectation under policy φ. We skip the subscript when the policy is clear from the context.

A. Stability, Network Cost, and the Optimization Problem

Definition 1 (Stability)**.** *A network is said to be* stable *under a policy φ if there exist constants $\bar{Q}$, $\rho > 0$ such that for any initial condition $(\mathbf{J}(0), \mathbf{Q}(1))$,*

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}_\varphi\left[\sum_{m\in[m],u\in[n]} Q_{m,u}(t) \le \bar{Q} \,\middle|\, \mathbf{J}(0), \mathbf{Q}(1)\right] > \rho. \tag{3}$$

Remark 1. *The above definition of stability is applicable for a general network state process that is not necessarily Markov. It is motivated by the fact that for an aperiodic and irreducible DTMC, Definition 1 implies positive recurrence. Indeed, for such a DTMC, we can conclude from (3) that*

$$\limsup_{T\to\infty} \mathbb{P}_\varphi\left[\sum_{m\in[m],u\in[n]} Q_{m,u}(t) \le \bar{Q} \,\middle|\, \mathbf{J}(0), \mathbf{Q}(1)\right] > \rho \tag{4}$$

holds and hence the DTMC is recurrent; further (4) violates the necessary condition for null recurrence ([29, Th. 21.17]):

$$\lim_{t\to\infty} \mathbb{P}_\varphi\left[\mathbf{Q}(t) = \mathbf{q} \,\middle|\, \mathbf{J}(0), \mathbf{Q}(1)\right] = 0, \ \forall \mathbf{q},$$

and hence the DTMC is positive recurrent.

Consider the set of all ergodic Markov policies $\mathfrak{M}$, including those that know the arrival and channel statistics. A policy $\varphi \in \mathfrak{M}$ if and only if it makes (possibly randomized) allocation decisions at time t based only on the current state $(\mathbf{J}(t-1), \mathbf{Q}(t))$ (and possibly the arrival and channel statistical parameters), and the resulting network state process is an ergodic Markov chain. Later, in Section IV-C, we discuss why it is sufficient to restrict attention to this class of policies. We now define the support region of a policy and the capacity region.

Definition 2 (Support Region of a Policy φ)**.** *The support region $\Lambda^\varphi(\boldsymbol{\mu})$ of a policy φ is the set of all arrival rate vectors for which the network is stable under the policy φ.*

Definition 3 (Capacity Region)**.** *The capacity region $\Lambda(\boldsymbol{\mu})$ is the set of all arrival rate vectors for which the network is stable under some policy in $\mathfrak{M}$, i.e., $\Lambda(\boldsymbol{\mu}) := \bigcup_{\varphi\in\mathfrak{M}} \Lambda^{\varphi}(\boldsymbol{\mu})$.*

Definition 4 (Network Cost of a Policy φ)**.** *The network cost $C^{\varphi}(\boldsymbol{\mu},\boldsymbol{\lambda})$ under a policy φ is the long term average network cost (BS switching and activation costs) per time-slot, i.e.,*

$$C^{\varphi}(\boldsymbol{\mu},\boldsymbol{\lambda}) := \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\varphi}\left[C(t) \,\middle|\, \mathbf{J}(0), \mathbf{Q}(1)\right].$$

We formulate the resource allocation problem in a network cost minimization framework. Consider the problem of network cost minimization under Markov policies $\mathfrak{M}$ subject to stability. The optimal network cost is given by

$$C^{\mathfrak{M}}(\boldsymbol{\mu},\boldsymbol{\lambda}) := \inf_{\{\varphi\in\mathfrak{M}:\boldsymbol{\lambda}\in\Lambda^{\varphi}(\boldsymbol{\mu})\}} C^{\varphi}(\boldsymbol{\mu},\boldsymbol{\lambda}). \tag{5}$$

B. Markov-Static-Split Rules

The capacity region $\Lambda(\boldsymbol{\mu})$ will naturally be characterized by only those Markov policies that maintain all the BSs active in all the time-slots, i.e., $\mathbf{J}(t) = \mathbf{1}\,\forall t$. In the traditional scheduling problem without BS switching, it is well-known that the capacity region can be characterized by the class of *static-split* policies [21] that allocate rates in a random i.i.d. fashion given the current channel state. An arrival rate vector $\boldsymbol{\lambda}\in\Lambda(\boldsymbol{\mu})$ iff there exists convex combinations $\left\{\boldsymbol{\alpha}(\mathbf{1},h)\in\mathcal{P}_{|\mathcal{R}(\mathbf{1},h)|}\right\}_{h\in\mathcal{H}}$ such that

$$\boldsymbol{\lambda} < \sum_{h\in\mathcal{H}} \mu(h) \sum_{\mathbf{r}\in\mathcal{R}(\mathbf{1},h)} \alpha_{\mathbf{r}}(\mathbf{1},h)\mathbf{r}.$$

But note that static-split rules in the above class, in which BSs are not switched *OFF*, do not optimize the network cost.

We now describe a class of activation policies called the *Markov-static-split + static-split* rules which are useful in handling the network cost. A policy is a Markov-static-split + static-split rule if it uses a time-homogeneous Markov rule for BS activation in every time-slot, and an i.i.d. static-split rule for rate allocations. For any $l\in\mathbb{N}$, let $\mathcal{W}_l$ denote the set of all stochastic matrices of size l with a single ergodic class. A Markov-static-split + static-split rule is characterized by

1) a stochastic matrix $\mathbf{P}\in\mathcal{W}_{|\mathcal{J}|}$ with a single ergodic class,
2) convex combinations $\left\{\boldsymbol{\alpha}(j,h)\in\mathcal{P}_{|\mathcal{R}(j,h)|}\right\}_{j\in\mathcal{J},h\in\mathcal{H}}$.

Here $\mathbf{P}$ represents the transition probability matrix that specifies the jump probabilities from one activation state to another in successive time-slots. $\{\boldsymbol{\alpha}(j,h)\}_{j\in\mathcal{J},h\in\mathcal{H}}$ specify the static-split rate allocation policy given the activation state and the network channel-state.

Let $\mathfrak{MS}$ denote the class of all Markov-static-split + static-split rules. For a rule $\left(\mathbf{P},\boldsymbol{\alpha}=\{\boldsymbol{\alpha}(j,h)\}_{j\in\mathcal{J},h\in\mathcal{H}}\right)\in\mathfrak{MS}$, let $\boldsymbol{\sigma}$ denote the invariant probability distribution corresponding to the stochastic matrix $\mathbf{P}$. Then the expected switching and activation costs are given by $\mathsf{C}_0\sum_{j',j\in\mathcal{J}}\sigma_{j'}P_{j',j}\|(j'-j)^{+}\|_1$ and $\mathsf{C}_1\sum_{j\in\mathcal{J}}\sigma_j\|j\|_1$ respectively. We prove in the following theorem that the class $\mathfrak{MS}$ can achieve the same performance as $\mathfrak{M}$, the class of all ergodic Markov policies.

Theorem 1. *For any $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and $\varphi\in\mathfrak{M}$ such that $\boldsymbol{\lambda}\in\Lambda^{\varphi}(\boldsymbol{\mu})$, there exists a $\varphi'\in\mathfrak{MS}$ such that $\boldsymbol{\lambda}\in\Lambda^{\varphi'}(\boldsymbol{\mu})$ and $C^{\varphi'}(\boldsymbol{\mu},\boldsymbol{\lambda}) = C^{\varphi}(\boldsymbol{\mu},\boldsymbol{\lambda})$. Therefore,*

$$C^{\mathfrak{M}}(\boldsymbol{\mu},\boldsymbol{\lambda}) = \inf_{\varphi'\in\mathfrak{MS},\boldsymbol{\lambda}\in\Lambda^{\varphi'}(\boldsymbol{\mu})} C^{\varphi'}(\boldsymbol{\mu},\boldsymbol{\lambda}).$$

Proof Outline. The proof of this theorem is similar to the proof of characterization of the stability region using the class of static-split policies. It maps the time-averages of BS activation transitions and rate allocations of the policy $\varphi\in\mathfrak{M}$ to a Markov-static-split rule $\varphi'\in\mathfrak{MS}$ that mimics the same time-averages. (Detailed proof is in the Appendix.) □

From the characterization of the class $\mathfrak{MS}$, Theorem 1 shows that the optimal cost $C^{\mathfrak{M}}(\boldsymbol{\mu},\boldsymbol{\lambda})$ is equal to the optimal value of the optimization problem $V(\boldsymbol{\mu},\boldsymbol{\lambda})$, which is given by

$$\inf_{\mathbf{P},\boldsymbol{\alpha}} \mathsf{C}_0\sum_{j',j\in\mathcal{J}}\sigma_{j'}P_{j',j}\|(j'-j)^{+}\|_1 + \mathsf{C}_1\sum_{j\in\mathcal{J}}\sigma_j\|j\|_1$$

such that $\mathbf{P}\in\mathcal{W}_{|\mathcal{J}|}$ with unique invariant distribution $\boldsymbol{\sigma}\in\mathcal{P}_{|\mathcal{J}|}$, and $\boldsymbol{\alpha}(j,h)\in\mathcal{P}_{|\mathcal{R}(j,h)|}\,\forall j\in\mathcal{J},h\in\mathcal{H}$ with

$$\boldsymbol{\lambda} < \sum_{j\in\mathcal{J}}\sigma_j\sum_{h\in\mathcal{H}}\mu(h)\sum_{\mathbf{r}\in\mathcal{R}(j,h)}\alpha_{\mathbf{r}}(j,h)\mathbf{r}. \tag{6}$$

C. A Modified Optimization Problem

Now, consider the linear program given by

$$\min_{\boldsymbol{\sigma},\boldsymbol{\beta}} \mathsf{C}_1\sum_{j\in\mathcal{J}}\sigma_j\|j\|_1, \quad \text{such that}$$

$$\begin{aligned}
&\boldsymbol{\sigma}\in\mathcal{P}_{|\mathcal{J}|}\\
&\beta_{j,h,\mathbf{r}}\geq 0 \quad \forall\mathbf{r}\in\mathcal{R}(j,h),\forall j\in\mathcal{J},h\in\mathcal{H},\\
&\sigma_j = \sum_{\mathbf{r}\in\mathcal{R}(j,h)}\beta_{j,h,\mathbf{r}} \quad \forall j\in\mathcal{J},h\in\mathcal{H}, &(7)\\
&\boldsymbol{\lambda}\leq \sum_{\substack{j\in\mathcal{J},h\in\mathcal{H},\\ \mathbf{r}\in\mathcal{R}(j,h)}}\beta_{j,h,\mathbf{r}}\mu(h)\mathbf{r}. &(8)
\end{aligned}$$

The constraint (7) forces the right-hand side to be a constant over $h\in\mathcal{H}$.

Let $d := |\mathcal{J}| + \sum_{j\in\mathcal{J},h\in\mathcal{H}}|\mathcal{R}(j,h)|$ be the number of variables in the above linear program. We denote by $L_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda})$, a linear program with constraints as above and with $\mathbf{c}\in\mathbb{R}^d$ as the vector of weights in the objective function. Thus, the feasible set of the linear program $L_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda})$ is specified by the parameters $\boldsymbol{\mu},\boldsymbol{\lambda}$ and the objective function is specified by the vector $\mathbf{c}$. Let $C^{*}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda})$ denote the optimal value of $L_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda})$ and $\mathcal{O}^{*}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda})$ denote the optimal solution set. Also, let

$$\mathcal{S} := \{(\boldsymbol{\mu},\boldsymbol{\lambda}) : \boldsymbol{\lambda}\in\Lambda(\boldsymbol{\mu})\},$$

$$\mathcal{U}_{\mathbf{c}} := \{(\boldsymbol{\mu},\boldsymbol{\lambda})\in\mathcal{S} : L_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda}) \text{ has a unique solution}\}.$$

We claim that $L_{\mathbf{c}^0}(\boldsymbol{\mu},\boldsymbol{\lambda})$, with

$$\mathbf{c}^0 := ((\mathsf{C}_1\|j\|_1)_{j\in\mathcal{J}},\mathbf{0}) \tag{9}$$

provides a lower bound on the value of the original optimization problem $V(\boldsymbol{\mu},\boldsymbol{\lambda})$. To see this, observe that we can lower bound the value by removing the switching cost from

the objective. Then change variables $\beta_{j,h,\mathbf{r}} = \sigma_j \alpha_{\mathbf{r}}(j,h)$ to reach the new form, but with strict inequality in the last constraint on (8), and then relax this inequality. Finally, (7) is met because $\sum_{\mathbf{r}\in\mathcal{R}(j,h)} \alpha_{\mathbf{r}}(j,h) = 1$. These observations establish the claim. Therefore

$$C^*_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda}) \le C^{\mathfrak{M}}(\boldsymbol{\mu},\boldsymbol{\lambda}). \tag{10}$$

We use results from [30], [31] to show (in the Lemma below) that the solution set and the optimal value of the linear program are continuous functions of the input parameters.

Lemma 1. *(I) As a function of the weight vector* $\mathbf{c}$ *and the parameters* $\boldsymbol{\mu},\boldsymbol{\lambda}$*, the optimal value* $C^*_{(\cdot)}(\cdot)$ *is continuous at any* $(\mathbf{c},(\boldsymbol{\mu},\boldsymbol{\lambda})) \in \mathbb{R}^d \times \mathcal{S}$.

(II) For any weight vector $\mathbf{c}$*, the optimal solution set* $\mathcal{O}^*_{\mathbf{c}}(\cdot)$*, as a function of the parameters* $(\boldsymbol{\mu},\boldsymbol{\lambda})$*, is continuous at any* $(\boldsymbol{\mu},\boldsymbol{\lambda}) \in \mathcal{U}_{\mathbf{c}}$.

Remark 2. *Since* $\mathcal{O}^*_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda})$ *is a singleton if* $(\boldsymbol{\mu},\boldsymbol{\lambda}) \in \mathcal{U}_{\mathbf{c}}$*, the definition of continuity in this context is unambiguous.*

D. A Feasible Solution: Static-Split + Max-Weight

We now discuss how we can use the linear program L to obtain a feasible solution for the original optimization problem (5). We need to deal with two modified constraints:

(i) Single Ergodic Class – Spectral Gap: For any $\boldsymbol{\sigma} \in \mathcal{P}_{|\mathcal{J}|}$ and $\epsilon_s \in (0,1)$, the stochastic matrix

$$\mathbf{P}(\boldsymbol{\sigma},\epsilon_s) := \epsilon_s \mathbf{1}_{|\mathcal{J}|}\boldsymbol{\sigma} + (1-\epsilon_s)\mathbf{I}_{|\mathcal{J}|} \tag{11}$$

is aperiodic and has a single ergodic class given by $\{j : \sigma_j > 0\}$ with $\boldsymbol{\sigma}$ as the invariant distribution. Therefore, given any optimal solution $(\boldsymbol{\sigma},\boldsymbol{\beta})$ for the relaxed problem $L_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\lambda})$, we can construct a feasible solution $(\mathbf{P}(\boldsymbol{\sigma},\epsilon_s),\boldsymbol{\alpha})$ for the original optimization problem $V(\boldsymbol{\mu},\boldsymbol{\lambda})$ such that the network cost for this solution is at most $\epsilon_s M \mathsf{C}_0$ more than the optimal cost. Note that ϵ_s is the spectral gap of the matrix $\mathbf{P}(\boldsymbol{\sigma},\epsilon_s)$.

(ii) Stability – Capacity Gap: To ensure stability, it is necessary that the arrival rate is strictly less than the service rate (inequality (6)). It can be shown that an optimal solution to the linear program satisfies the constraint (8) with equality, and therefore cannot guarantee stability. An easy remedy to this problem is to solve a modified linear program with a fixed small gap ϵ_g between the arrival rate and the offered service rate. We refer to the parameter ϵ_g as the *capacity gap*. Continuity of the optimal cost of the linear program L (from part (I) of Lemma 1) ensures that the optimal cost of the modified linear program is close to the optimal cost of the original optimization problem for sufficiently small ϵ_g.

To summarize, if the statistical parameters $\boldsymbol{\mu},\boldsymbol{\lambda}$ were known, one could adopt the following scheduling policy:

(a) BS activation: Compute an optimal solution $(\boldsymbol{\sigma}^*,\boldsymbol{\beta}^*)$ for the linear program $L_{\mathbf{c}^0}(\boldsymbol{\mu},\boldsymbol{\lambda}+\epsilon_g)$. At every time-slot, with probability $1-\epsilon_s$, maintain the BSs in the same state as the previous time-slot, i.e., no switching. With probability ϵ_s, choose a new BS state according to the static-split rule given by $\boldsymbol{\sigma}^*$. The network can be operated at a cost close to the optimal by choosing ϵ_s, ϵ_g sufficiently small.

(b) Rate allocation: To ensure stability, use a queue-based rule such as the Max-Weight rule to allocate rates given the observed channel state:

$$\mathbf{S}(t) = \underset{\mathbf{r}\in\mathcal{R}(\mathbf{J}(t),H(t))}{\arg\max} \mathbf{Q}(t)\cdot\mathbf{r}. \tag{12}$$

We denote the above static-split + Max-Weight rule with parameters ϵ_s, ϵ_g by $\varphi(\boldsymbol{\mu},\boldsymbol{\lambda}+\epsilon_g,\epsilon_s)$. Theorem 2 shows that the static-split + Max-Weight policy achieves close to optimal cost while ensuring queue stability.

Theorem 2. *For any* $\boldsymbol{\mu},\boldsymbol{\lambda}$ *such that* $(\boldsymbol{\mu},\boldsymbol{\lambda}+2\epsilon_g) \in \mathcal{S}$*, and for any* $\epsilon_s \in (0,1)$*, under the static-split + Max-Weight rule* $\varphi(\boldsymbol{\mu},\boldsymbol{\lambda}+\epsilon_g,\epsilon_s)$*,*

1) the network cost satisfies

$$C^{\varphi(\boldsymbol{\mu},\boldsymbol{\lambda}+\epsilon_g,\epsilon_s)}(\boldsymbol{\mu},\boldsymbol{\lambda}) \le C^{\mathfrak{M}}(\boldsymbol{\mu},\boldsymbol{\lambda}) + \kappa\epsilon_s + \gamma(\epsilon_g),$$

for some constant κ *that depends on the network size and* C_0*,* C_1*, and for some increasing function* $\gamma(\cdot)$ *such that* $\lim_{\epsilon_g\to 0}\gamma(\epsilon_g) = 0$*, and*

2) the network is stable, i.e.,

$$\boldsymbol{\lambda} \in \Lambda^{\varphi(\boldsymbol{\mu},\boldsymbol{\lambda}+\epsilon_g,\epsilon_s)}(\boldsymbol{\mu}).$$

Proof Outline. Since $\mathbf{P}(\boldsymbol{\sigma}^*,\epsilon_s)$ has a single ergodic class, the marginal distribution of the activation state $(\mathbf{J}(t))_{t>0}$ converges to $\boldsymbol{\sigma}^*$. Part 1 of the theorem then follows from (10) and the continuity of the optimal value of L (Lemma 1(I)). Part 2 relies on the strict inequality gap enforced by ϵ_g in (6). Therefore, it is possible to serve all the arrivals in the long-term. We use a standard Lyapunov argument which shows that the T-step quadratic Lyapunov drift for the queues is strictly negative outside a finite set for some $T > 0$. A complete proof of this theorem can be found in the Appendix. □

One can also achieve the above guarantees with a static-split + static-split rule which has BS activations as above, but channel allocation through a static-split rule with convex combinations given by $\boldsymbol{\alpha}^*$ such that

$$\alpha^*_{\mathbf{r}}(j,h) = \frac{\beta^*_{j,h,\mathbf{r}}}{\sigma^*_j} \quad \forall \mathbf{r} \in \mathcal{R}(j,h), \forall j \in \mathcal{J}, h \in \mathcal{H}. \tag{13}$$

E. Effect of Parameter Choice on Performance

The constants ϵ_s and ϵ_g can be used as control parameters to trade-off between two desirable but conflicting features — small queue lengths and low network cost.

(i) Spectral gap, ϵ_s: ϵ_s is the spectral gap of the transition probability matrix $\mathbf{P}(\boldsymbol{\sigma}^*,\epsilon_s)$ and, therefore, impacts the mixing time of the activation state $(\mathbf{J}(t))_{t>0}$. Since the average available service rate is dependent on the distribution of the activation state, the time taken for the queues to stabilize depends on the mixing time, and consequently, on the choice of ϵ_s. With $\epsilon_s = 1$, we are effectively ignoring switching costs, as this corresponds to a rule that chooses the activation sets in an i.i.d. manner according to the distribution $\boldsymbol{\sigma}^*$. Thus, stability is ensured but at a penalty of larger average costs. At the other extreme, when $\epsilon_s = 0$, the transition probability matrix $\mathbf{I}_{|\mathcal{J}|}$ corresponds to an activation rule that never switches the BSs from their initial activation state. This extreme naturally achieves zero switching cost, but does not

guarantee queue stability as the initial activation set is frozen for all time and may not be large enough to ensure stable queues.

(ii) Capacity gap, ϵ_g: Recall that ϵ_g is the gap enforced between the arrival rate and the allocated service rate in the linear program $L_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g)$. Since the mean queue-length is known to vary inversely as the capacity gap, the parameter ϵ_g can be used to control queue-lengths. A small ϵ_g results in low network cost and large mean queue-lengths.

IV. Policy with Unknown Statistics

In the setting where arrival and channel statistics are unknown, our interest is in designing policies that learn the arrival and channel statistics to make rate allocation and BS activation decisions. As described in Section II-B, channel rates are observed in every time-slot after activation of the BSs. Since only channel rates of activated BSs can be obtained in any time-slot, the problem naturally involves a trade-off between activating more BSs to get better channel estimates versus maintaining low network cost. Our objective is to design policies that achieve network cost close to $C^{\mathfrak{M}}$, while learning the statistics well enough to stabilize the queues.

A. An Explore-Exploit Policy

Algorithm 1 Policy $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$ with parameters ϵ_p, ϵ_s, ϵ_g

1: Generate a uniformly distributed random direction $\boldsymbol{\upsilon} \in \mathbb{R}^d$, $\|\boldsymbol{\upsilon}\|_2 = 1$.
2: Construct a perturbed weight vector
$$\mathbf{c}^{\epsilon_p} \leftarrow \mathbf{c}^0 + \epsilon_p \boldsymbol{\upsilon}.$$
3: Initialize $\hat{\boldsymbol{\mu}} \leftarrow \mathbf{0}$, $\hat{\boldsymbol{\lambda}} \leftarrow \mathbf{0}$ and $\tilde{\mathbf{J}}(0) \leftarrow \mathbf{J}(0)$.
4: **for all** $t > 0$ **do**
5: Generate $E_s(t)$, an indep. Bernoulli(ϵ_s) sample.
6: **if** $E_s(t) = 0$ **then** ▷ *No Switching*
7: $\tilde{\mathbf{J}}(t) \leftarrow \tilde{\mathbf{J}}(t-1)$.
8: **else**
9: Solve $L_{\mathbf{c}^{\epsilon_p}}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\lambda}} + \epsilon_g\right)$.
10: Select an optimal solution $\left(\hat{\boldsymbol{\sigma}}(t), \hat{\boldsymbol{\beta}}(t)\right)$.
11: Select $\tilde{\mathbf{J}}(t)$ according to the distribution $\hat{\boldsymbol{\sigma}}(t)$.
12: **end if**
13: Set $\epsilon_l(t) \leftarrow \frac{2\log t}{t}$.
14: Generate $E_l(t)$, an indep. Bernoulli($\epsilon_l(t)$) sample.
15: **if** $E_l(t) = 1$ **then** ▷ *Explore*
16: $\mathbf{J}(t) \leftarrow \mathbf{1}$ (Activate all the BSs).
17: Observe the channel state $H(t)$.
18: Update empirical distributions $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\lambda}}$.
19: **else** ▷ *Exploit*
20: $\mathbf{J}(t) \leftarrow \tilde{\mathbf{J}}(t)$.
21: Observe the channel state $H(t)|_{\mathbf{J}(t)}$.
22: **end if**
23: Allocate channels according to the Max-Weight Rule,
$$\mathbf{S}(t) \leftarrow \underset{\mathbf{r} \in \mathcal{R}(\mathbf{J}(t), H(t))}{\arg\max} \mathbf{Q}(t) \cdot \mathbf{r}.$$
24: **end for**

Algorithm 1 gives a policy $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$, which is an explore-exploit strategy similar to the ϵ-greedy policy in the multi-armed bandit problem. Here, $\epsilon_p, \epsilon_s, \epsilon_g$ are fixed parameters of the policy. If an iterative scheme is used to solve the LP (line 15 of Algorithm 1), one could initialize the iteration at the solution parameterized by the previously obtained empirical distributions.

1) Initial Perturbation of the Cost Vector: Given the original cost vector $\mathbf{c}^0$ (given by (9)), the policy first generates a slightly perturbed cost vector $\mathbf{c}^{\epsilon_p}$ by adding to $\mathbf{c}^0$ a random perturbation uniformly distributed on the ϵ_p-ball. It is easily verified that, for any $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{S}$,

$$|C^*_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda}) - C^*_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda})| \leq \sqrt{|\mathcal{H}| + 1}\mathsf{C}_1 \epsilon_p.$$

In addition, the following lemma shows that the perturbed linear program has a unique solution with probability 1.

Lemma 2. *For any* $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{S}$,

$$\mathbb{P}\left[(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{U}_{\mathbf{c}^{\epsilon_p}} \mid \mathbf{J}(0), \mathbf{Q}(1)\right] = 1.$$

2) BS Activation:

Estimated Markov-static-split rule: The policy attempts to mimic the Markov-static-split rule using the empirical means $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\lambda}})$. The vector $\tilde{\mathbf{J}}(t)$ is used to keep track of the BS activations according to the estimated Markov-static-split rule. To be precise, with probability $1 - \epsilon_s$, the policy chooses to keep the same activation set as the previous time-slot's candidate, i.e., $\tilde{\mathbf{J}}(t-1)$. With probability ϵ_s, it solves the linear program $L_{\mathbf{c}^{\epsilon_p}}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\lambda}} + \epsilon_g\right)$ with the perturbed cost vector $\mathbf{c}^{\epsilon_p}$ and parameters $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\lambda}} + \epsilon_g$ given by the empirical distribution. From an optimal solution $\left(\hat{\boldsymbol{\sigma}}(t), \hat{\boldsymbol{\beta}}(t)\right)$ of the linear program, it chooses the BS activation vector $\tilde{\mathbf{J}}(t)$ according to the distribution $\hat{\boldsymbol{\sigma}}(t)$.

Explore-Exploit: At each time, the policy chooses to either *explore* or *exploit* and accordingly selects the actual BS activation vector $\mathbf{J}(t)$. The probability that it explores, $\epsilon_l(t) = \frac{2\log t}{t}$, decreases with time.

- In the *explore* phase, the policy activates all the BSs and observes the channel. It maintains $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\lambda}}$, the empirical distribution of the channel and the empirical mean of the arrival vector respectively, obtained from samples in the explore phase.
- In the *exploit* phase, it simply chooses the activation vector given by the estimated Markov-static-split rule, i.e., $\mathbf{J}(t) = \tilde{\mathbf{J}}(t)$.

3) Rate Allocation: The policy uses the Max-Weight Rule given by (12) for channel allocation.

B. Performance Guarantees

In Theorem 3, we give stability and network cost guarantees for the proposed learning-cum-scheduling rule $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$.

Theorem 3. *For any* $\boldsymbol{\mu}, \boldsymbol{\lambda}$ *such that* $(\boldsymbol{\mu}, \boldsymbol{\lambda} + 2\epsilon_g) \in \mathcal{S}$, *and for any* $\epsilon_p, \epsilon_s \in (0, 1)$, *under the policy* $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$,

1) the network cost satisfies

$$C^{\phi(\epsilon_p, \epsilon_s, \epsilon_g)}(\boldsymbol{\mu}, \boldsymbol{\lambda}) \leq C^{\mathfrak{M}}(\boldsymbol{\mu}, \boldsymbol{\lambda}) + \kappa(\epsilon_p + \epsilon_s) + \gamma(\epsilon_g),$$

for some constant κ that depends on the network size and C_0, C_1, and for some increasing function $\gamma(\cdot)$ such that $\lim_{\epsilon_g \to 0} \gamma(\epsilon_g) = 0$, and

2) *the network is stable, i.e.,*

$$\boldsymbol{\lambda} \in \Lambda^{\phi(\epsilon_p, \epsilon_s, \epsilon_g)}(\boldsymbol{\mu}).$$

Proof Outline. As opposed to known statistical parameters for the arrivals and the channel in the Markov-static-split rule, the policy uses empirical statistics that change dynamically with time. Thus, the activation state process $(\mathbf{J}(t))_{t>0}$, in this case, is not a time-homogeneous Markov chain. However, we note that $\mathbf{J}(t)$ along with the empirical statistics forms a time-inhomogeneous Markov chain with the empirical statistics converging to the true statistics almost surely. Specifically, we show that the time taken by the algorithm to learn the parameters within a small error has a finite second moment.

We then use convergence results for time-inhomogeneous Markov chains (derived in Lemma 3 in Section V) to show convergence of the marginal distribution of the activation state $(\mathbf{J}(t))_{t>0}$. As in Theorem 2, Part 1 then follows from (10) and the continuity of the optimal value of L (Lemma 1(I)).

Part 2 requires further arguments. The queues have a negative Lyapunov drift only after the empirical estimates have converged to the true parameters within a small error. To bound the Lyapunov drift before this time, we use boundedness of the arrivals along with the existence of a second moment for the convergence time of the estimated parameters. By using a telescoping argument as in Foster's theorem, we show that this implies stability as per Definition 1. For the complete proof, please see the Appendix. □

C. Optimality of static-split + max-weight policies

We now address the restriction to ergodic Markov policies and the question of its optimality. Recall Definition 1. When there is only activation cost, and no switching cost, it is easy to see that the class of static-split policies is both cost and throughput optimal. This scenario is represented by the linear program $L_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\lambda})$ in Section III-C. The optimal cost for the problem with switching cost cannot be lower than the value of $L_{\mathbf{c}}$; see (10). The static-split + max-weight policies in Section III-D can get arbitrarily close to this value; see Theorem 2. We can thus conclude that it is sufficient to consider the class of all ergodic Markov policies. Theorem 3 finally asserts that a Markov-static-split policy for BS activation can be implemented using estimated parameters.

D. Discussion: Other Potential Approaches

Recall that our system consists of two distinct time-scales: (a) exogenous fast dynamics due the channel variability, that occurs on a per-time-slot basis, and (b) endogenous slow dynamics of learning and activation due to base-station active-sleep state dynamics. By 'exogenous', we mean that the time-scale is controlled by nature (channel process), and by 'endogenous', we mean that the time-scale is controlled by the learning-cum-activation algorithm (slowed dynamics where activation states change only infrequently). To place this in perspective, consider the following alternate approaches, each of which has defects.

1. Virtual queues + MaxWeight: As is now standard [11], [13], suppose that we encode the various costs through virtual queues (or variants there-of), and apply a MaxWeight algorithm to this collection of queues. Due to the switching cost, the effective channel, i.e., the vector of channel rates on the active collection of base-stations, has dependence across time (coupled dynamics of channel and queues) through the activation set scheduling, and voids the standard Lyapunov proof approach for showing stability. Specifically, we cannot guarantee that the time average of various activation sets chosen by this (virtual + actual queue) MaxWeight algorithm equals the corresponding optimal fractions computed using a linear program with known channel and arrival parameters.

2. Ignoring Switching Costs with Fast Dynamics: Suppose we use virtual queues to capture only the activation costs. In this case, a MaxWeight approach (selecting a new activation set and channel allocation in each time-slot) will ensure stability, but will not provide any guarantees on cost optimality as there will be frequent switching of the activation set.

3. Ignoring Switching Costs with Slowed Dynamics: Again, we use virtual queues for encoding only activation costs, and use block scheduling. In other words, re-compute an activation + channel schedule once every R time-slots, and use this fixed schedule for this block of time (pick-and-compare, periodic, frame-based algorithms [32], [33], [34], [35]). While this approach minimizes switching costs (as activation changes occur infrequently), stability properties are lost as we are not making use of opportunism arising from the wireless channel variability (the schedule is fixed for a block of time and does not adapt to instantaneous channel variations).

Our approach avoids the difficulties in each of these approaches by explicitly slowing down the time-scale of the activation set dynamics (an engineered slow time-scale), thus minimizing switching costs. However, it allows channels to be opportunistically re-allocated in each time-slot based on the instantaneous channel state (the fast time-scale of nature). The channel allocations are based on observations of channel state but only on the activated BSs. This fast-slow co-evolution of learning, activation sets and queue lengths requires a new proof approach. We combine new results (see Section V) on convergence of inhomogeneous Markov chains with Lyapunov analysis to show both stability and cost (near) optimality.

V. Convergence of a Time-Inhomogeneous Markov Process

In this section, we derive some convergence bounds for perturbed time-inhomogeneous Markov chains which are useful in proving stability and cost optimality. Let $\mathcal{P} := \{\mathbf{P}_\delta, \delta \in \Delta\}$ be a collection of stochastic matrices in $\mathbb{R}^{N \times N}$, with $\{\boldsymbol{\sigma}_\delta, \delta \in \Delta\}$ denoting the corresponding invariant probability distributions. Also, let $\mathbf{P}_*$ be an $N \times N$ aperiodic stochastic matrix with a single ergodic class and invariant probability distribution $\boldsymbol{\sigma}_*$.

Recall that for a stochastic matrix $\mathbf{P}$ the coefficient of ergodicity [36] $\tau_1(\mathbf{P})$ is defined by

$$\tau_1(\mathbf{P}) := \max_{\mathbf{z}^\mathsf{T}\mathbf{1}_N = 0\,,\, \|\mathbf{z}\|_1 = 1} \|\mathbf{P}^\mathsf{T}\mathbf{z}\|_1. \tag{14}$$

It has the following basic properties [36]:

1) $\tau_1(\mathbf{P}_1\mathbf{P}_2) \leq \tau_1(\mathbf{P}_1)\tau_1(\mathbf{P}_2)$,
2) $|\tau_1(\mathbf{P}_1) - \tau_1(\mathbf{P}_2)| \leq \|\mathbf{P}_1 - \mathbf{P}_2\|_\infty$,
3) $\|\mathbf{x}\mathbf{P} - \mathbf{y}\mathbf{P}\|_1 \leq \tau_1(\mathbf{P})\,\|\mathbf{x} - \mathbf{y}\|_1 \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathcal{P}_N$, and
4) $\tau_1(\mathbf{P}) < 1$ if and only if $\mathbf{P}$ has no pair of orthogonal rows (i.e., if it is a scrambling matrix).

By the results in [37], if $\mathbf{P}_*$ is aperiodic and has a single ergodic class then there exists an integer $\hat{m}$ such that $\mathbf{P}_*^k$ is scrambling for all $k \geq \hat{m}$. Therefore, $\tau_1(\mathbf{P}_*^k) < 1\ \forall k \geq \hat{m}$.

Define

$$\epsilon := \sup_{\delta \in \Delta} \|\mathbf{P}_\delta - \mathbf{P}_*\|_1. \tag{15}$$

Now, consider a time-inhomogeneous Markov chain $(X(t))_{t\geq 0}$ with initial distribution $\mathbf{y}(0)$, and transition probability matrix at time t given by $\mathbf{P}_{\delta_t} \in \mathcal{P}\ \forall t > 0$. Let $\{\mathbf{y}(t)\}_{t\geq 0}$ be the resulting sequence of marginal distributions. The following lemma gives a bound on the convergence of the limiting distribution of such a time-inhomogeneous DTMC to $\boldsymbol{\sigma}_*$. Additional results are available in the Appendix.

Lemma 3. *For any* $\mathbf{y}(0)$,

(a) the marginal distribution satisfies

$$\|\mathbf{y}(n) - \boldsymbol{\sigma}_*\|_1 \leq \tau_1(\mathbf{P}_*^n)\|\mathbf{y}(0) - \boldsymbol{\sigma}_*\|_1 + \epsilon \sum_{\ell=0}^{n-1} \tau_1(\mathbf{P}_*^\ell)\,, \tag{16}$$

(b) and the limiting distribution satisfies

$$\limsup_{n\to\infty} \|\mathbf{y}(n) - \boldsymbol{\sigma}_*\|_1 \leq \epsilon\, \Upsilon(\mathbf{P}_*)$$

where $\Upsilon(\mathbf{P}_*) := \sum_{\ell=0}^{\infty} \tau_1(\mathbf{P}_*^\ell) \leq \frac{\hat{m}}{1-\tau_1(\mathbf{P}_*^{\hat{m}})}$.

Proof. The trajectory $(\mathbf{y}(n))_{n>0}$ satisfies $\forall n \geq 1$,

$$\mathbf{y}(n) = \mathbf{y}(n-1)\mathbf{P}_* + \mathbf{y}(n-1)(\mathbf{P}_{\delta_{n-1}} - \mathbf{P}_*). \tag{17}$$

Using (17) recursively, we have

$$\mathbf{y}(n) = \mathbf{y}(0)\mathbf{P}_*^n + \sum_{k=1}^{n} \mathbf{y}(n-k)(\mathbf{P}_{\delta_{n-k}} - \mathbf{P}_*)\mathbf{P}_*^{k-1},$$

which gives us

$$\mathbf{y}(n) - \boldsymbol{\sigma}_* = (\mathbf{y}(0) - \boldsymbol{\sigma}_*)\mathbf{P}_*^n + \sum_{k=1}^{n} \mathbf{y}(n-k)(\mathbf{P}_{\delta_{n-k}} - \mathbf{P}_*)\mathbf{P}_*^{k-1}. \tag{18}$$

Now, taking norms and using the definitions in (14) and (15), we obtain

$$\|\mathbf{y}(n) - \boldsymbol{\sigma}_*\|_1 \leq \tau_1(\mathbf{P}_*^n)\|\mathbf{y}(0) - \boldsymbol{\sigma}_*\|_1 + \epsilon \sum_{\ell=0}^{n-1} \tau_1(\mathbf{P}_*^\ell)\,.$$

This proves part (a) of the lemma. Now, note that

$$\tau_1(\mathbf{P}_*^k) \leq \left(\tau_1(\mathbf{P}_*^m)\right)^{\lfloor k/m \rfloor} \tag{19}$$

for any positive integers k, m. Since $\tau_1(\mathbf{P}_*^{\hat{m}}) < 1$, it follows that $\lim_{n\to\infty} \tau_1(\mathbf{P}_*^n) = 0$, and

$$\Upsilon(\mathbf{P}_*) = \sum_{\ell=0}^{\infty} \tau_1(\mathbf{P}_*^\ell) \leq \frac{\hat{m}}{1-\tau_1(\mathbf{P}_*^{\hat{m}})}.$$

Using this in (16), we have

$$\limsup_{n\to\infty} \|\mathbf{y}(n) - \boldsymbol{\sigma}_*\|_1 \leq \epsilon\, \Upsilon(\mathbf{P}_*) \leq \frac{\epsilon\, \hat{m}}{1-\tau_1(\mathbf{P}_*^{\hat{m}})}\,,$$

which proves part (b) of the lemma. □

VI. Simulation Results

We present simulations that corroborate the theoretical results in this paper. The setting is as follows. There are five users and three BSs in the system. BS 1 can service users

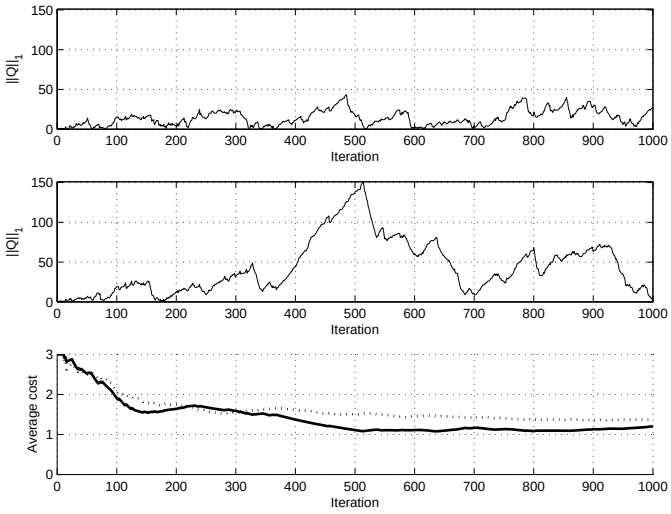

Fig. 1. The top two plots show the total queue size as a function of time when $\epsilon_s = 0.2$ and $\epsilon_s = 0.05$, respectively. The bottom plot shows the corresponding average costs (with the solid curve for $\epsilon_s = 0.05$). A smaller ϵ_s yields a lower average cost but has higher queue occupancy.

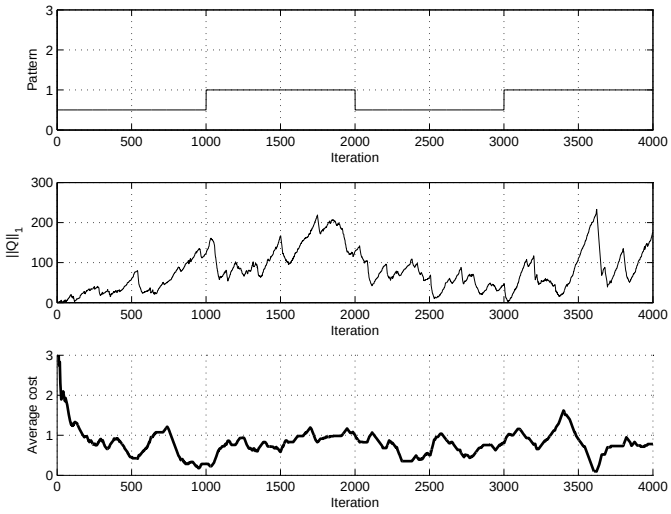

Fig. 2. The top plot shows a time-varying traffic pattern. The middle plot shows total queue size as a function of time when $\epsilon_s = 0.05$. The bottom plot shows the corresponding short-term averaged cost. The learning algorithm is modified and employs a constant learning rate so as to track the regime changes. Due to a constant learning rate, queue occupancy is a little higher, but the algorithm tracks the changes and stabilizes queue. The short-term average cost is kept small through the regime changes. The larger fluctuation in comparison to the bottommost plot in Figure 1 is due to the short-term nature of the average.

1, 2, and 5. BS 2 can service users 1, 2, 3, and 4. BS 3 can service users 3, 4, and 5. The Bernoulli arrival rates on each queue (which have to be learned by the algorithm) is 0.1 packets/slot on each mobile-BS service connection. The total arrival rate to the system is thus 0.1 packet/slot $\times$ 10 connections, or 1 packet/slot. A good channel yields a service of 2 packets/slot while a bad channel yields 1 packet/slot. In our correlated fading model, either all channels are bad, or all connections to exactly one BS are good while the others bad. This yields four correlated channel states and all four are equiprobable (the probabilities being unknown to the algorithm). The fading process is independent and identically distributed over time. The activation constraint is that each BS can service at most one mobile per slot. The per BS switching cost C_0 and activation cost C_1 are both taken to be 1.

Figure 1 provides the instantaneous queue sizes (first two plots) and time-averaged costs (third plot) for two values of ϵ_s, namely, 0.2 (first plot) and 0.05 (second plot). The plots show that a smaller ϵ_s yields a lower average cost and stabilizes the queue, but has higher queue occupancy.

Figure 2 considers a situation with regime changes (see top plot). A value 0.5 indicates that all instantaneous arrival rates are lowered by a factor 0.5. The parameter $\epsilon_s = 0.05$. The middle plot shows instantaneous total queue occupancy. The bottom plot is a short-term average cost (averaged over the past 200 slots). The algorithm was modified to keep the learning rate for estimating $\hat{\lambda}$ and $\hat{\mu}$ not below a threshold (0.001) to help track regime changes. Figure 2 indicates that the queues are stabilized but have a higher occupancy due to the use of a constant learning rate in comparison to the middle plot in Figure 1. But the short-term average cost (bottom plot) is kept small through the regime changes.

VII. Conclusion

We study the problem of jointly activating base-stations along with channel allocation, with the objective of minimizing energy costs (activation + switching) subject to packet queue stability. Our approach is based on timescale decomposition, consisting of fast-slow co-evolution of user queues (fast) and base-station activation sets (slow). We develop a learning-cum-scheduling algorithm that can achieve an average cost that is arbitrarily close to optimal, and simultaneously stabilize the user queues (shown using convergence results for inhomogeneous Markov chains).

Acknowledgements

This work was partially supported by NSF grants CNS-1017549, CNS-1161868, CNS-1343383, CNS-1731658 and DMS-1715210, Army Research Office grant W911NF-17-1-0019, the US DoT supported D-STOP Tier 1 University Transportation Center, and the Robert Bosch Centre for Cyber Physical Systems.

Appendix A
Proof of Theorem 1

Proof of Theorem 1. Consider an ergodic Markov policy $\varphi \in \mathfrak{M}$ such that $\boldsymbol{\lambda} \in \Lambda^{\varphi}(\boldsymbol{\mu})$. We use the notation $\mathbb{P}_\pi$ to denote probabilities corresponding to the stationary distribution under policy φ. Let for all $j', j \in \mathcal{J}$,

$$\sigma_j = \mathbb{P}_\pi\left[\mathbf{J}(t) = j\right],$$
$$P_{j',j} = \mathbb{P}_\pi\left[\mathbf{J}(t) = j \mid \mathbf{J}(t-1) = j'\right] \mathbb{1}\left\{\sigma_j \sigma_{j'} > 0\right\},$$

and

$$\alpha_{\mathbf{r}}(j,h) = \mathbb{P}_\pi\left[\mathbf{S}(t) = \mathbf{r} \mid \mathbf{J}(t) = j, H(t) = h\right] \quad \forall \mathbf{r} \in \mathcal{R}(j,h),\ \forall j \in \mathcal{J}, h \in \mathcal{H}.$$

We first show that $\mathbf{P} \in \mathcal{W}_{|\mathcal{J}|}$. Since $\varphi \in \mathfrak{M}$, the network state process $\{X(t)\}_{t>0}$, where $X(t) = (\mathbf{J}(t-1), \mathbf{Q}(t))$, under policy φ is an ergodic Markov chain. Therefore, for any $j', j \in \mathcal{J}$ such that $\sigma_j, \sigma_{j'} > 0$, there exists a constant $k \in \mathbb{N}$ and k states $(j_0 = j', \mathbf{q}_1), (j_1, \mathbf{q}_2), \ldots, (j_{k-1} = j, \mathbf{q}_k) \in \mathcal{J} \times \mathbb{Z}^{M \times n}$ such that for all $1 \le l \le k-1$,

$$\mathbb{P}_\pi\left[X(l+1) = (j_l, \mathbf{q}_{l+1}) \mid X(l) = (j_{l-1}, \mathbf{q}_l)\right] > 0,$$

and

$$\mathbb{P}_\pi\left[X(l) = (j_{l-1}, \mathbf{q}_l)\right] > 0.$$

This gives us that

$$P_{j_{l-1}, j_l} = \mathbb{P}_\pi\left[\mathbf{J}(l) = j_l \mid \mathbf{J}(l-1) = j_{l-1}\right] > 0.$$

Therefore, for any $j', j \in \mathcal{J}$ such that $j \neq j'$ and $\sigma_j, \sigma_{j'} > 0$, there exists a $k \in \mathbb{N}$ such that $P^k_{j',j} > 0$. A similar argument shows that P is aperiodic. In addition, we also have $\boldsymbol{\sigma} = \boldsymbol{\sigma}\mathbf{P}$, from which we can conclude that $\mathbf{P}\mathbf{1}_{\mathcal{J}} = \mathbf{1}_{\mathcal{J}}$. This proves that $\mathbf{P}$ is a stochastic matrix with a single ergodic class, i.e., $\mathbf{P} \in \mathcal{W}_{|\mathcal{J}|}$.

Further, it is easy to verify that

$$\mathsf{C}_0 \sum_{j',j \in \mathcal{J}} \sigma_{j'} P_{j',j} \|(j'-j)^+\|_1 + \mathsf{C}_1 \sum_{j \in \mathcal{J}} \sigma_j \|j\|_1 = C^{\varphi}(\boldsymbol{\mu}, \boldsymbol{\lambda}),$$

and

$$\sum_{j \in \mathcal{J}} \sigma_j \sum_{h \in \mathcal{H}} \mu(h) \sum_{\mathbf{r} \in \mathcal{R}(j,h)} \alpha_{\mathbf{r}}(j,h)\mathbf{r} = \mathbb{E}_\pi\left[\mathbf{S}(t)\right].$$

Since $\boldsymbol{\lambda} \in \Lambda^{\varphi}(\boldsymbol{\mu})$, we have $\boldsymbol{\lambda} < \mathbb{E}_\pi\left[\mathbf{S}(t)\right]$. Therefore, for

$$\varphi' := \left(\mathbf{P}, \boldsymbol{\alpha} = \{\boldsymbol{\alpha}(j,h)\}_{j \in \mathcal{J}, h \in \mathcal{H}}\right),$$

we have $\varphi' \in \mathfrak{MS}$, $\boldsymbol{\lambda} \in \Lambda^{\varphi'}(\boldsymbol{\mu})$ and $C^{\varphi'}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = C^{\varphi}(\boldsymbol{\mu}, \boldsymbol{\lambda})$. □

Appendix B
Proofs of Lemmas 1 and 2

Proof of Lemma 1. Let $\mathcal{F}$ and $\mathcal{D}$ denote the feasible sets of L and its dual respectively. By Theorem 2 in [30], to prove (I), it is sufficient to establish that $\mathcal{F}$ and $\mathcal{D}$ are continuous multifunctions on $\mathbb{R}^d \times \mathcal{S}$. The feasible set of the linear program depends only on $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ and not on $\mathbf{c}$. By Proposition 6 in [30], $\mathcal{F}$ is continuous on $\mathcal{S}$ if

(i) the dimension of $\mathcal{F}$ is constant on $\mathcal{S}$, and
(ii) for any $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{S}$, there exists a neighborhood $\mathcal{V}$ of $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ such that, if a particular inequality constraint is tight (satisfied with equality) for all $x \in \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\lambda})$, then

for any $(\boldsymbol{\mu}', \boldsymbol{\lambda}') \in \mathcal{V}$, the corresponding constraint is tight for all $x \in \mathcal{F}(\boldsymbol{\mu}', \boldsymbol{\lambda}')$.

The above two conditions are satisfied if

(i) the equality constraints are the same for every $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\lambda})$, and

(ii) for any $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{S}$, no inequality constraint is tight for every $x \in \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\lambda})$.

These can be verified to be true for all $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{S}$. Therefore, $\mathcal{F}$ is continuous on $\mathcal{S}$.

According to Corollary 11 in [30], $\mathcal{D}$ is continuous on $\mathbb{R}^d \times \mathcal{S}$ if $\mathcal{F}$ is bounded. This is again true since any feasible solution is a set of probability mass functions. Therefore, by Theorem 2 in [30], C^* is continuous on $\mathbb{R}^d \times \mathcal{S}$.

To prove (II), i.e., that the optimal solution set $\mathcal{O}^*_{\mathbf{c}}(\cdot)$ is continuous on $\mathcal{U}_{\mathbf{c}}$, we first note that $\mathcal{O}^*_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is the feasible set for a family of linear constraints, which is same as that for $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\lambda})$, in addition to the equality constraint

$$\mathbf{c} \cdot (\boldsymbol{\sigma}, \boldsymbol{\beta}) = C^*_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\lambda}).$$

By definition, the set $\mathcal{O}^*_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is non-empty for any $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{S}$, and is a singleton for any $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{U}_{\mathbf{c}}$. Now consider any $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{U}_{\mathbf{c}}$. Using (I) and Theorem 3.1 in [31], the extreme point set of $\mathcal{O}^*_{\mathbf{c}}$ is continuous at $(\boldsymbol{\mu}, \boldsymbol{\lambda})$. Since $\mathcal{O}^*_{\mathbf{c}}$ is convex and is a singleton at $(\boldsymbol{\mu}, \boldsymbol{\lambda})$, it follows that $\mathcal{O}^*_{\mathbf{c}}$ is continuous at $(\boldsymbol{\mu}, \boldsymbol{\lambda})$. □

Proof of Lemma 2. For any $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{S}$, it holds that $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{U}_{\mathbf{c}^{\epsilon_p}}$ if the vector $\mathbf{c}^{\epsilon_p}$ is not perpendicular to any of the faces of the polytope given by the feasible set of $L_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda})$. For any $1 \le k \le d-1$, consider any k-dimensional face of this polytope. The probability that the vector $\mathbf{c}^{\epsilon_p}$ lies in the $d-k$ dimensional space orthogonal to this face is zero. Since there are only a finite number of faces, by the union bound, we have

$$\mathbb{P}\left[(\boldsymbol{\mu}, \boldsymbol{\lambda}) \notin \mathcal{U}_{\mathbf{c}^{\epsilon_p}} \mid \mathbf{J}(0), \mathbf{Q}(1)\right] = 0.$$

□

Appendix C
Proof of Theorem 2

We use continuity of the linear program L (Lemma 1) to prove part 1 of Theorem 2. To prove part 2 of Theorem 2, we show that the long term Lyapunov drift is negative.

A. Cost Optimality

Part 1 of the theorem follows easily from the continuity of the optimal value of the linear program L. Since $\mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)$ has a single ergodic class, the marginal distribution of the Markov chain $\{\mathbf{J}(t)\}_{t\ge 0}$ converges to $\boldsymbol{\sigma}^*$. This gives us

$$\begin{aligned}
&\limsup_{t\to\infty} \mathbb{E}\left[C(t) \mid \mathbf{J}(0), \mathbf{Q}(1)\right] \\
&\le (1-\epsilon_s) \sum_{j' \in \mathcal{J}} \sigma^*_{j'} \mathsf{C}_1 \|j'\|_1 + \epsilon_s \left(M\mathsf{C}_0 + \sum_{j\in\mathcal{J}} \sigma^*_j \mathsf{C}_1 \|j\|_1 \right) \\
&\le C^*_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g) + M\mathsf{C}_0 \epsilon_s \\
&\le C^*_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda}) + M\mathsf{C}_0 \epsilon_s + \gamma(\epsilon_g),
\end{aligned}$$

for some increasing function $\gamma(\cdot)$ such that $\lim_{\epsilon_g \to 0} \gamma(\epsilon_g) = 0$. This follows from the continuity of $C^*_{\mathbf{c}^0}(\boldsymbol{\mu}, \cdot)$ (part (I) of Lemma 1). Therefore, for $\kappa = M\mathsf{C}_0$,

$$\begin{aligned}
\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} &\mathbb{E}\left[C(t) \mid \mathbf{J}(0), \mathbf{Q}(1)\right] \\
&\le C^*_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda}) + \kappa\epsilon_s + \gamma(\epsilon_g) \\
&\le C^{\mathfrak{M}}(\boldsymbol{\mu}, \boldsymbol{\lambda}) + \kappa\epsilon_s + \gamma(\epsilon_g),
\end{aligned}$$

where the last inequality follows from (10). This proves part 1 of Theorem 2.

B. Stability: Negative Lyapunov Drift

We show stability in the sense of Definition 1 by showing that the quadratic Lyapunov drift for the Markov policy $\varphi(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g, \epsilon_s)$ is negative outside a finite set. Let $V(\mathbf{q}) := \sum_{m,u} q^2_{m,u}$ be the Lyapunov function. For any $T > 0$, $t > 0$, let $\Delta_T(t) := V(\mathbf{Q}(t+T)) - V(\mathbf{Q}(t))$ be the T-step Lyapunov drift. Due to Foster's theorem, it is sufficient to prove the following lemma to prove part 2 of Theorem 2.

Lemma 4. *For any $\boldsymbol{\mu}, \boldsymbol{\lambda}$, there exists constants T, B such that for any $t \in \mathbb{N}$,*

$$\begin{aligned}
&\mathbb{E}_{\varphi(\boldsymbol{\mu}, \boldsymbol{\lambda}+\epsilon_g, \epsilon_s)}\left[\Delta_T(t) \mid \mathbf{J}(t-1), \mathbf{Q}(t)\right] \\
&\le BT - \epsilon_g T \sum_{m,u} Q_{m,u}(t).
\end{aligned}$$

Proof. Since the marginal distribution of the Markov chain $\{\mathbf{J}(t)\}_{t\ge 0}$ converges to $\boldsymbol{\sigma}^*$, we can choose a constant $T \in \mathbb{N}$ such that

$$\max_{j' \in \mathcal{J}} \sum_{l=0}^{T-1} \sum_{j\in\mathcal{J}} \left| \mathbb{P}\left[\mathbf{J}(l) = j \mid \mathbf{J}(0) = j'\right] - \sigma^*_j \right| \le \frac{T\epsilon_g}{2\bar{\mathsf{R}}}. \quad (20)$$

Since we have bounded arrivals and service, for $B' = nM \max\{\bar{\mathsf{A}}^2, \bar{\mathsf{R}}^2\}$, the T-step drift satisfies

$$\Delta_T(t) \le B'T + 2\sum_{l=0}^{T-1} (\mathbf{A}(t+l) - \mathbf{S}(t+l)) \cdot \mathbf{Q}(t+l).$$

Let $\boldsymbol{\alpha}^*$ be the set of convex combinations related to the unique optimal solution $(\boldsymbol{\sigma}^*, \boldsymbol{\beta}^*)$ through (13). Since the policy $\varphi(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g, \epsilon_s)$ allocates rates according to the Max-Weight rule, we have

$$\begin{aligned}
&\mathbf{S}(t+l) \cdot \mathbf{Q}(t+l) \\
&= \max_{\mathbf{r} \in \mathcal{R}(\mathbf{J}(t+l), H(t+l))} \mathbf{r} \cdot \mathbf{Q}(t+l) \\
&\ge \sum_{\mathbf{r} \in \mathcal{R}(\mathbf{J}(t+l), H(t+l))} \alpha^*_{\mathbf{r}}(\mathbf{J}(t+l), H(t+l)) \mathbf{r} \cdot \mathbf{Q}(t+l),
\end{aligned}$$

for any $l \in \{0, 1, \ldots, T-1\}$. Using the above inequality and that the arrivals and service per time-slot are bounded at every queue, for $B = B' + nM \max\{\bar{\mathsf{A}}^2, \bar{\mathsf{R}}^2\}(T-1)$, we obtain

$$\begin{aligned}
&\Delta_T(t) \\
&\le BT + 2\sum_{l=0}^{T-1} \mathbf{A}(t+l) \cdot \mathbf{Q}(t) \\
&\quad -2\sum_{l=0}^{T-1} \sum_{\mathbf{r} \in \mathcal{R}(\mathbf{J}(t+l), H(t+l))} \alpha^*_{\mathbf{r}}(\mathbf{J}(t+l), H(t+l)) \mathbf{r} \cdot \mathbf{Q}(t).
\end{aligned}$$

Taking averages in the above inequality, we get

$$\mathbb{E}\left[\Delta_T(t) \,\middle|\, \mathbf{J}(t-1), \mathbf{Q}(t)\right] \leq BT + 2(T\boldsymbol{\lambda} - Z)\cdot \mathbf{Q}(t), \tag{21}$$

where

$$Z := \sum_{l=0}^{T-1} \mathbb{E}\left[\sum_{\mathbf{r}\in\mathcal{R}(\mathbf{J}(t+l),H(t+l))} \alpha^*_{\mathbf{r}}(\mathbf{J}(t+l), H(t+l))\mathbf{r} \,\middle|\, \mathbf{J}(t-1)\right].$$

Now, for any $l \in [0, T-1]$, let

$$y(t+l)_j := \mathbb{P}\left[\mathbf{J}(t+l) = j \,\middle|\, \mathbf{J}(t-1)\right].$$

Then,

$$\begin{aligned}
&\sum_{l=0}^{T-1} \mathbb{E}\left[\sum_{\mathbf{r}\in\mathcal{R}(\mathbf{J}(t+l),H(t+l))} \alpha^*_{\mathbf{r}}(\mathbf{J}(t+l), H(t+l))\mathbf{r} \,\middle|\, \mathbf{J}(t-1)\right] \\
&= \sum_{l=0}^{T-1}\sum_{j\in\mathcal{J}} y(t+l)_j \left(\sum_{h\in\mathcal{H}}\mu(h) \sum_{\mathbf{r}\in\mathcal{R}(j,h)} \alpha^*_{\mathbf{r}}(j,h)\mathbf{r}\right) \\
&\geq T\left(\sum_{j\in\mathcal{J}}\sigma^*_j \sum_{h\in\mathcal{H}}\mu(h) \sum_{\mathbf{r}\in\mathcal{R}(j,h)} \alpha^*_{\mathbf{r}}(j,h)\mathbf{r}\right) \\
&\quad - \sum_{l=0}^{T-1} \|\mathbf{y}(t+l) - \boldsymbol{\sigma}^*\|_1 \bar{\mathsf{R}} \\
&\geq T(\boldsymbol{\lambda} + \epsilon_g/2),
\end{aligned}$$

where the last inequality follows from (20) and the fact that any solution to the linear program $L_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda}+\epsilon_g)$ satisfies its constraints

$$\sum_{j\in\mathcal{J}}\sigma^*_j \sum_{h\in\mathcal{H}}\mu(h) \sum_{\mathbf{r}\in\mathcal{R}(j,h)} \alpha^*_{\mathbf{r}}(j,h)\mathbf{r} \geq \boldsymbol{\lambda} + \epsilon_g.$$

Substituting this inequality in (21), we get the required result

$$\mathbb{E}\left[\Delta_T(t) \,\middle|\, \mathbf{J}(t-1), \mathbf{Q}(t)\right] \leq BT - T\epsilon_g \sum_{m,u} Q_{m,u}(t).$$

□

Appendix D
Proof of Theorem 3

As in the proof of Theorem 2, we use continuity of the linear program L (Lemma 1) to prove part 1 of Theorem 3. To prove stability (part 2 of Theorem 3), we show that the long term Lyapunov drift is negative outside a finite set given the event

$$\mathcal{E}^0 := (\boldsymbol{\mu}, \boldsymbol{\lambda}+\epsilon_g) \in \mathcal{U}_{\mathbf{c}^{\epsilon_p}}. \tag{22}$$

This negative Lyapunov drift, as in the Foster's theorem for time-homogeneous Markov chains, is then used to prove stability as per Definition 1.

A. Cost Optimality

According to the BS activation rule of policy $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$, we have $\mathbf{J}(t) = (1 - E_l(t))\tilde{\mathbf{J}}(t) + E_l(t)\mathbf{1}$ and $\mathbf{J}(t) \geq \tilde{\mathbf{J}}(t)$, which in turn implies that

$$\begin{aligned}
&\|(\mathbf{J}(t-1) - \mathbf{J}(t))^+\|_1 \\
&\leq ME_l(t-1) + \|(\mathbf{J}(t-1) - \mathbf{J}(t))^+\|_1 (1 - E_l(t-1)) \\
&\leq ME_l(t-1) + \|\left(\tilde{\mathbf{J}}(t-1) - \tilde{\mathbf{J}}(t)\right)^+\|_1,
\end{aligned}$$

where the last inequality can be checked via an straightforward case-by-case analysis of $E_l(t-1)$ and $E_l(t)$. Let

$$\mathbf{z}(t) := \left(\mathbb{1}\left\{\tilde{\mathbf{J}}(t) = j\right\}\right)_{j\in\mathcal{J}},$$

$$\mathbf{y}(t) := \left(\mathbb{1}\left\{\mathbf{J}(t) = j\right\}\right)_{j\in\mathcal{J}}.$$

Then, the expected cost at time t under policy $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$ is given by

$$\begin{aligned}
&\mathbb{E}\left[C(t) \,\middle|\, \mathbf{J}(0), \mathbf{Q}(1)\right] \\
&= \mathbb{E}\left[\mathsf{C}_0\|(\mathbf{J}(t-1) - \mathbf{J}(t))^+\|_1 + \mathsf{C}_1\|\mathbf{J}(t)\|_1 \,\middle|\, \mathbf{J}(0)\right] \\
&\leq \mathsf{C}_0\left(M\epsilon_l(t-1) + \mathbb{E}\left[\|\left(\tilde{\mathbf{J}}(t-1) - \tilde{\mathbf{J}}(t)\right)^+\|_1 \,\middle|\, \mathbf{J}(0)\right]\right) \\
&\quad + \mathsf{C}_1\mathbb{E}\left[\|\mathbf{J}(t)\|_1 \,\middle|\, \mathbf{J}(0)\right] \\
&= \mathsf{C}_0 M\epsilon_l(t-1) \\
&\quad + \mathsf{C}_0 \sum_{j',j\in\mathcal{J}} \mathbb{E}\left[z(t-1)_{j'} z(t)_j \,\middle|\, \mathbf{J}(0)\right] \|(j'-j)^+\|_1 \\
&\quad + \mathsf{C}_1 \sum_{j\in\mathcal{J}} \mathbb{E}[y(t)_j \,|\, \mathbf{J}(0)]\|j\|_1.
\end{aligned} \tag{23}$$

In the rest of the proof, we will suppress the dependence on the initial state $\mathbf{J}(0)$ for convenience of notation.

Let $(\hat{\boldsymbol{\mu}}(t), \hat{\boldsymbol{\lambda}}(t))$ be the estimated parameters at the beginning of time-slot t. Observe that $\hat{\boldsymbol{\mu}}(\cdot)$, $\hat{\boldsymbol{\lambda}}(\cdot)$, and consequently $\hat{\boldsymbol{\sigma}}(\cdot)$ and $\hat{\boldsymbol{\beta}}(\cdot)$, are modified only at times t when $E_l(t) = 1$. Now, consider a sample path that fixes $(E_l(\cdot), \hat{\boldsymbol{\mu}}(\cdot), \hat{\boldsymbol{\lambda}}(\cdot))$. Conditioned on this sample path, the process $\boldsymbol{z}(\cdot)$ is a time-inhomogeneous Markov chain with transition probability matrix $\mathbf{P}(\hat{\boldsymbol{\sigma}}(l), \epsilon_s)$ at time l. Hence

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{z}(t) \,\middle|\, (E_l(\cdot), \hat{\boldsymbol{\mu}}(\cdot), \hat{\boldsymbol{\lambda}}(\cdot))\right] \\
&= \mathbb{E}\left[\mathbf{z}(0)\prod_{l=1}^{t}\mathbf{P}(\hat{\boldsymbol{\sigma}}(l), \epsilon_s) \,\middle|\, (E_l(\cdot), \hat{\boldsymbol{\mu}}(\cdot), \hat{\boldsymbol{\lambda}}(\cdot))\right],
\end{aligned}$$

which when unconditioned yields

$$\mathbb{E}\left[\mathbf{z}(t)\right] = \mathbb{E}\left[\mathbf{z}(0)\prod_{l=1}^{t}\mathbf{P}(\hat{\boldsymbol{\sigma}}(l), \epsilon_s)\right].$$

Given $\mathcal{E}^0$ as defined in (22), let $(\boldsymbol{\sigma}^*, \boldsymbol{\beta}^*) \in \mathcal{O}^*_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda}+\epsilon_g)$ be the unique solution to the linear program $L_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda}+\epsilon_g)$. Since $\lim_{t\to\infty}\sum_{s=1}^{t} E_l(s) \overset{\text{a.s.}}{=} \infty$, we have $\lim_{t\to\infty}\left(\hat{\boldsymbol{\mu}}(t), \hat{\boldsymbol{\lambda}}(t)\right) \overset{\text{a.s.}}{=} (\boldsymbol{\mu}, \boldsymbol{\lambda})$ and from part (II) of Lemma 1, $\lim_{t\to\infty}\hat{\boldsymbol{\sigma}}(t) \overset{\text{a.s.}}{=} \boldsymbol{\sigma}^*$ and $\lim_{t\to\infty}\mathbf{P}(\hat{\boldsymbol{\sigma}}(t), \epsilon_s) \overset{\text{a.s.}}{=} \mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)$. Furthermore, using Lemma 3(b) and the limit law under a

sample path $(E_l(\cdot), \hat{\boldsymbol{\mu}}(\cdot), \hat{\boldsymbol{\lambda}}(\cdot))$ with all these properties, we also have

$$\lim_{t\to\infty} \mathbf{z}(0) \prod_{l=1}^{t} \mathbf{P}(\hat{\boldsymbol{\sigma}}(l), \epsilon_s) \overset{\text{a.s.}}{=} \boldsymbol{\sigma}^*,$$

which gives us $\lim_{t\to\infty} \mathbb{E}[\mathbf{z}(t)] = \boldsymbol{\sigma}^*$ by the bounded convergence theorem, and

$$\lim_{t\to\infty} \mathbb{E}[\mathbf{y}(t)] = \lim_{t\to\infty} (1-\epsilon_l(t))\mathbb{E}[\mathbf{z}(t)] + \epsilon_l(t)\boldsymbol{\eta} = \boldsymbol{\sigma}^*.$$

Similarly, for any $j', j \in \mathcal{J}$,

$$\begin{aligned}\lim_{t\to\infty} \mathbb{E}[z(t-1)_{j'} z(t)_j] &= \lim_{t\to\infty} \mathbb{E}[z(t-1)_{j'} \mathbf{P}(\hat{\boldsymbol{\sigma}}(t), \epsilon_s)_{j',j}] \\ &= \mathbb{E}\left[\lim_{t\to\infty} z(t-1)_{j'} \mathbf{P}(\hat{\boldsymbol{\sigma}}(t), \epsilon_s)_{j',j}\right] \\ &= \lim_{t\to\infty} \mathbb{E}[z(t-1)_{j'}]\, \mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)_{j',j} \\ &= \sigma^*_{j'} \mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)_{j',j},\end{aligned}$$

where the second equality is once again due to the bounded convergence theorem. Applying these along with Lemma 2 to (23) yields

$$\begin{aligned}&\limsup_{t\to\infty} \mathbb{E}\left[C(t) \mid \mathbf{J}(0), \mathbf{Q}(1)\right] \\ &\le \limsup_{t\to\infty} \mathsf{C}_0 \sum_{j',j\in\mathcal{J}} \mathbb{E}[z(t-1)_{j'} z(t)_j] \left\|(j'-j)^+\right\|_1 \\ &\quad + \limsup_{t\to\infty} \mathsf{C}_1 \sum_{j\in\mathcal{J}} \mathbb{E}[y(t)_j] \|j\|_1 \\ &= \mathsf{C}_0 \sum_{j',j\in\mathcal{J}} \sigma^*_{j'} \mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)_{j',j} \left\|(j'-j)^+\right\|_1 + \mathsf{C}_1 \sum_{j\in\mathcal{J}} \sigma^*_j \|j\|_1 \\ &\le C^*_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g) + M\mathsf{C}_0\epsilon_s \\ &\le C^*_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g) + \sqrt{|\mathcal{H}|+1}\mathsf{C}_1\epsilon_p + M\mathsf{C}_0\epsilon_s \\ &\le C^*_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda}) + \kappa(\epsilon_p + \epsilon_s) + \gamma(\epsilon_g),\end{aligned}$$

for some increasing function $\gamma(\cdot)$ such that $\lim_{\epsilon_g\to 0} \gamma(\epsilon_g) = 0$ and $\kappa = \max(\sqrt{|\mathcal{H}|+1}\mathsf{C}_1, M\mathsf{C}_0)$. This gives us

$$\begin{aligned}&\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[C(t) \mid \mathbf{J}(0), \mathbf{Q}(1)\right] \\ &\le C^*_{\mathbf{c}^0}(\boldsymbol{\mu}, \boldsymbol{\lambda}) + \kappa(\epsilon_p + \epsilon_s) + \gamma(\epsilon_g) \\ &\le C^{\mathfrak{M}}(\boldsymbol{\mu}, \boldsymbol{\lambda}) + \kappa(\epsilon_p + \epsilon_s) + \gamma(\epsilon_g),\end{aligned}$$

where the last inequality follows from (10). This proves part 1 of Theorem 3.

B. Stability: Negative Lyapunov Drift

Similar to Theorem 2, we show in Lemma 5 that the long term Lyapunov drift is negative outside a finite set. But unlike in Theorem 2, this negative drift condition holds only after a random time that has bounded second moment.

Lemma 5. *For any $\boldsymbol{\mu}, \boldsymbol{\lambda}$, there exists constants T, B and a random time Γ such that $\mathbb{E}\left[\Gamma^2 \mid \mathbf{J}(0), \mathbf{Q}(1)\right] < \infty$, and for any $t > \Gamma$,*

$$\begin{aligned}&\mathbb{E}_{\phi(\epsilon_p,\epsilon_s,\epsilon_g)}\left[\Delta_T(t) \mid \tilde{\mathbf{J}}(t-1), \mathbf{Q}(t), \mathbf{J}(0), \mathbf{Q}(1), \Gamma, \mathcal{E}^0\right] \\ &\le BT - \epsilon_g T \sum_{m,u} Q_{m,u}(t). \end{aligned} \tag{24}$$

Before we prove Lemma 5, we show below that (24) implies stability as per Definition 1.

Lemma 6. *If the condition given by (24) is satisfied, then for any $b > 0$,*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{l=1}^{t} \mathbb{P}_{\phi(\epsilon_p,\epsilon_s,\epsilon_g)}\left[\mathbf{Q}(l) \in \mathcal{A} \,\middle|\, \mathbf{J}(0), \mathbf{Q}(1)\right] > \frac{b}{\bar{B}},$$

where $\bar{B} = B + b$, $\bar{Q} = \frac{\bar{B}}{\epsilon_g}$ and $\mathcal{A} = \left\{\mathbf{Q} \in \mathbb{R}^{M\times n} : \sum_{m,u} Q_{m,u} < \bar{Q}\right\}$. Therefore, the network is stable under the policy $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$.

Proof. For ease of notation, we do not explicitly write the conditioning on $\mathbf{J}(0), \mathbf{Q}(1)$. Let the condition given by (24) be true. Then under policy $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$, we have from (24) that

$$\mathbb{E}\left[\Delta_T(t) \mid \tilde{\mathbf{J}}(t-1), \mathbf{Q}(t), \Gamma, \mathcal{E}^0\right] \le -bT + \bar{B}T\mathbb{1}\left\{\mathbf{Q}(t) \in \mathcal{A}\right\},$$

for any $t > \Gamma$. Now, let $I^* := \min\{i : (i-1)T \ge \Gamma\}$. Consider, for any $k \in \mathbb{N}$,

$$\begin{aligned}&\sum_{l=1}^{T} \mathbb{E}\left[V(\mathbf{Q}(kT+l)) - V(\mathbf{Q}(l))\right] \\ &= \sum_{l=1}^{T} \mathbb{E}\left[V(\mathbf{Q}(kT+l) - V(\mathbf{Q}((I^*-1)T+l))\right] \\ &\quad + \mathbb{E}\left[V(\mathbf{Q}((I^*-1)T+l)) - V(\mathbf{Q}(l))\right]. \end{aligned} \tag{25}$$

Now,

$$\begin{aligned}&\sum_{l=1}^{T} \mathbb{E}\left[V(\mathbf{Q}(kT+l)) - V(\mathbf{Q}((I^*-1)T+l))\right] \\ &= \sum_{l=1}^{T} \mathbb{E}\left[\sum_{i=I^*-1}^{k-1} \Delta_T(iT+l)\right] \\ &= \sum_{l=1}^{T} \mathbb{E}\left[\sum_{i=I^*-1}^{k-1} \mathbb{E}\left[\Delta_T(iT+l) \mid \mathbf{Q}(iT+l), \Gamma, \mathcal{E}^0\right]\right] \\ &\le \sum_{l=1}^{T} \mathbb{E}\left[\sum_{i=I^*-1}^{k-1} \left(-bT + \bar{B}T\mathbb{1}\left\{\mathbf{Q}(iT+l) \in \mathcal{A}\right\}\right)\right] \\ &\le -b(k-1)T^2 + bT\mathbb{E}[\Gamma] + \bar{B}T\mathbb{E}\left[\sum_{t=\Gamma+1}^{kT} \mathbb{1}\left\{\mathbf{Q}(t) \in \mathcal{A}\right\}\right]. \end{aligned} \tag{26}$$

In (26), we have used that $\mathbb{P}\left[\mathcal{E}^0\right] = 1$ from Lemma 2.

Moreover, since $(I^*-1)T < \Gamma + T$, we have

$$\mathbf{Q}((I^*-1)T + l) \le \mathbf{Q}(l) + \bar{\mathsf{A}}(\Gamma + T)$$

for any $l \in \mathbb{N}$, which gives

$$\begin{aligned}
&\sum_{l=1}^{T} \mathbb{E}\left[V(\mathbf{Q}((I^*-1)T+l)) - V(\mathbf{Q}(l))\right] \\
&\leq \sum_{l=1}^{T} \mathbb{E}\left[\sum_{m,u}\left((\bar{\mathsf{A}}(\Gamma+T)+Q_{m,u}(l))^2 - Q_{m,u}(l)^2\right)\right] \\
&\leq \mathbb{E}\left[nMT(\bar{\mathsf{A}}(\Gamma+T))^2 + 2\sum_{l=1}^{T}\sum_{m,u}\bar{\mathsf{A}}(\Gamma+T)Q_{m,u}(l)\right] \\
&\leq \mathbb{E}\left[nMT(\bar{\mathsf{A}}(\Gamma+T))^2\right] \\
&\quad + \mathbb{E}\left[2\bar{\mathsf{A}}(\Gamma+T)\sum_{l=1}^{T}\sum_{m,u}\left(\bar{\mathsf{A}}(l-1)+Q_{m,u}(1)\right)\right] \\
&\leq T\mathbb{E}\left[nM\bar{\mathsf{A}}^2(\Gamma^2+3T\Gamma+2T^2)\right] \\
&\quad + T\mathbb{E}\left[2\bar{\mathsf{A}}(\Gamma+T)\sum_{m,u}Q_{m,u}(1)\right].
\end{aligned}$$

Applying the above two inequalities in (25) and rearranging the terms, we get

$$\begin{aligned}
&\bar{B}\mathbb{E}\left[\sum_{t=\Gamma+1}^{kT}\mathbb{1}\left\{\mathbf{Q}(t)\in\mathcal{A}\right\}\right] \\
&\geq bkT - \mathbb{E}\left[Y\right] \\
&\quad + \frac{1}{T}\sum_{l=1}^{T}\mathbb{E}\left[V(\mathbf{Q}(kT+l)) - V(\mathbf{Q}(l))\right],
\end{aligned}$$

where

$$\begin{aligned}
Y = nM\bar{\mathsf{A}}^2\Gamma^2 + \left(3nM\bar{\mathsf{A}}^2T + 2\bar{\mathsf{A}}\sum_{m,u}Q_{m,u}(1) + b\right)\Gamma \\
+ 2T^2 + 2\bar{\mathsf{A}}T\sum_{m,u}Q_{m,u}(1) + bT.
\end{aligned}$$

We have $\mathbb{E}\left[Y\right] < \infty$ since $\mathbb{E}\left[\Gamma^2\right] < \infty$. Additionally, we have

$$\begin{aligned}
&\limsup_{k\to\infty}\frac{1}{kT}\mathbb{E}\left[\frac{1}{T}\sum_{l=1}^{T}V(\mathbf{Q}(l))\right] \\
&\leq \limsup_{k\to\infty}\frac{1}{kT}\mathbb{E}\left[\frac{1}{T}\sum_{l=1}^{T}V(\bar{\mathsf{A}}(l-1)+\mathbf{Q}(1))\right] = 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\bar{B}\liminf_{k\to\infty}\frac{1}{kT}\sum_{t=1}^{kT}\mathbb{P}\left[\mathbf{Q}(t)\in\mathcal{A}\right] \\
&\geq \bar{B}\liminf_{k\to\infty}\frac{1}{kT}\mathbb{E}\left[\sum_{t=\Gamma+1}^{kT}\mathbb{1}\left\{\mathbf{Q}(t)\in\mathcal{A}\right\}\right] \\
&\geq b - \limsup_{k\to\infty}\frac{1}{kT}\mathbb{E}\left[Y + \frac{1}{T}\sum_{l=1}^{T}V(\mathbf{Q}(l))\right],
\end{aligned}$$

which gives us the required result

$$\liminf_{k\to\infty}\frac{1}{kT}\sum_{t=1}^{kT}\mathbb{P}\left[\mathbf{Q}(t)\in\mathcal{A}\,\middle|\,\mathbf{J}(0),\mathbf{Q}(1)\right] \geq \frac{b}{\bar{B}}.$$

□

Before we proceed to prove Lemma 5, we will prove an intermediate result which shows that the transition probability matrices used by the policy to select the activation vector converge to a matrix that is close to the optimal. This result along with Lemma 3 allows us to show that the distribution of the activation vector converges to the optimal invariant distribution.

For any $k > 0$, let $\tilde{\boldsymbol{\mu}}(k)$, $\tilde{\boldsymbol{\lambda}}(k)$ denote the empirical distributions of channels and the empirical means of arrivals respectively obtained from the first k explore samples. For every $t > 0$, $\delta > 0$, define the events

$$\begin{aligned}
\mathcal{E}_l(t) &:= \left\{\sum_{s=1}^{t}E_l(s) \geq \frac{1}{2}\log^2 t\right\}, \\
\mathcal{E}_{\tilde{\mu}}(t,\delta) &:= \left\{\|\tilde{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}\|_1 \leq \delta\right\}, \\
\mathcal{E}_{\tilde{\lambda}}(t,\delta) &:= \left\{\|\tilde{\boldsymbol{\lambda}}(t) - \boldsymbol{\lambda}\|_1 \leq \delta\right\},
\end{aligned}$$

and for $\delta' = \frac{1}{2}\min(\delta, \epsilon_g/\bar{\mathsf{R}})$,

$$\mathcal{E}(t,\delta) := \mathcal{E}_l(t) \bigcap_{k\geq\frac{1}{2}\log^2 t}\left\{\mathcal{E}_{\tilde{\mu}}(k,\delta')\cap\mathcal{E}_{\tilde{\lambda}}(k,\delta')\right\}.$$

Lemma 7. *For any $\delta > 0$, let*

$$T_1(\delta) := \min\left\{t : \mathcal{E}(t,\delta) \text{ is true}\right\}.$$

Then, $\mathbb{E}\left[T_1(\delta)^2 \mid \mathbf{J}(0),\mathbf{Q}(1)\right] < \infty$.

Proof. For ease of notation, we do not explicitly write the conditioning on $\mathbf{J}(0),\mathbf{Q}(1)$.

Consider the mean number of explore samples in the first t slots.

$$\begin{aligned}
\mathbb{E}\left[\sum_{s=1}^{t}E_l(s)\right] &= \sum_{s=2}^{t}\frac{2\log s}{s} \geq \int_{s=e}^{t+1}\frac{2\log s}{s}\,\mathsf{d}s \\
&= \log^2(t+1) - 1 \geq \frac{3}{4}\log^2(t),
\end{aligned}$$

for all $t \geq 3$. Using the Chernoff bound for Bernoulli random variables, we have $\forall t \geq 3$,

$$\mathbb{P}\left[\mathcal{E}_l(t)^c\right] \leq \exp\left(-\frac{1}{32}\log^2 t\right).$$

Using the Hoeffding's inequality for Bernoulli random variables, for any k, $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}_{\tilde{\lambda}}(k,\epsilon)^c\right] &\leq \sum_{u=1}^{n}\mathbb{P}\left[|\tilde{\lambda}_u(k) - \lambda_u| \geq \frac{1}{n}\epsilon\right] \\
&\leq n\exp\left(-2k\left(\frac{\epsilon}{n\bar{\mathsf{A}}}\right)^2\right).
\end{aligned}$$

In addition, using Pinsker's inequality it can be shown [38] that for any k, $\epsilon > 0$,

$$\mathbb{P}\left[\mathcal{E}_{\tilde{\mu}}(k,\epsilon)^c\right] \leq (k+1)^{|\mathcal{H}|}\exp\left(-\frac{\epsilon^2}{2}k\right).$$

Now, let $\delta' = \frac{1}{2}\min(\delta, \epsilon_g/\bar{\mathsf{R}})$. Using the above inequalities, we have $\forall t \geq 3$,

$$\begin{aligned}\mathbb{P}\left[T_1(\delta) > t\right] &\leq \mathbb{P}\left[\mathcal{E}(t,\delta)^c\right] \\ &\leq \mathbb{P}\left[\mathcal{E}_l(t)^c\right] \\ &\quad + \sum_{k=\frac{1}{2}\log^2 t}^{\infty} \left(\mathbb{P}\left[\mathcal{E}_{\tilde{\mu}}(k,\delta')^c\right] + \mathbb{P}\left[\mathcal{E}_{\tilde{\lambda}}(k,\delta')^c\right]\right) \\ &= o\left(\frac{1}{t^3}\right),\end{aligned}$$

which gives us

$$\mathbb{E}\left[T_1(\delta)^2\right] = 2\sum_{t=0}^{\infty} t\mathbb{P}\left[T_1(\delta) > t\right] < \infty.$$

□

Given $\mathcal{E}^0$, let $(\boldsymbol{\sigma}^*, \boldsymbol{\beta}^*) \in \mathcal{O}^*_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g)$ be the unique solution to the linear program $L_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g)$. Due to the continuity of the solution set of $L_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g)$ given $\mathcal{E}^0$ (from Lemma 1), there exists a positive function $\delta_1 \mapsto f(\delta_1)$ such that, if $(\boldsymbol{\mu}', \boldsymbol{\lambda}' + \epsilon_g) \in \mathcal{S}$ and

$$\|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_1 + \|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\|_1 \leq f(\delta_1),$$

then for any $(\boldsymbol{\sigma}', \boldsymbol{\beta}') \in \mathcal{O}^*_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}', \boldsymbol{\lambda}' + \epsilon_g)$,

$$\|\boldsymbol{\sigma}' - \boldsymbol{\sigma}^*\|_1 \leq \delta_1.$$

The following lemma shows that continuity of the solution of $L_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu}, \boldsymbol{\lambda} + \epsilon_g)$ implies convergence of the activation vector transition probability matrices.

Lemma 8. *If $(\boldsymbol{\mu}, \boldsymbol{\lambda} + 2\epsilon_g) \in \mathcal{S}$, then for any δ_1, t, the event $\mathcal{E}^0 \cap \mathcal{E}(t, f(\delta_1))$ implies the event*

$$\tilde{\mathcal{E}}(t, \delta_1) := \left\{\|\mathbf{P}(\hat{\boldsymbol{\sigma}}(l), \epsilon_s) - \mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)\|_1 \leq \delta_1 \ \forall l > t\right\}.$$

Proof. The event $\mathcal{E}(t, f(\delta_1))$ implies that for any $l > t$, $(\hat{\boldsymbol{\mu}}(l), \hat{\boldsymbol{\lambda}}(l) + \epsilon_g) \in \mathcal{S}$ and

$$\|\hat{\boldsymbol{\mu}}(l) - \boldsymbol{\mu}\|_1 + \|\hat{\boldsymbol{\lambda}}(l) - \boldsymbol{\lambda}\|_1 \leq f(\delta_1).$$

Therefore, we have

$$\|\hat{\boldsymbol{\sigma}}(l) - \boldsymbol{\sigma}^*\|_1 \leq \delta_1,$$

which gives us

$$\|\mathbf{P}(\hat{\boldsymbol{\sigma}}(l), \epsilon_s) - \mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)\|_1 \leq \|\hat{\boldsymbol{\sigma}}(l) - \boldsymbol{\sigma}^*\|_1 \leq \delta_1.$$

□

We now prove the negative Lyapunov drift condition (Lemma 5).

Proof of Lemma 5. Fix constants $\delta_1 \in (0,1)$ and $T \in \mathbb{N}$ such that

$$\frac{2 + T\delta_1}{\epsilon_s} \leq \frac{T\epsilon_g}{2\bar{\mathsf{R}}}.$$

Define the random time $\Gamma := T_1(f(\delta_1))$. Using the fact that the policy $\phi(\epsilon_p, \epsilon_s, \epsilon_g)$ allocates rates according to the Max-Weight rule and that $\mathbf{J}(t+l) \geq \tilde{\mathbf{J}}(t+l)$, we have

$$\begin{aligned}&\mathbf{S}(t+l)\cdot\mathbf{Q}(t+l) \\ &= \max_{\mathbf{r}\in\mathcal{R}(\mathbf{J}(t+l), H(t+l))} \mathbf{r}\cdot\mathbf{Q}(t+l) \\ &\geq \max_{\mathbf{r}\in\mathcal{R}(\tilde{\mathbf{J}}(t+l), H(t+l))} \mathbf{r}\cdot\mathbf{Q}(t+l) \\ &\geq \sum_{\mathbf{r}\in\mathcal{R}(\tilde{\mathbf{J}}(t+l), H(t+l))} \alpha^*_{\mathbf{r}}(\tilde{\mathbf{J}}(t+l), H(t+l))\mathbf{r}\cdot\mathbf{Q}(t+l),\end{aligned}$$

for any $l \in \{0, 1, \ldots, T-1\}$. Following the same argument in the proof of Theorem 2, i.e., using the above inequality and that the arrivals and service per time-slot are bounded at every queue, for $B = B' + nM\max\{\bar{\mathsf{A}}^2, \bar{\mathsf{R}}^2\}(T-1)$, we obtain

$$\begin{aligned}&\Delta_T(t) \\ &\leq BT + 2\sum_{l=0}^{T-1} \mathbf{A}(t+l)\cdot\mathbf{Q}(t) \\ &\quad -2\sum_{l=0}^{T-1} \sum_{\mathbf{r}\in\mathcal{R}(\tilde{\mathbf{J}}(t+l), H(t+l))} \alpha^*_{\mathbf{r}}(\tilde{\mathbf{J}}(t+l), H(t+l))\mathbf{r}\cdot\mathbf{Q}(t).\end{aligned}$$

Taking averages in the above inequality, we have for any $t > \Gamma$,

$$\begin{aligned}&\mathbb{E}\left[\Delta_T(t) \mid \tilde{\mathbf{J}}(t-1), \mathbf{Q}(t), \mathbf{J}(0), \mathbf{Q}(1), \Gamma, \mathcal{E}^0\right] \\ &\leq BT + 2(T\boldsymbol{\lambda} - Z)\cdot\mathbf{Q}(t),\end{aligned}$$

where Z is given by (27).

To prove (24), it is sufficient to prove that $\mathbb{E}\left[\Gamma^2 \mid \mathbf{J}(0), \mathbf{Q}(1)\right] < \infty$, and for any $t > \Gamma$,

$$Z \geq T(\boldsymbol{\lambda} + \epsilon_g/2). \tag{29}$$

From Lemma 7, we have that $\mathbb{E}\left[\Gamma^2 \mid \mathbf{J}(0), \mathbf{Q}(1)\right] < \infty$.

Now to prove (29), let $Y(t) = (\tilde{\mathbf{J}}(t-1), \hat{\boldsymbol{\mu}}(t), \hat{\boldsymbol{\lambda}}(t), \Gamma)$, and for any $l \in [0, T-1]$,

$$y(t+l)_j := \mathbb{P}\left[\tilde{\mathbf{J}}(t+l) = j \mid Y(t), \mathcal{E}^0\right].$$

From Lemma 8, given $\mathcal{E}^0$, for any $t > \Gamma$ and $l \in [0, T-1]$, we have

$$\|\mathbf{P}(\hat{\boldsymbol{\sigma}}(t+l), \epsilon_s) - \mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)\|_1 \leq \delta_1.$$

Recall that

$$\mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s) = \epsilon_s \mathbf{1}_{|\mathcal{J}|}\boldsymbol{\sigma}^* + (1-\epsilon_s)\mathbf{I}_{|\mathcal{J}|},$$

and for any $l \in \mathbb{N}$,

$$\mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)^l = \left(1 - (1-\epsilon_s)^l\right)\mathbf{1}_{|\mathcal{J}|}\boldsymbol{\sigma}^* + (1-\epsilon_s)^l\mathbf{I}_{|\mathcal{J}|}.$$

Since $\tau_1(\mathbf{1}_{|\mathcal{J}|}\boldsymbol{\sigma}^*) = 0$, using the definition in (14), it can be verified that $\tau_1(\mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)^l) = (1-\epsilon_s)^l$. Therefore,

$$\Upsilon(\mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)) = \sum_{l=0}^{\infty} \tau_1(\mathbf{P}(\boldsymbol{\sigma}^*, \epsilon_s)^l) = \frac{1}{\epsilon_s}.$$

$$Z := \sum_{l=0}^{T-1} \mathbb{E}\left[\sum_{\mathbf{r}\in\mathcal{R}(\tilde{\mathbf{J}}(t+l),H(t+l))} \alpha^*_{\mathbf{r}}(\tilde{\mathbf{J}}(t+l),H(t+l))\mathbf{r} \,\middle|\, \tilde{\mathbf{J}}(t-1),\mathbf{Q}(t),\mathbf{J}(0),\mathbf{Q}(1),\Gamma,\mathcal{E}^0 \right] \tag{27}$$

$$= \sum_{l=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[\sum_{\mathbf{r}\in\mathcal{R}(\tilde{\mathbf{J}}(t+l),H(t+l))} \alpha^*_{\mathbf{r}}(\tilde{\mathbf{J}}(t+l),H(t+l))\mathbf{r} \,\middle|\, Y(t),\mathcal{E}^0 \right] \,\middle|\, \tilde{\mathbf{J}}(t-1),\Gamma,\mathcal{E}^0 \right]$$

$$= \sum_{l=0}^{T-1} \mathbb{E}\left[\sum_{j\in\mathcal{J}} y(t+l)_j \left(\sum_{h\in\mathcal{H}} \mu(h) \sum_{\mathbf{r}\in\mathcal{R}(j,h)} \alpha_{\mathbf{r}}(j,h)\mathbf{r} \right) \,\middle|\, \tilde{\mathbf{J}}(t-1),\Gamma,\mathcal{E}^0 \right]$$

$$\geq \sum_{l=0}^{T-1} \mathbb{E}\left[\left(\sum_{j\in\mathcal{J}} \sigma^*_j \sum_{h\in\mathcal{H}} \mu(h) \sum_{\mathbf{r}\in\mathcal{R}(j,h)} \alpha_{\mathbf{r}}(j,h)\mathbf{r} \right) - \|\mathbf{y}(t+l)-\boldsymbol{\sigma}^*\|_1 \bar{\mathsf{R}} \,\middle|\, \tilde{\mathbf{J}}(t-1),\Gamma,\mathcal{E}^0 \right]. \tag{28}$$

From Lemma 3(a), we have

$$\sum_{l=0}^{T-1} \|\mathbf{y}(t+l)-\boldsymbol{\sigma}^*\|_1 \leq \sum_{l=0}^{T-1} \left(2\tau_1(\mathbf{P}(\boldsymbol{\sigma}^*,\epsilon_s)^l) + \delta_1 \Upsilon(\mathbf{P}(\boldsymbol{\sigma}^*,\epsilon_s))\right) \leq \frac{2+T\delta_1}{\epsilon_s} \leq \frac{T\epsilon_g}{2\bar{\mathsf{R}}}.$$

Since any solution to the linear program $L_{\mathbf{c}^{\epsilon_p}}(\boldsymbol{\mu},\boldsymbol{\lambda}+\epsilon_g)$ satisfies its constraints, we also have

$$\sum_{j\in\mathcal{J}} \sigma^*(\boldsymbol{\mu},\boldsymbol{\lambda}+\epsilon_g)_j \sum_{h\in\mathcal{H}} \mu(h) \sum_{\mathbf{r}\in\mathcal{R}(j,h)} \alpha^*_{\mathbf{r}}(j,h)\mathbf{r} \geq \boldsymbol{\lambda}+\epsilon_g.$$

Using the above two inequalities in (28) gives us

$$Z \geq T(\boldsymbol{\lambda}+\epsilon_g/2),$$

and this proves (29). □

Combining Lemmas 5 and 6, we have part 2 of Theorem 3.

REFERENCES

[1] S. Krishnasamy, P. T. Akhil, A. Arapostathis, S. Shakkottai, and R. Sundaresan, "Augmenting max-weight with explicit learning for wireless scheduling with switching costs," in *Proceedings of IEEE Infocom*, Atlanta, GA, April 2017.

[2] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, and A. Damnjanovic, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, 2014.

[3] G. Wu, C. Yang, S. Li, and G. Y. Li, "Recent advances in energy-efficient networks and their application in 5g systems," *IEEE Wireless Communications*, vol. 22, no. 2, pp. 145–151, 2015.

[4] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 56–61, 2011.

[5] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *2009 IEEE International Conference on Communications Workshops*. IEEE, 2009, pp. 1–5.

[6] J. Wu, Y. Zhang, M. Zukerman, and E. K.-N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Comm. Surveys & Tutorials*, vol. 17, no. 2, 2015.

[7] G. Jie, Z. Sheng, and N. Zhisheng, "A dynamic programming approach for base station sleeping in cellular networks," *IEICE transactions on communications*, vol. 95, no. 2, pp. 551–562, 2012.

[8] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *2010 Future Network & Mobile Summit*. IEEE, 2010.

[9] A. Abbasi and M. Ghaderi, "Distributed base station activation for energy-efficient operation of cellular networks," in *Proceedings of the 16th ACM international conference on Modeling, analysis & simulation of wireless and mobile systems*. ACM, 2013, pp. 427–436.

[10] J. Zheng, Y. Cai, X. Chen, R. Li, and H. Zhang, "Optimal base station sleeping in green cellular networks: A distributed cooperative framework based on game theory," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4391–4406, 2015.

[11] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource Allocation and Cross-Layer Control in Wireless Networks*. NOW Publishers, Foundations and Trends in Networking, 2006.

[12] X. Lin, N. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE Journal on Selected Areas in Comm.*, 2006.

[13] R. Srikant and L. Ying, *Communication Networks – An Optimization, Control, and Stochastic Networks Perspective*. Cambridge University Press, 2014.

[14] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Comm. surveys & tutorials*, vol. 13, no. 4, 2011.

[15] F. Han, Z. Safar, W. S. Lin, Y. Chen, and K. R. Liu, "Energy-efficient cellular network operation via base station cooperation," in *2012 IEEE Intl. Conf. on Comm. (ICC)*. IEEE, 2012, pp. 4374–4378.

[16] J. Gong, J. S. Thompson, S. Zhou, and Z. Niu, "Base station sleeping and resource allocation in renewable energy powered cellular networks," *IEEE Trans. on Communications*, vol. 62, no. 11, pp. 3801–3813, 2014.

[17] I. Kamitsos, L. Andrew, H. Kim, and M. Chiang, "Optimal sleep patterns for serving delay-tolerant jobs," in *Proceedings of the 1st Intl. Conf. on Energy-Efficient Computing and Networking*. ACM, 2010, pp. 31–40.

[18] X. Guo, Z. Niu, S. Zhou, and P. Kumar, "Delay-constrained energy-optimal base station sleeping control," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1073–1085, 2016.

[19] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE transactions on wireless communications*, vol. 12, no. 5, pp. 2126–2136, 2013.

[20] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 4, pp. 1936–1948, December 1992.

[21] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel conditions," *Bell Labs Tech. Memo*, April 2000.

[22] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89–103, 2005.

[23] A. L. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Systems*, vol. 50, no. 4, pp. 401–457, 2005.

[24] V. J. Venkataramanan and X. Lin, "On wireless scheduling algorithms for minimizing the queue-overflow probability," *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 788–801, June 2010.

[25] M. J. Neely, S. T. Rager, and T. F. L. Porta, "Max weight learning algorithms for scheduling in unknown environments," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1179–1191, May 2012.

[26] A. Gopalan, C. Caramanis, and S. Shakkottai, "On wireless scheduling

with partial channel-state information," *IEEE Transactions on Information Theory*, vol. 58, no. 1, pp. 403–420, Jan 2012.

[27] L. Ying and S. Shakkottai, "On throughput optimality with delayed network-state information," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5116–5132, Aug 2011.

[28] C. Manikandan, S. Bhashyam, and R. Sundaresan, "Cross-layer scheduling with infrequent channel and queue measurements," *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, 2009.

[29] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. Providence, RI: American Mathematical Society, 2009.

[30] R. J. B. Wets, "On the continuity of the value of a linear program and of related polyhedral-valued multifunctions," *Mathematical programming study*, no. 24, pp. 14–29, 1985.

[31] M. Davidson, "Stability of the extreme point set of a polyhedron," *Jnl. of optim. theory and appl.*, vol. 90, no. 2, pp. 357–380, 1996.

[32] L. Tassiulas, "Linear complexity algorithms for maximum throughput in radio networks and input queued switches," in *IEEE Infocom*, 1998.

[33] M. J. Neely, E. Modiano, and C. E. Rohrs, "Tradeoffs in delay guarantees and computation complexity for $n \times n$ packet switches," in *Proceedings of CISS*, 2002.

[34] P. Chaporkar and S. Sarkar, "Stable scheduling policies for maximizing throughput in generalized constrained queueing," in *IEEE Infocom*, 2006.

[35] Y. Yi, A. Proutiere, and M. Chiang, "Complexity in wireless scheduling: Impact and tradeoffs," in *Proc. of the 9th ACM International Symp. on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2008.

[36] E. Seneta, *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

[37] J. M. Anthonisse and H. Tijms, "Exponential convergence of products of stochastic matrices," *Journal of Mathematical Analysis and Applications*, vol. 59, no. 2, pp. 360–364, 1977.

[38] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the L1 deviation of the empirical distribution," 2003, HP Labs Technical Report.